Topomers: A Validated Protocol for Their Self-Consistent Generation

Robert J. Jilek and Richard D. Cramer*

Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri 63144

Received January 20, 2004

The hypothesis underlying topomer development is that describing molecular structures consistently may be at least as productive as describing them more realistically but incompletely. A general protocol is detailed for deterministically generating self-consistent shapes of molecular fragments from their topologies. These and other extensions to the topomer methodology are validated by repetition of earlier benchmark studies.

INTRODUCTION

A topomer is an invariant 3D representation of a molecular fragment, generated from its 2D topology by deterministic rules that produce absolute configuration, conformation, and orientation.[1] Topomers were originally devised to enable shape similarity searching of very large virtual libraries.[2,3] It was then demonstrated in retrospective[4] and prospective[5] studies that topomer shape similarity provides at least as powerful a predictor of biological similarity as other molecular descriptors investigated. More recently it was reported[6] that topomers may provide a surprisingly robust and general solution to the vexing challenge of CoMFA alignments.

The efficacy of the topomer description while searching virtual libraries motivated creation of the "dbtop" program,[7] to perform topomer similarity searching within typical compound collections. However, the topomer generators in dbtop and in ChemSpace did not always produce identical 3D structures from a given 2D topology. Investigation showed that different orderings of the atoms in the intermediate 2D connection tables could yield different topomer geometries. Clearly the existing topomer generating algorithm was leaving indeterminant many important 3D "degeneracies". Resolving these discrepant 3D structures was so unexpectedly difficult and time-consuming that we decided to publish the first complete and detailed account of topomer generation.[1] Although the topomer description is itself proprietary,[8] any other researcher who also seeks to produce deterministic 3D structures from 2D topologies will encounter similar issues and so may benefit from our findings.

These significant improvements of the original topomer generation rules, along with extensions of topomer similarity to include pharmacophoric feature similarity[7] as well as shape similarity, and special provisions when comparing polyvalent fragments as also described below, might affect the ability of topomers to predict biological properties. So we have also repeated the two major retrospective validations[4,6] of topomer efficacy, with results reported here.

METHODS

The procedure for generating the topomer of a monovalent molecular fragment may be very briefly summarized as follows:

- A structurally distinctive "cap" is attached to the open valence, and a Concord or similar model is generated for the resulting complete structure;
- This model is oriented to superimpose the "cap" attachment bond ("root") onto a vector fixed in Cartesian space;
- Proceeding away from this "root" attachment bond, only as required to place the "most important" (typically the largest) unprocessed group farthest from the root and the next most important to the "right" of the largest (when looking away from the root along the current bond), acyclic torsional angles may be adjusted, "stereocenters" inverted,[9] and ring "puckerings" standardized;
- Removal of the cap completes the topomer conformation.

Note that the conventional force field energy (intramolecular enthalpy) of a topomer is altogether immaterial. For example any steric clashes that may result from topomer generation are completely ignored.

It is only the third step in this topomer generation procedure that requires a detailed account. The following description is "bottom-up", beginning with more precise definitions of the various structural elements that may need adjustment (torsions, (pro)chiralities, "puckerings") and the natures of those adjustments. "Precedence" rules for ordering a set of attached groups (to any atom of interest) are also set forth. The entire topomer generation procedure can then be described, supported by a structural example.

Bond "Torsions". During topomer generation, the "torsion" or dihedral angle of almost every acyclic skeletal bond is measured and usually altered. This operation requires the identification of four consecutively connected atoms, i.e., the two atoms that define the bond, plus a selected atom attached to each of those defining atoms. This assembly will be referenced below as **a**−**b**−**c**−**d**, with **b** always designating the end of the bond that is topologically nearer the "root". Precedence rules to be presented below determine which of the candidate atoms, attached to **b** or **c,** become designated as **a** and **d**. The final setting of an **a**−**b**−**c**−**d** dihedral angle depends on whether the **a**−**b** and **c**−**d** bonds are cyclic or

* Corresponding author phone: (505)995-4425 or (314)647-1099; fax: (505)995-4439; e-mail: cramer@tripos.com.

10.1021/ci049961d CCC: $27.50 © 2004 American Chemical Society
Published on Web 05/22/2004

not (**b**−**c** of course is acyclic). If both are acyclic (or if **b**−**c** is an amide or multiple bond), **a**−**b**−**c**−**d** becomes 180 degrees; if neither is acyclic, **a**−**b**−**c**−**d** becomes 60 degrees; if only one is acyclic, **a**−**b**−**c**−**d** becomes 90 degrees.

(Pro)chirality. The atoms whose attachments require assessment for left/right(ness) are a superset of those that are formally chiral within 3D molecular models. Not only must pyramidal sp^3 nitrogen be included but also enantiotopic[10] atoms. For example, within an isopropyl group (−CH-(CH_3)CH_3), after one of the methyl groups is designated as having highest precedence, the remaining −H and −CH_3 groups become nonequivalent. Furthermore and in contrast to torsions, chiral atoms within rings are also inspected and adjusted whenever both attachment bonds are acyclic.

The left/right(ness) of a particular attachment proved easy to assess on the computer screen but surprisingly difficult to determine computationally. Note that the objective is control of "local" left/right(ness), relative to the other atoms being processed concurrently, not "global" left/right(ness). For example, if the overall fragment were a phenyl with a complex ortho substituent, local "right(ness)" within the ortho substituent will be "left(ness)" within the global coordinate system. After fruitless searches for congruent algorithms within the computer graphics and interactive gaming literature, the following heuristic was empirically derived, which though somewhat ad hoc does correctly perceive the geometry in the dozens of cases examined.

The geometric objective is to determine whether a fourth point **a4** is to the "right" of a plane **a1-a2-a3**, where the **a1-a2-a3** order also establishes the viewing direction (along **a1=>a3**). The first test is whether the **a1-a2-a3** plane is parallel to any of the three global coordinate planes (e.g., perpendicular to any of the three Cartesian axes), which in practice means that none of the individual **a1**, **a2**, or **a3** *x*-, *y*-, or *z*-coordinate values differs from their mean by more than 0.2 A. If so, then determining the left(right)ness of **a4** requires only the comparison of the appropriate **a4** coordinate value to the mean coordinate value, while considering the viewing direction.

If the **a1-a2-a3** plane is not parallel to a global coordinate plane, then the cross-product of **a1=>a2** (becoming **v1**) and **a2=>a3** (becoming **v2**) is formed, as usual except that if any of the values of **a1=>a2** or **a2=>a3** are within 0.1 of 0.0, they become exactly 0.0. The vector **a2=>a4** (if the atom **a4** is bonded to **a2**) or **a3=>a4** (if **a4** is bonded to **a3**) is also formed (with the same rounding of values <0.1−0.0), to become **v3**. The remaining steps depend on which of the three elements of the cross-product **v3** has the largest magnitude. If the first element is largest, then **a4** is on the right of the plane (or on the left if **v1.z** times **v3.x** is greater than 0). If the second element is largest, then **a4** is on the right if **v1.x** times **v3.y** is less than 0. Finally, if the magnitude of the third cross-product element is largest, then **a4** is on the right if **v1.x** times **v3.z** is greater than 0.

To correct any attachment **a4** that is thus found to be on the wrong (left) side of the plane, all atoms in all groups that are attached to the atom to which **a4** is attached are reflected through that **a1-a2-a3** plane.

Puckering. Ring systems that have some degree of nonplanarity can, much like chiral atoms, exist in either of two geometric forms having equivalent internal strain energy (though nonbonded interactions with atoms external to that ring system are usually not equivalent). For example, consider the reflection of a chair conformation of the (possibly substituted) cyclohexyl fragment through the plane formed by its 1, 2, and 6 carbon atoms to create an alternate chair form. To obtain an invariant geometry for such fragments, the puckering must be standardized in much the same way that (pro)chirality is standardized, as follows. Whenever the topomer generation process encounters the first bond within a new ring system, an "entry plane" is generated from the two atoms at the end of that first cyclic bond and the last atom in the encountering chain. Then the entire ring system is enumerated (by growing the largest tree that can be grown from that ring bond without including any acyclic bonds). Two summary scores for the ring system are determined, a nonplanarity (sum of the angstrom distances of those ring atoms from the entry plane) and a nonplanarity-weighted centroid (sums of *x*-, *y*-, and *z*-coordinates over all ring atoms, but with each coordinate multiplied by the distance of that ring atom from the entry plane, divided by the nonplanarity score). The ring system is considered planar, and no action is taken if its nonplanarity score is less than 0.5. Otherwise the dihedral angle is calculated between the entry plane and a second plane defined by the same two initial ring atoms and the nonplanarity-weighted centroid. If this angle is less than 180 degrees no action need be taken. Otherwise all the atoms in the ring system, this time including all its attachments except the entry attachment, are reflected through the entry plane.

Precedence Rules. These determine which, among a set of attachments to an atom of interest, the torsional and (pro)-chiral operations are to affect. The precedence of an attachment is mostly determined by the relative properties of its "path", so the first step in establishing the relative precedence among a set of attachments is to enumerate their paths. Whenever the attachment bond is acyclic, its path will simply include all the atoms in the implicit "side chain". However, whenever the attachment bond is cyclic, its path is defined to consist only of those atoms with a "topological distance" (number of separating bonds) to the attachment point that is less than any alternative topological distance back to the starting point. To take as an example the (possibly substituted) phenyl side chain, with the two "attachment" atoms to consider being its 2 and 6 carbons, the path generated by the 2 carbon will include the 2 and 3 carbons plus their attachments, and the path generated by the 6 carbon will include the 5 and 6 carbons plus their attachments, but the 4 carbon plus its attachment (being topologically equidistant by both paths to the starting point) will belong to neither path. A more complicated example is the 2-naphthyl side chain. The path originating from the 3 carbon will include only the 3 and 4 carbons and their attachments. The other path from the 1 carbon will include all the other naphthyl carbons and their attachments, except only the fusion carbon between the 4 and 5 carbons that is equidistant from the starting point.

There is one other major determinant of attachment precedence. The overall topomer generation process usually imposes an additional distinction, between those attachment atom(s) whose path(s) include the root atom of the topomer and those whose paths do not. Depending on the current step in topomer generation, such "rooted attachments" will either be completely suppressed or else given the highest precedence. In the latter case, each individual path ends whenever the root atom is encountered, and the highest precedent attachment will be the one whose topological distance to that root is shortest.

Once such paths have been generated for each of the candidate attachments, the precedence order of those attachments is determined. This process applies an ordered set of rules in a strictly hierarchical fashion, such that lower rules are applied only to break ties that remain after the application of higher rules. In order of application, these rules are as follows:

(1) only when a root atom is specified, the path whose topological distance to that root atom is shortest has precedence;

(2) the path containing the largest number of atoms has precedence;

(3) the path having the largest molecular weight (sum of atomic weights) has precedence;

(4) the path having the highest sum of atomic weights, each atomic weight being divided by its topological distance to the root atom, has precedence (thus within 2,5-dimethylphenyl the path rooted at atom 2 has precedence over that rooted at atom 6);

(5) only when a root atom is specified, an ambiguity can remain that the following rule addresses. First an illustrative example. Assume the fragment is 4-methoxyphenyl, requiring an adjustment of the torsion of the phenyl-oxygen bond, and thereby the position of the methoxy methyl. The 3- and the 5-carbons have equal precedence according to rules 1−4, so either could become atom **a** (recalling the **a**−**b**−**c**−**d** nomenclature to define a torsion), thereby producing two very different possible locations for the methyl (and especially for any attachments to the methyl group). To resolve this ambiguity, the path whose attachment atom is to the local right (determined as described above) of the plane defined by the root atom and the **b**−**c** atoms is assigned the higher precedence. (If the root atom and the **b**−**c** atoms are collinear, as indeed in the example, and so cannot define a plane, then the *x*−*y* plane is used instead.)

Topomer Generation. The overall geometry adjustment process, the third step in topomer generation, can now be described, referring as necessary to the above descriptions of individual entities. The first step is to construct a ”(pro)-chiral list” of all tetrahedral acyclic atoms attached to at most one hydrogen. This (pro)chiral list is combined with a list of all atoms that are nonterminal, that are the end point of an acyclic and nontriple “qualifying” bond, and that also are topologically nearer the root than the alternative end point of that bond. This “qualifying bond list” also includes the temporary bond from the root atom to the temporary cap fragment, because the setting of its torsion fixes the third degree of freedom in the overall orientation of the topomer. Also the attaching atom within the cap is added to the list of “current atoms”. This “current atom” list is ordered by increasing topological distance to the attaching atom within the cap and processed in that atom-by-atom order, as follows:

(1) if the current atom is included on the (pro)chiral list, perform the (pro)chirality operation on its attachments (with paths to the root atom being omitted from its precedence candidates)

(2) On each of the current bonds between the current atom and an attachment farther from the root, perform the torsional operation as follows. No action if the bond is not a “qualified bond”, otherwise the **a** atom will be the highest in precedence among the current atom attachments, specifically the shortest atom path to the root (no action being taken if this **a**−**b** bond is triple). The **d** atom will be the highest in precedence among the attachments to **b**, of course excluding **a** itself.

(3) For each such torsional operation, only if the **c**−**d** bond is cyclic. Ensure that **d** is indeed local right of the **a**−**b**−**c** plane, and if not, reset the **a**−**b**−**c**−**d** torsion to 180 degrees greater than its current value. (Some bonds between non-planar rings will require this additional action.) Then perform the pucker operation with respect to the **c**−**d** bond (the “entry plane” being **b**−**c**−**d**).

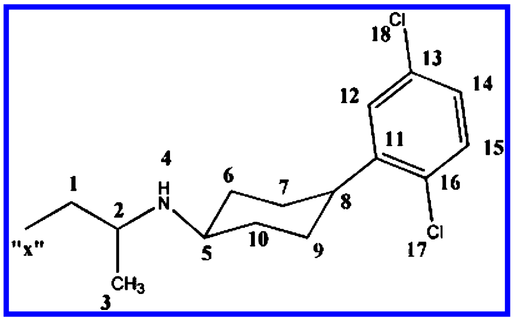

Figure 1. Structure used to exemplify the process of topomer generation. See text for explanation.

Completely processing the “current atom list” in this fashion yields the final topomer conformation.

Example. To further illustrate the process of topomer generation, the topomeric alignment of the fragment example in Figure 1 will be described. The cap is on the left, the cap attaching atom being atom “1” and the remainder of the cap being denoted by “x”. (Thus the input structure and the final topomer after removal of the cap do not include atom “1”.) Hydrogens that do not affect the resulting topomer conformation are omitted for clarity. It may be seen that the (pro)-chiral list (all tetrahedral acyclic atoms attached to at most one hydrogen) contains atoms 2 and 4. The acyclic bonds between atoms 1−2, 2−3, 2−4, 4−5, and 8−11 become the “qualified bond” list, so the final list of “current atoms” to be traversed during topomer generation is 1, 2, 4, and 8. Proceeding in order down this list:

Atom 1. No (pro)chirality operation is needed since atom 1 is not on the (pro)chiral list. The 1−2 bond needs its torsional angle set, and so the precedence of the attachments to the 1−2 bonds must be established. The first precedence rule “take the path to the root” establishes the position designated by “x” as the **a** atom (within the **a**−**b**−**c**−**d** specification**)**, while the next precedence rule “take the attachment with the most atoms” clearly favors atom 4 over atom 3 as the **d** atom. Therefore the dihedral angle of the x-1−2−4 bond is changed to 180 degrees, to appear much as shown. The 2−4 bond is not in a ring so no pucker adjustment need be considered.

Atom 2. Its highest precedent attachment (excluding the root) is again the remainder of the fragment. However atom 2 is on the (pro)chiral list, so the (pro)chirality standardization procedure described above is applied to atom 2. There are then two bonds away from atom 2 whose torsions need attention, 2−3 and 2−4. In both cases atom 1 as the head of the shortest path to the root becomes the **a** atom. All of the attachments to atom 3 (hydrogens) are equivalent in precedence, so the selection of the **d** atom is completely arbitrary, the topomer geometry of course becoming the same regardless of which hydrogen becomes the **d** atom in the setting of 1−2−3-H to 180 degrees. Finally the dihedral angle about 2−4 can be addressed. There are two attachments to atom 4, the hydrogen and the rest of the fragment, the latter having higher precedence because it has more atoms so that it is the 1−2−4−5 dihedral which is set to 180 degrees.

Atom 4. Its highest precedent attachment (excluding the root) is again the remainder of the fragment. There is only one other attachment to 4, the hydrogen as shown. Since 4 is found in the (pro)chiral list, it must be adjusted as described above (to ensure that the hydrogen is located to the right of the main chain). There is one torsional angle to be established. Its **a** atom is again determined by the shortest path back to the route. However the selection of the **d** atom is complicated. It will be evident that the paths away from atom 5, beginning with atoms 6 and 10, are topologically identical. (As noted earlier, path generation stops when another path is encountered, any overlapping atom(s) being discarded. In this case atom 8 ends both paths and so it and its attached atoms appear in neither.) However the paths are not geometrically equivalent, in that a rotation about 2−4−5−6 will yield a geometry different from the equivalent rotation about 2−4−5−10. Therefore it is the last of the precedence rules outlined earlier that yields an unambiguous geometry, selecting 6 as the candidate that is locally to the right of the 2−4−5 plane (assuming atom 6 to be closer to the viewer than atom 10). The 2−4−5−6 dihedral value will thus be the one altered; however, because the 5−6 bond is in a ring, this value is adjusted to 90 degrees, not 180 degrees.

The other consequence of the 5−6 bond being in a ring is that the pucker state of that ring must be standardized. The ring system is found to include atoms 5 through 10 (since the 8−11 bond is not in a ring, the phenyl group is not part of this ring system). The ring pucker adjustment method will indicate that atoms 5 through 10 do not lie in the 4−5−10 plane, and so the dihedral angle 1−5−10-(nonplanarity-weighted-centroid of this ring system) is evaluated. If this value is greater than 180 degrees, the coordinates of all the remaining atoms 5 through 18 are reflected through the 4−5−10 plane.

Atom 8. Atom 8 is not on the (pro)chiral list. To establish the dihedral angle about bond 8−11, the precedence rules must choose between atoms 7 and 9 as the **a** atom and between atoms 12 and 16 as the **d** atom. Because the paths leading from the 7 and 9 atoms are topologically identical, the final precedence rule will be invoked, determining that it is atom 7 that is on the right of the 1−8−11 plane and so becomes **a**. The paths leading away from atoms 12 and 16 have the same numbers of atoms and the same molecular weights. However the sums of the atomic weights divided by the bond separations will not be equal (as a consequence of their topological difference), and so atom 16 will have higher precedence and become the **d** atom of the dihedral angle. The complete dihedral angle to be set is 7−8−11−16, and the value that its dihedral will take is 60 degrees, since both the 7−8 and 11−16 bonds are contained within rings.

Because the 11−16 bond is contained within a ring, the ring system including atoms 11 through 16 will be evaluated, found to be planar, and thereby require no pucker adjustment for standardization.

Topomer Similarity. This section combines an overview of the similarity comparison of two monovalent ("side chain") topomers with more detail on the similarity comparison of polyvalent ("core" or "scaffold") topomers. (Actually topomer similarities are calculated and expressed as dissimilarities.) An odd and completely general distinction of all topomer dissimilarities is their method of combination. Analogously to Euclidean distances between vectors, topomeric dissimilarities combine in geometric fashion, instead of the more intuitive arithmetic fashion. Thus two dissimilarity values A and B yield a combined dissimilarity value of $(A^2 + B^2)^{1/2}$ rather than A + B.

The dissimilarity between a pair of monovalent topomers has two contributions, steric fields and pharmacophoric features, as detailed elsewhere.[1,7] To summarize, the steric component combines the differences in the CoMFA-like steric fields exerted by the topomers over the intersections in a standard 2-Å spaced lattice. However, these steric fields are softer than those in CoMFA because the steric effects of individual atoms diminish with distance from the root atom, by being scaled according to the formula 0.85^n where n is essentially the number of intervening rotatable bonds. The pharmacophoric feature component is based on the usual designations of aromatic, positive, negative, accepting, and donating feature classes.[11] Topomer feature dissimilarity combines class-specific penalty values for feature instances that are either not duplicated in kind or are else too distant in spatial location. Furthermore feature dissimilarity is not symmetric, with the penalty for an unmatched feature in a candidate hit being much smaller than the penalty for an unmatched feature in the query.[7]

The overall shape of a complete molecule is far more affected by any central polyvalent fragments than by its monovalent fragments, and so major methodological extensions have been made to handle polyvalency, somewhat ad hoc although used for many years. These extensions have two major aspects, first some modifications to topomer generation, required for handling of shape/feature comparisons with multiple root atoms, and second inclusion of the relative locations of the attachment bonds, with their powerful leverage on the locations of monovalent fragments and overall molecular shape, in the dissimilarity calculation. (Note that the topomer dissimilarity between fragments having different numbers of valences is not defined.)

A polyvalent root produces as many topomers as there are open valences, one for each root atom, the other open valences being temporarily closed with methyl groups that then become a part of that particular topomer's structure. The only difference in individual topomer generation involves the precedence rules. Within a polyvalent fragment an additional rule is "inserted between" rules 1 and 2 within the hierarchy given earlier; namely, the shortest path that ends at one of the nonroot open valences takes precedence.

The topomer difference between two polyvalent fragments (having the same valency) has two complications. First there are the multiple possible mappings of the open valences onto each other, e.g., two for divalent fragments and six for trivalent fragments. Each of these potential dissimilarities ("symmetries") must be evaluated independently and tracked, while ensuring that the comparisons among sets of appropriately corresponding monovalent “side chains” are also evaluated, the final topomer dissimilarity being taken as the minimum dissimilarity from among these mappings. Second, for any mapping there are as many topomer dissimilarities to evaluate and combine as there are open valences. So to avoid an overemphasis on the core relative to the side chains, the sum of squared topomer dissimilarities over the *n* open valences in the core is then divided by *n* before being combined with other (squared) dissimilarities.

Also the dissimilarity of two polyvalent topomers is strongly increased by the relative position of the open valences, in two ways. First, the sum of six squared differences between the *x*, *y*-, and *z*-coordinates, for both the capping methyl carbon and the open valence atom to which the carbon is attached, becomes an additional dissimilarity. (With trivalent fragments having two capped open valences, this sum becomes the average over the two possible mappings.) This sum is scaled by 100.0 (default value) before being combined with the other squared dissimilarities. Second, the difference in angles between the vectors along the bonds from the open valence to the capping methyl carbon becomes another topomer dissimilarity, as the squared sine of the angle between the two vectors after the two open valence atoms have been made coincident, scaled by 10000.0 (default value) and within trivalent fragments averaging the contributions from the two open valences, as just described.[12]

Revalidation. Using the above rules for generating and comparing all the topomers, but otherwise exactly the same data and procedures as previously described, the two major published validation studies were repeated. To briefly summarize, the first of these compared the “neighborhood behavior” of topomers to that of several other descriptors, using “neighborhood plots” of *differences* between all pairs of biological potencies against *differences* between the correspondingly paired molecular descriptors. The more infrequent the observations that small differences in descriptor values are associated with large differences in biology, the better (more “valid”) the descriptor. The outcome of this study might have been affected both by the improved consistency in topomer generation and also by the intervening extension of topomer dissimilarity to consider feature as well as shape differences. The second major study examined whether topomer conformations might serve as automatic alignments for CoMFA, by comparing the statistical qualities of models based on topomer alignments with the original CoMFA models, over 15 trials. The revalidation study differs only in the methodological extensions in topomer generation. More experimental details can be found in the original publications.

RESULTS

The two completely code-independent implementations of the topomer generation protocol mentioned above were steadily refined, based on automatic comparisons of the topomer structures that each produced from various test sets. Toward the end of this protocol evolution, the test set was a large sublibrary of synthons constructed by removing a hydrogen from the nitrogen of all commercially available primary anilines, including 8489 structures, each containing at least one of the more demanding ring moieties. (Data sets with more structural variety at the topomer root were considered during earlier iterations.) Protocol evolution ended when only 65 of the 8489 structures (0.8%) yielded pairs of topomers differing by at least 40 units (equivalent to the steric shape difference between phenyl and pyridyl). Furthermore, among the 26 of those 65 discrepancies that were individually examined, none of the differences seemed to have originated within the adjustment protocol itself (22 of the 26 definitely being caused by various differences in the initial 3D models, and the four others apparently by issues elsewhere in the code).

Table 1. Neighborhood Distribution Enhancement Ratios for Various Implementations of the Topomer and Tanimoto 2D Fingerprint Descriptors

	topomeric		Tanimoto 2D fgpt	
	now	previous (steric)	whole (usual)	side chn only
Uehling	1.78	1.69	1.55	1.90
Strupczewski	1.46	1.39	1.41	1.73
Siddiqi	1.78	1.50	1.04	1.04
Garratt1	1.66	1.65	1.07	1.59
Garratt2	1.33	1.39	1.08	1.85
Heyl	1.12	1.04	1.01	1.71
Cristalli	1.50	1.40	1.31	1.76
Stevenson	1.45	1.06	1.07	1.08
Doherty	1.84	1.62	1.06	1.72
Penning	1.67	1.44	1.53	1.92
Lewis	1.42	1.05	1.01	1.63
Krystek	1.65	1.63	1.23	1.23
Yokoyama1	1.43	1.19	1.01	1.48
Yokoyama2	1.28	1.23	1.70	1.76
Svensson	1.36	1.26	1.02	1.68
Tsutsumi	1.35	1.38	1.58	1.74
Chang	1.46	1.33	1.13	1.68
Rosowsky	1.67	1.71	1.02	1.02
Thompson	1.63	1.47	1.17	1.70
Depreux	1.11	1.22	1.18	1.65
mean	1.50	1.38	1.21	1.59
standard dev	0.21	0.21	0.22	0.28
ratios > 1.1	20/20	17/20	10/20	17/20

Results from repeating the “neighborhood plot” validation of topomers are shown in the first column of Table 1. The other three columns repeat the comparable values from Table 2 of the original study,[4] the second column representing steric-only topomer differences, the third the conventional Tanimoto result, and the fourth the Tanimoto result if the underlying 2D fingerprints ignore the constant part of the structures. (Note that in the original study the topomer and Tanimoto were much the best descriptors among the eleven considered.) Each of the rows in Table 1 corresponds to a congeneric literature data set, further detailed and referenced within the original study. The values themselves are “neighborhood distribution enhancements”, ratios of the density of data points within an algorithmically defined “lower right trapezoid” region of the difference plot to the overall density of data points. Larger ratios indicate fewer deviations from the desirable tendency for similar structures to have similar biological effects, with ratios of 1.0 or lower indicating no such tendency existent and the largest value possible for such a ratio being 2.0.

Comparison of the values in the first and second columns of Table 1 shows that the topomer similarity enhancements, including the augmentation of steric shape by pharmacophoric features, have appreciably improved the original association between topomeric similarity and biological similarity. In particular, as shown in the last line of Table 1,

Table 2. Statistical Parameters of Model Derivation and the External Prediction Errors, for the 15 3D QSAR Literature Studies and Their Repetitions with Topomeric CoMFA, Generating Topomers by Either the Previous or the Current Protocol

		CoMFA model construction												CoMFA prediction			
data set	no. of	x-validated q^2			x-val SDEP			no. of compnts			final r^2			no. of	RMS pred error		
ID name	compds	Lit	TopB[e]	TopC[f]	Lit[b]	TopB	TopC	Lit	TpB	TopC	Lit	TopB	TopC	compds	Lit[a]	TopB	TopC
1 ICEc	36	0.630	0.362	0.418	0.816	1.002	0.913	6	5	2	0.970	0.883	0.788	9	0.568	0.740	1.210
2 ICEb	38	0.630	0.433	0.434	0.816	0.951	0.964	6	3	4	0.970	0.806	0.872	10	0.553	0.595	0.686
3 thrombin	72	0.687	0.533	0.519	0.594	0.726	0.736	4	4	4	0.881	0.838	0.842	16	0.673	0.619	0.596
4 trypsin	72	0.629	0.657	0.671	0.556	0.531	0.520	5	4	4	0.916	0.886	0.894	16	0.524	0.523	0.482
5 factorXa	72	0.374	0.186	0.159	0.515	0.591	0.592	3	4	2	0.680	0.747	0.529	16	0.278	0.340	0.413
6 MAOa	71	0.440	0.566	0.544	1.025	0.926	0.949	2	4	4	0.680	0.813	0.806				
7 MAOb	71	0.430	0.483	0.497	1.253	1.214	1.198	2	2	2	0.880	0.640	0.647				
8 hiv	25	0.680	0.389	0.703	0.571	0.845	0.619	3	3	5	0.950	0.878	0.984	7	0.823	0.449	0.090
9 a2a	78	0.541	0.226	0.220	0.563	0.742	0.740	4	3	2	0.817	0.555	0.446	23	0.668	0.761	0.712
10 d4	29	0.739	0.636	0.699	0.734	0.802	0.729	7	5	5	0.996	0.957	0.961				
11 flav	38	0.752	0.763	0.794	0.475	0.495	0.448	4	5	3	0.969	0.952	0.898	4	0.337	1.314	0.978
12 cannab	61	0.592	0.423	0.326	0.570	0.696	0.747	4	3	2	0.905	0.777	0.536	6	0.452	0.540	0.632
13 ACEest	41	0.937	0.746	0.755	0.346	0.726	0.728	4	3	2	0.990	0.916	0.867	7	0.413	0.478	0.473
14 5ht3	61	0.645	0.295	0.317	1.193	1.804	1.761	5	2	1	0.913	0.519	0.422				
15 rvtrans	82	0.837	0.830	0.841	0.567	0.587	0.574	4	4	6	0.936	0.916	0.947	19	0.791	0.608	0.496
total/av	847	0.636	0.502	0.526	0.581[c]	0.717[c]	0.689[c]	4.2	3.6	3.2	0.897	0.806	0.763	133	0.553	0.633	0.615
															0.574[d]	0.565[d]	0.579[d]

[a] For ICEc, ICEb, hiv, and a2a, the individual prediction values were read from the graphs in Figure 3, Figure 3, Figure 6, and Figure 3, respectively, of the original publications. Others were taken directly from tables. [b] For MAOa, MAOb, hiv, a2a, flac, cannab, and ACEest the SDEP was calculated from the original variance in biological activity and the reported q^2. Other values were taken directly from the tables. [c] Average excluding MAOa, MAOb, d4, and 5ht3 (to permit comparison with RMS CoMFA prediction error). [d] Average excluding flav (see text in original paper for discussion). [e] The TopB protocol represents "standard topomeric CoMFA" settings and the previous topomer generation protocol. [f] The TopC protocol represents "standard topomeric CoMFA" settings and the current topomer generation protocol.

every one of the neighborhood distribution enhancement values in the first column is greater than 1.1 (where 1.1 had previously been found to be the smallest ratio consistent with a statistically significant neighborhood enhancement). In the third column, the standard Tanimoto (using the whole structure to generate 2D fingerprints), though substantially better in the original study than any descriptor other than topomers, now provides only half the consistency (with 10 of 20 ratios > 1.1 rather than 20 of 20) and averaged performance (1.21, i.e., 21% better than "no association", rather than 1.50/50%) of the current topomer implementation ($p \gg 0.99$; $t = 5.9$). The last column of Table 1 repeats that an ad hoc definition of 2D fingerprints, applicable only within such congeneric series, also yielded excellent neighborhood behaviors, though now only comparable rather than superior to topomers ($p < 0.9$; $t = 1.6$). (It may be noted that this substantial improvement of column 4 over 3 seems to arise only from "bit-alias'ing", the forcing of many otherwise informative fingerprint bits to "1" by substructures within the larger constant fragments in these congeneric series.)

Turning to the revalidation of topomer CoMFA,[6] the various statistical metrics from this revalidation study appear within the five columns labeled "TopC" within Table 2. Each "TopC" column may be compared with two other columns in the same block, "Lit" providing results from one of the original papers using a conventional alignment protocol and "TopB" the results from CoMFA alignments using the previous topomer generation rules. Overall the average statistical quality resulting from more consistent topomer generation seems modestly better than that previously reported. The CoMFA models utilize a more concise 3.2 instead of 3.6 components to achieve an improved mean cross validated q^2 value of 0.526 rather than 0.502 and an improved mean SDEP value of 0.689 instead of 0.717. The drop in r^2 value to 0.763 from 0.806 seems simply a side effect of the fewer components. The average error of prediction for 133 compounds not included in any of the CoMFA derivations is also a bit better, 0.615 instead of 0.633. However these results are so similar as to reverse when a single questionable prediction is removed, the average error of prediction then increasing to 0.579 from 0.565. In summary, none of these observed slight improvements is statistically significant. Furthermore each of these slight overall improvements, particularly that in prediction, is mostly caused by substantial gains within just the one *hiv* data set.

DISCUSSION

Most methodological development in the 3D modeling of chemical structures has been aimed, quite understandably, toward greater physicochemical realism. However, that physicochemical reality is that biologically interesting molecules exhibit a formidable multiplicity of shapes and states, which in practical applications necessitates various summarizations so approximate as to perhaps be self-defeating. These considerations have led to the topomer methodology with its alternative goal of consistency rather than realism, in the positioning of similar structural features into similar regions of an arbitrary geometric space. Certainly the results so far achieved by application of this methodology (more of them unpublished than published) have encouraged its further development.

Although the desired self-consistency in geometry has been harder to achieve than was first appreciated, the essentially perfect agreement in structural output between two implementations of the current protocol encourages us to believe that the major difficulties have been successfully identified and addressed. Hopefully these experiences will be helpful to other workers who choose to pursue 3D self-consistency in other modeling applications.

Results from the repeated validation studies suggest that the protocol extensions needed to improve self-consistency have had a beneficial, though usually slight, effect on topomer applications. A substantial improvement was found in the "neighborhood behavior" of topomer similarity, although this is probably more attributable to the inclusion of pharmacophoric features than to the protocol extensions (recall that the original validation work considered only steric shape similarity). Nevertheless, the now unmistakable superiority in neighbor behavior of topomers compared to conventional Tanimoto 2D fingerprints is noteworthy, especially considering that Tanimoto 2D fingerprints were previously reported to have outperformed all other 3D similarity approaches.[13−15]

Despite the apparently modest effects of the topomer enhancements on these two validation outcomes, these enhancements have proven critical to success in other topomer applications. One reason for this apparent contradiction is that the validation structures did not include many of the (pro)chiral atoms and polycyclic or unsaturated ring systems that these enhancements directly address. A far more important reason is that the structure generation methods underlying the original validations had produced a fortuitous consistency in the numberings of atoms within the connection tables. Thus the many unresolved numbering-dependent degeneracies not addressed by the original topomer generation rules went undetected.

In summary, even though the particular shape displayed by a topomer may seem arbitrary, the utility of these shapes depends on their reproducibility. A protocol such as this seems mandatory to achieve the self-consistency and comparability required of topomeric shapes.

ACKNOWLEDGMENT

We thank Robert Clark, David Dunn, and Edmond Abrahamian for helpful discussions, particularly related to perceiving local right-handedness.

REFERENCES AND NOTES

(1) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069. This, the only previously published description of topomer generation, includes only torsion adjustment, because the need for control of prochirality and puckering, and of course their sequencing within the overall process, had not yet been recognized.

(2) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(3) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* **1998**, *6*, 1010−1023.

(4) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−60.

(5) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective Identification of Biologically Active Structures by Topomer Shape Similarity Searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(6) Cramer, R. D. Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. *J. Med. Chem.* **2003**, *46*, 374−389.

(7) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer Similarity Searching of Conventional Databases. *J. Mol. Graph. Modeling* **2002**, *20*, 447−462.

(8) Cramer; R. D.; Patterson; D. E.; Clark; R. D.; Ferguson; A. M. Method for selecting an optimally diverse library of small molecules based on validated molecular structural descriptors. U.S. Patent 6,185,506, 2001. Cramer; R. D.; Patterson, D. E. Further method of creating and rapidly searching a virtual library of potential molecules using validated molecular structural descriptors. U.S. Patent 6,240,374, 2001. Others pending.

(9) In practice, known stereocenters are treated no differently from unknown stereocenters, which seems paradoxical for a methodology addressing shape similarity. The reason is that topomeric similarity is always being assessed within a context where far more stereoforms are unknown than known. Whether a fragment structure is part of a query or a potentially matching fragment, because the far more numerous unassigned stereoforms will always have been standardized topomerically, any known stereocenter has a roughly 50% chance of being the nontopomeric stereoisomer. As a result, structurally identical fragments would be topomerically dissimilar and unrecognized 50% of the time. Faced with this very unattractive outcome, it was agreed that known stereocenters would be structurally registered but ignored in topomer modeling. Racemic fragments thus become two distinct registered "substances" mapping to the same topomer.

(10) "Two atoms or groups that upon replacement with a third group give enantiomers" are denoted as "enantiotopic". March, J. *Advanced Organic Chemistry, Reactions, Mechanisms, and Structure*, 4th ed.; John Wiley: NYC, 1991; p 135.

(11) For one set of structurally specific pharamcophoric point definitions: Abrahamian, E.; Fox, P. C.; Naerum, L.; Christensen, I. T.; Thogersen, H.; Clark, R. D. Efficient Generation, Storage and Manipulation of Fully Flexible Pharmacophore Multiplets and their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 458−468.

(12) These current treatments of the special dissimilarities between polyvalent topomers could be further improved. For example, surely the effect of differences in valence bond geometries on overall shape differences is better expressed as some function of the sizes of the side chains to be attached, rather than as a constant scaling factor.

(13) Martin, Y. C.; Kofron, J. L.; and Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activities? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(14) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases − a Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(15) Martin, Y, C.; Willett, P.; Lajiness, M.; Johnson, M.; Maggiora, G.; Martin, E.; Bures, M. G.; Gasteiger, J.; Cramer, R. D.; Pearlman, R. S.; Mason, J. S.; Diverse Viewpoints on Computational Aspects of Molecular Diversity. *J. Comb. Chem.* **2001**, *3*, 231−250.

CI049961D