J Comput Aided Mol Des (2010) 24:617–620
DOI 10.1007/s10822-010-9330-z

Tautomers and topomers: challenging the uncertainties of direct physicochemical modeling

Richard D. Cramer

Received: 3 February 2010 / Accepted: 10 March 2010 / Published online: 21 March 2010

Abstract To address the goal of improved discovery decision making, the uncertainties of physicochemical modeling, as exemplified by tautomer identification, are contrasted with methods focused exclusively on the sole experimental system variable, the changes in ligand structures, as exemplified by topomers.

Keywords Tautomers · Topomers · CoMFA · R-group virtual screening · Topomer CoMFA

Let us start a perhaps iconoclastic discourse by recalling the central process during drug discovery—the acquisition and testing of candidate molecules. How should computer-aided drug design activities benefit this process? Presumably we would generally agree with the response: “by better identifying those candidate molecules that are most likely to satisfy the therapeutic goals of the discovery program”.

The most important of these therapeutic goals, efficacy at a primary receptor [1, 2], is also the target that our CADD methodologies, training, and skills most strongly address. Of course mention might also be made of additional factors such as:

- To provide efficacy at that primary receptor, a satisfactory candidate molecule must also reach the receptor, a journey with dozens of possible detours and blind alleys [3, 4].
- To usefully affect the decision making within a discovery program, our CADD compound identification activities should be as timely, and therefore as rapid and convenient, as possible (the synthetic chemists will not remain idle).
- The easier that an identified structure will be to synthesize (the lower the cost to a discovery program of placing a synthetic “bet”), the more likely the compound is to be synthesized.
- The more candidate structures that can be constructively considered [5], the more novel (non-obvious and/or patentable) and the more highly valued any identified structure is likely to be.
- The closer that the candidates of a discovery program are to preclinical evaluation, the greater the potential impact and value of any success in identifying a superior structure (while recalling that during lead optimization improving efficacy at the primary receptor is seldom the major challenge).
- Performing more lengthy and expensive calculations would seem justifiable only by a reasonable expectation of an increase in the confidence of whatever decision making results [6].

Nevertheless efficacy at a primary receptor remains the focus of most of our CADD activities, typically involving (receptor) structure-based modeling. Successful modeling of course requires (among many other things) that the correct tautomer (for the receptor and probably also the ligand) has been identified. Unfortunately this requirement is much easier to state than to fulfill. For the ligand the most tautomer-determinant information, the relative pKa’s of its competing protonatable sites in dilute aqueous solution, may be entirely unknown. For both ligand and receptor, the preferred tautomeric states will depend,

R. D. Cramer (✉)
Tripos, Inc., 1699 South Hanley Road, St. Louis,
MO 63144, USA
e-mail: cramer@tripos.com

dynamically, on all the uncomfortable uncertainties of receptor-based CADD calculation. To recall only some of the more significant areas of uncertainty:

- State sampling (At bottom, the dynamic nature of any biological process unavoidably contrasts with the static character of the experimental receptor structure usually available for CADD calculations [7].)
- Potential energy functions (Electrostatics uncertainties, including local and global dielectric attenuations, and charge magnitudes and positions, have not diminished despite numerous capable attempts over many decades [8].)
- Effects of “3rd-species” intrinsic to any actual biological milieu (including any discontinuities in solvent/membrane environments, ionic strength and pH effects, and competitive binding, direct or allosteric, by endogeneous molecules) [9].

And of course tautomer identification, while surely under-appreciated, is only another among the many such uncertainties of receptor-based CADD calculations. An outsider would probably assume that a CADD specialist would want to evaluate any fundamentally different methodology, particularly if accompanied by substantial evidence of effectiveness in “better identifying those candidate molecules that are most likely to satisfy the therapeutic goals of the discovery program”.

Let us take a step back and recall that “better identifying those candidate molecules that are most likely to satisfy the therapeutic goals of the discovery program” usually requires predicting only the relative pIC50’s within a structure/activity table, not their absolute values. The single experimental factor that produces such potency variations is of course the structures of the candidate ligands themselves. True, the effect of any change in ligand structure on the observed biological potency may have been mediated by profound receptor effects, involving its static structure and/or its dynamic behavior. But the only cause of any change in the reproducibly observed potency of an administered ligand, no matter how mediated, is the known intentional change to its structure. Focusing exclusively on the effects of these intentional changes to ligand structure on observed potency could obviate the plaguing uncertainties in direct calculation of ligand-receptor interactions.

Such a focus may implicitly account for the strong “neighborhood behavior” [10] of similarity in such 2D descriptors as “Tanimoto fingerprints” [11, 12], and indeed of the as yet unexcelled effectiveness of “medicinal chemists’ intuition”, despite the self-evident truth that “receptors cannot read structural diagrams” [13]. Receptors do “read” molecular shape, however, and the recently introduced topomer methodologies explicitly and uniquely pursue shape consistency within the structurally invariant regions of candidate ligands, thereby focusing the analysis of their overall shape description differences and any SAR that results on their intentional structural differences.

A topomer is defined as a molecular fragment having a single internal geometry (conformation). The underlined words highlight the two means by which the topomer methodologies produce high shape consistencies. By definition, a molecular fragment possesses at least one open valence, so a topomer is oriented in space, or “rooted”, simply by superimposing its open valence onto a fixed Cartesian vector. The single conformation of a topomer is determined only by its own topology (“2D” structure), and not by either direct comparison with other structures or intramolecular energy. Yet the overall goal is that similar fragment topologies should afford similarly shaped topomers. Thus topomer valence geometries are generated by a 3D-model builder such as Concord, followed by canonically-determined adjustments to acyclic single bond torsions, stereochemistries, and ring “puckerings”. Then topomer CoMFA [14], a topomer-based derivation of (3D-Q)SAR, simply uses topomers from the fragmented training set as the 3D-QSAR-requisite inputs, and otherwise differs from “standard CoMFA” only in the use of multiple “CoMFA columns”, one for each set of fragments (aka “R-groups”) .

The strongest justification for any more empirical approach such as topomers is published evidence of effectiveness in “better identifying those candidate molecules that are most likely to satisfy the therapeutic goals of the discovery program”. Such support for topomer similarity as a superior predictor of the existence (not the magnitude) of similar biological properties currently consists of three retrospective studies [15–17], including a general hERG model [17], and two prospective applications [18, 19], one reporting success in thirteen of fifteen attempted “lead hops” [18]. For topomer CoMFA as a further predictor of biological property magnitudes, the superior accuracy of 3D-QSAR-based predictions generally is extensively documented [14, 20], and two retrospective studies [14, 21] and one prospective application [22, 23] show that the completely objective and remarkably facile topomer alignments perform comparably to the tedious and necessarily somewhat subjective alignment alternatives. Finally, although the number of additional yet-to-be-published topomer CoMFA applications is still modest, their cumulative success rates are not (so far, five unpublished topomer CoMFA leading to compound acquisition and testing have also been encountered by the author, from discovery programs in four different organizations. In all five cases, and thus quite remarkably, the encouraging topomer CoMFA predictions of biological activity were experimentally confirmed).

Thus the surprisingly general effectiveness of topomers as input alignments for 3D-QSAR can be understood as a consequence of how topomer shape differences focus on intentional ligand modifications. Yet that extreme emphasis on self-consistency [24, 25] may produce structures containing physicochemical absurdities. To illustrate this point with an extreme example, the rules that establish the torsional angle of a ring system attachment take no notice of any steric clashes with a bulky ortho group that may result. Energetically resolving such clashes would increase the torsional angle, and thus accentuate the shape differences between an ortho-substituted ring and other rings. Indeed any experimentally observed strong tendency for ortho substitution to particularly affect biological activity is usually so rationalized. Nevertheless, it was empirically found that on average such physicochemically absurd topomer shape behavior improves the prediction of biosimilarity; apparently a relatively greater certainty in detection of other shape dissimilarities is far more often influential when recognizing biological dissimilarity than is an otherwise overwhelming dissimilarity produced from "ortho bulkiness increasing torsional angles". Moreover, note that during topomer CoMFA, at least, any strong and consistent effect of ortho substitution on potency will still be captured, to appear within the resulting 3D contours enclosing the ortho group, colored either yellow (ortho substitution is unfavorable) or green (ortho substitution is favorable). However, any further interpretation of such yellow or green contours, such as "yellow suggests that the active conformation is flatter", would be far more speculative. The goal when performing topomer CoMFA is guidance useful for further structural improvements, not mechanistic inference.

Another important counter-intuitive behavior of the topomer generator is the standardization of each stereocenter, such that for example all hexose (monosaccharide) fragments produce identical topomers. The argument for this behavior is explicit, to ensure that shape similarities are not overlooked in the most common topomer application, searching for probable biosimilarity within large collections of structures. Otherwise the usual lack of explicit stereochemical assignments within such large collections, necessarily generating topomers with arbitrary configurations, means that every individual chiral center present in any candidate topomer will reduce its chances of being recognized as shape-similar by 50%. (For specialized situations this topomer standardization of stereocenters could easily be made suppressable.)

Tautomers too should be made as self-consistent as possible during topomer generation. (Albeit it must be confessed that until now the tautomer self-consistency issue had been overlooked during topomer methodology development.) For example, for both topomer CoMFA model generation and any topomer searching, all 2-pyridones and 2-hydroxypyridines should be converted to the same tautomer. Whether that tautomer is pyridone or hydroxypyridine will seldom matter, although each possibility may of course be tried. However, again please note that it is not expected that the "correct" tautomer will be identified by trying both. During topomer CoMFA, if there are actual differences in the predominant tautomeric state within a series, these differences must have been caused by the other intentional structural changes, just as with ortho substitution, with any consistent and substantial effects of tautomeric state differences on biological activities similarly being referred back to those originally causative changes by the resulting 3D-QSAR. And again, the only expected result is guidance useful in further structural improvement, not the structure of any preferred tautomer.

In current practice, the current topomer generator will automatically and insistently remove the charges associated with many common protomeric alternatives, for example converting every O[−] including COO[−] to OH and every NH[+] to N. Achieving further tautomeric self-consistency during topomer generation depends upon direct user control, by means of a preliminary and separate "ligand prep" process that includes substructure standardization. During standardization, each of a series of user-supplied substructural patterns (in SLN formats [26]) is repetitively sought within each processed structure until no match is obtained, with each matching pattern within the processed structure being converted into a second user-supplied pattern. Thus conversion of all 2-pyridones into 2-hydroxypyridines could be effected by supplying a matching pattern of "O=C[1]NHAny=Any=Any=Any@1" and a convert-into pattern of "HOC[1]:N:Any:Any:Any:Any:@1".

This entire discourse has omitted ligand-based SAR-type approaches, such as "classical CoMFA", that explicitly invoke tautomer ratios as a possible important cause of observed potency variations. It would be interesting to compare the results of such an approach with those of topomer CoMFA, in particular to confirm the admittedly unsupported belief that their guidance for structural improvements will not materially differ.

Returning to the comparison of the topomer approach with direct physicochemical modeling, it might be argued that "cancellation of errors" during physicochemical modeling of a series of structures may hope to provide a self-consistency benefit equivalent to that of topomers. However, the key words are "may hope to", as performing sufficient calculations on sufficient states of all the structures to assure such a cancellation is obviously even more computationally demanding than doing so for single structures. By comparison, topomer CoMFA offers robust and almost immediate guidance about a resulting model's overall trustworthiness, in the form of a q^2 value, and also about each of its predictions, by providing a topomer similarity score [25] as well as a predicted pIC50.

Of course the inherent scopes of direct physicochemical modeling and of the topomer methodologies differ. As ligand-based methodologies, the topomer approaches require at least one established ligand and depend primarily on shape similarity and successful extrapolation of structurally localized models to identify better candidates. Physicochemical modeling instead requires a known receptor structure, and therefore can suggest major changes to ligand shapes. Clearly these differing scopes tend to make the two approaches complementary. The solid theoretical basis of physicochemical modeling should ultimately make this approach superior (perhaps the advent of locally optimizable force fields [27] will help). However, CADD's current challenge is to best further today's discovery projects.

References

1. Hubbard RE (2006) Structure-based drug discovery: an overview. Royal Society of Chemistry
2. Lyne PD (2002) Drug Discovery Today 7:1047–1055
3. Gabrielson J, Weiner D (2007) Pharmacokinetic and pharmacodynamic data analysis: concepts and applications (4th edn)
4. Rydzewski RM (2008) Real world drug discovery. Elsevier, Amsterdam
5. Cramer RD, Soltanshahi F, Jilek R, Campbell B (2007) J Comp-Aided Drug Des 21:341–350
6. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912–5931
7. Grossfield A, Zuckerman D (2009) Ann Rep Comp Chem 5: 23–48
8. Amendola V, Boiochhi M, Fabbrizzi L, Palchetti A (2005) Chemistry 11:5648–5660
9. Weatherman RV, Fletterick RJ, Scanlan TS (1999) Ann Rev Biochem 68:559–581
10. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) J Med Chem 39:3049–3060
11. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350–4358
12. Brown RD, Martin YC (1996) J Chem Inf Comput Sci 36: 572–584
13. Cramer RD, Redl G, Berkoff CE (1974) J Med Chem 17:533–535
14. Cramer RD (2003) J Med Chem 46:374–389
15. Cramer RD, Jilek RJ, Andrews KM (2002) J Mol Graph Model 20:447–462
16. Cramer RD, Poss MA, Hermsmeier MA, Caulfield TJ, Kowala MC, Valentine MTJ (1999) Med Chem 42:3919–3933
17. Nisius B, Goeller AJ (2009) Chem Inf Model 49:247–256
18. Cramer RD, Jilek RJ, Guessregen S, Clark SJ, Wendt B, Clark RDJ (2004) Med Chem 47:6777–6791
19. Tresadern G, Bemporad D, Howe TJ (2009) Mol Graph Model 27:860–870
20. Doweyko AJ (2004) Comp Aided Mol Des 18:587–596
21. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi FJ (2008) Chem Inf Model 48:2180–2195
22. Wendt B, Uhrig U, Wang L (2009) ACS Abstracts 237, COMP 209
23. http://www.qsar2008.org/home/FA04-10-12-42_h6vpw99c3zxmfq28f4e9/qsar2008.org/public_html/File/Poster%20abstracts/Uhrig_Ulrike_Euroqsar2008.pdf
24. Cramer RD, Clark RD, Patterson DE, Ferguson AMJ (1996) Med Chem 39:3060–3069
25. Jilek RJ, Cramer RDJ (2004) J Chem Inf Comp Sci 44:1221–1227
26. Ash S, Cline MA, Homer RW, Hurst T, Smith GB (1997) J Chem Inf Comp Sci 37:71–79
27. Pham TA, Jain AN (2008) J Comp-Aided Mol Des 22:269–286