DOI: 10.1002/cmdc.200600161

Asymmetry of Chemical Similarity**

Xin Chen* and Frank K. Brown[a]

The concept of chemical similarity plays an important role in modern medicinal chemistry. Based on the similar property principle,[1] which states that compounds with similar chemical structures should possess similar physicochemical properties and biological activities, chemical similarity calculations have been conducted in many applications of computer-assisted drug discovery such as chemical database searches,[2] molecular diversity analysis,[3] and QSAR.[4] Interestingly, almost all the chemical similarity measures adopted so far are symmetric and do not depend on the order of comparison. In other words, the similarity value of compound *a* to compound *b* is always assumed to be equal to that of compound *b* to compound *a*. However, this assumption may be overstated and unnecessary.[5] Herein we present the first evidence of the presence of asymmetry in chemical similarity measures by an empirical study of two large pharmaceutical databases. The implications of the new findings on the practical use of chemical similarity searches in lead identification are discussed as well.

Although chemical similarity can be measured in numerous ways, the most popular usually involve two major components: structural descriptors, which represent chemical structures in a certain numerical way so that they can be easily compared, and similarity coefficients, which provide a mathematical formula for calculating similarity values based on the numerical values of all the structural descriptors.

Three popular structural descriptor sets, MACCS keys,[6] Daylight fingerprints,[7] and atom pairs,[8] were used in this study to avoid potential bias on one particular descriptor set. They are all topological fragment-based descriptors and have been implemented in the commercial or corporate chemical database search systems. The choice of these three descriptor sets also reflects our real interest in studying similarity measures in the context of searching large pharmaceutical databases.

The similarity coefficients used in this study are Tversky coefficients.[9] They can be defined by the set-theoretic approach as:

$$s(a,b) = \frac{(A \cap B)}{\alpha(A-B) + (1-\alpha)(B-A) + (A \cap B)} \quad (1)$$

Here, $s(a,b)$ denotes the similarity value of compound *a* to compound *b*. As illustrated in Figure 1, *A* and *B* represent the sets of structural descriptors for the two compared compounds *a* and *b*, respectively. $(A-B)$ and $(B-A)$ stand for the sets of structural descriptors that are unique to compounds *a* and *b*, respectively. $(A \cap B)$ represents the set of structural descriptors that are common to both compounds. α is an adjustable parameter within the range of [0,1] that controls the relative contributions from the two compared compounds. In fact, Tversky coefficients represent a class of association coefficients with adjustable relative weights on the two compared compounds. They offer the flexibility to model complicated similarity relationships beyond the simple metric assumption, which implies minimality,[10] symmetry,[11] and triangle inequality.[12] In this study, we focus only on using Tversky coefficients to study the symmetry–asymmetry characteristics of chemical similarity measures. Specifically, when $\alpha = 0.5$, Equation (1) becomes the more popular Dice coefficient, which is symmetric and also monotonic with another highly popular Tanimoto coefficient. When α is not equal to 0.5, $s(a,b)$ becomes different from $s(b,a)$, and therefore, the Tversky coefficient becomes asymmetric.

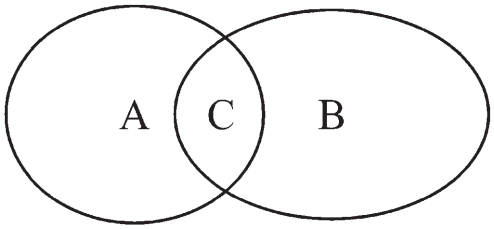

Figure 1. Illustration for the set-theoretic definition of Tversky coefficients.

Two large pharmaceutical databases, the NCI anti-AIDS database[13] and the J&J corporate database,[14] were used as the test data sets for this empirical study. After certain data cleaning, the former contains 37 942 compounds with 1067 being actives, and the latter contains 983 407 compounds including 70 833 actives belonging to 11 selected activity classes (Table 1).

Simulated similarity searches were conducted to investigate the performance of Tversky coefficients with different α values, that is, different degrees of asymmetry. Each active compound was used as the probe to rank the rest of the database (target

Table 1. Activity classes selected for study in the J&J corporate database.

No.	Activity Class	Target Family	No. of Actives
1	ZAP70 inhibitor	kinase	2871
2	VEGF receptor inhibitor	kinase	2830
3	JAK3 inhibitor	kinase	3407
4	ABL inhibitor	kinase	2092
5	5-HT_{2C} receptor inhibitor	GPCR	8183
6	H1 receptor inhibitor	GPCR	7884
7	opiate receptor μ inhibitor	GPCR	6102
8	MMP9 inhibitor	protease	2097
9	PTP1B inhibitor	phosphatase	23 990
10	NE transporter inhibitor	transporter	3368
11	hERG channel inhibitor	ion channel	8009

[a] *Dr. X. Chen, Dr. F. K. Brown*
Computer-Assisted Drug Discovery
Johnson & Johnson Pharmaceutical Research and Development, L.L.C.
920 Route 202, Raritan, NJ 08869 (USA)
Fax: (+01) 215-628-5047
E-mail: xin69chen@yahoo.com

[**] *We acknowledge Dr. Yuting Tang for helpful discussions about the biology data in the J&J database and Dr. Yves Wetzels and Dr. Christophe Buyck for generating Daylight fingerprints.*

compounds) according to their similarity values *s*(*probe*,*target*) calculated by Equation (1), from the most to least similar. The α parameter in Equation (1) was systematically adjusted from 0 to 1 with a small increment of 0.01. As a result, the average hit rate with the change of both the α value and the number of the nearest neighbors is demonstrated in Figures 2 and 3 for the NCI anti-AIDS and J&J corporate databases, respectively. The hit rate was calculated as *n*/*m*, for *n* active compounds present among the top *m* nearest neighbors. This is depicted in the color scale, with red representing a high hit rate and the blue representing low.

In Figures 2 and 3, all the hit-rate maps clearly show an asymmetric color pattern with the red region inclined to the right side, indicating that greater α values are generally more favored for higher hit rates. In other words, Tversky measures with a higher weight on the probe compound generally outperform those with a higher weight on the target compound in terms of the ability to retrieve active analogues. This can serve as evidence of the presence of asymmetry in chemical similarity measures. The observed asymmetry can be interpreted by the directionality or inequality that is inherent in a chemical similarity search: the probe compounds used in practice are always active, whereas the target compounds can be either active or inactive so that their unique structural descriptors may not be as important as those of the probe compounds regarding their contribution to biological activity.

Figures 2 and 3 also clearly indicate that the degree of asymmetry is positively correlated to the degree of similarity itself, in general. In other words, the red region becomes ever more inclined to the right side with an increasing number of nearest neighbors. Therefore, to retrieve more remotely similar active compounds, a more asymmetric similarity measure should be used, that is, more weight should be put on the probe compound. On the other hand, to retrieve more highly similar active compounds, approximately equal weights should be put on both the probe and target compounds, leading to a measure close to symmetric.

An additional observation from Figures 2 and 3 is that the maximum hit rates generally appear between α values of 0.5 and 1.0, not at $\alpha = 1.0$ as claimed by Blankley and Wild.[5c] This indicates that a weight scheme totally biased on probe compounds usually will not lead to optimal performance. Despite the diminished importance of the unique structural descriptors of target compounds, they do have the potential to positively contribute to biological activity. Totally ignoring their contribution may lead to suboptimal performance.

The implication of the findings presented herein is that the relative weights on probe and target compounds can be adjusted to more effectively achieve different purposes of a similarity search. One goal we frequently pursue in drug discovery is to find compounds that are highly similar to the known actives. For example, these actives may come from HTS screening or serendipity, and we want to quickly identify and test their analogues to confirm the series and hopefully find some preliminary SARs. Then, highly symmetric similarity measures should be the choice for this purpose, based on the results of this study. Another goal we often seek is to identify remotely similar analogues. For example, when we follow the lead structures reported by our competitors, we may wish to identify some new structures that are similar to these lead structures so that they will have better chances of being active as well, yet not so similar that they fall under protection of the competitors' patents. In this case, we now know highly asymmetric similarity measures should be adopted.

Asymmetry can be easily introduced into many other popular similarity measures, such as that recently proposed by Flinger et al.[15] We expect to see more studies on this concept and more applications of the related techniques in computer-assisted drug discovery in the future.

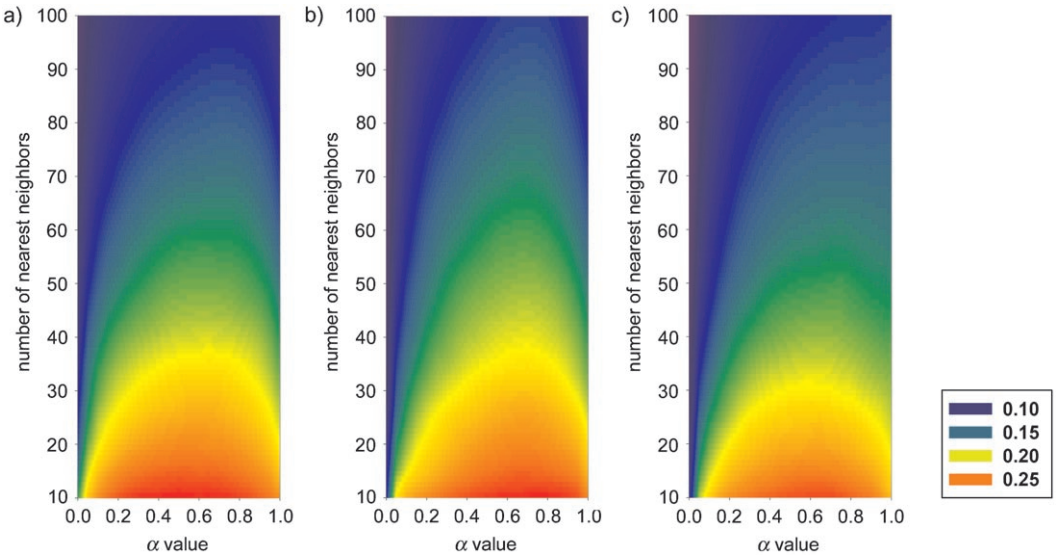

Figure 2. Hit-rate maps for the NCI anti-AIDS database using a) atom pairs, b) Daylight fingerprints, and c) MACCS keys as structural descriptors. Average hit rate is represented in the color scale shown.

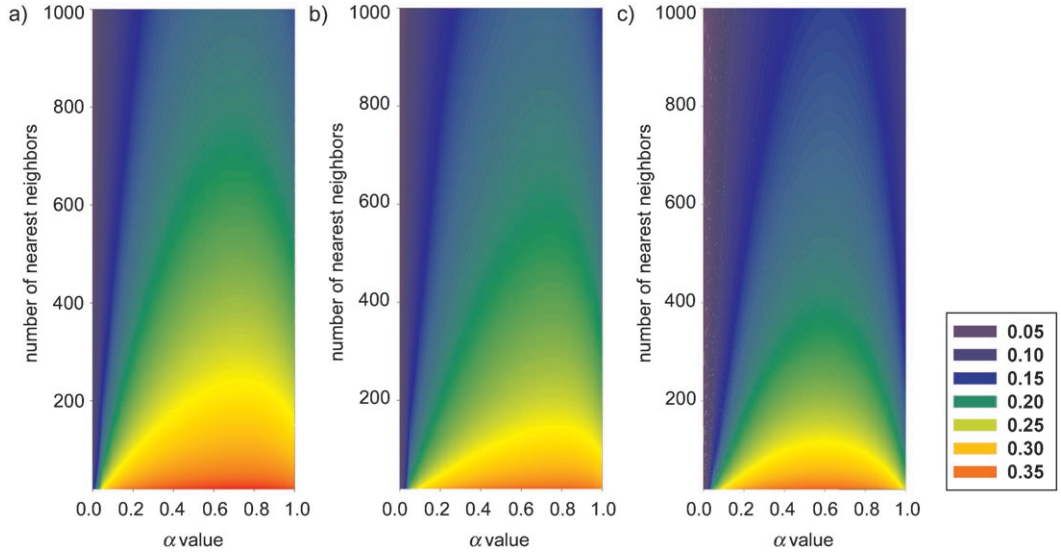

Figure 3. Hit-rate maps for the J&J corporate database using a) atom pairs, b) Daylight fingerprints, and c) MACCS keys as structural descriptors. Average hit rate is represented in the color scale shown.

Experimental Section

The MACCS keys, Daylight fingerprints, and atom pairs were generated by MOE,[16] Daylight toolkits,[17] and an in-house C++ program based on the report by Carhart et al.,[8] respectively. Both the databases were subject to data cleaning to remove compounds with molecular weights outside the range of 60–600 Da, resulting in more druglike databases. As a result, the NCI anti-AIDS database contains 251 conformed actives, 816 confirmed moderate actives, and 36875 confirmed inactives. Both the confirmed actives and confirmed moderate actives were treated equally as actives and the conformed inactives as inactives in this study. Table 1 lists 11 activity classes selected for study from the J&J corporate database. A cutoff of 10 μM was arbitrarily selected so that all the compounds with an IC_{50} or K_i value below 10 μM were treated as actives, while all the compounds with IC_{50} or K_i values above 10 μM and all the untested compounds were treated as inactives. All the simulated similarity searches were done by an in-house C++ program.

Keywords: asymmetry · chemical similarity · cheminformatics · Tversky coefficients · virtual screening

[1] G. M. Maggiora, M. A. Johnson, *Concepts and Applications of Molecular Similarity*, Wiley, New York, **1990**.
[2] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.
[3] Y. C. Martin, *J. Comb. Chem.* **2001**, *3*, 231 – 250.
[4] A. C. Good, W. G. Richards, *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 321 – 338.
[5] a) J. Bradshaw, "Introduction to Tversky Similarity Measure", 11th Daylight user group meeting, can be found under http://www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html, **1997**; b) G. M. Maggiora, J. Mestres, T. R. Hagadone, M. S. Lajiness, "Asymmetric Similarity and Molecular Diversity", 213th ACS National Meeting, San Francisco, April 13 – 17, **1997**; c) C. J. Blankley, D. J. Wild, "Asymmetric Similarity in Action", 221st ACS National Meeting, San Diego, April 1 – 5, **2001**; d) J. D. MacCuish, N. E. MacCuish, "Asymmetric Clustering of Chemical Datasets: An Investigation", 224th ACS National Meeting, Boston, August 18 – 22, **2002**.
[6] MACCS II Manual, MDL Information Systems, Inc. San Leandro, CA (USA).
[7] Daylight Theory Manual, Daylight Information Systems, Inc. Santa Fe, NM (USA).
[8] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64 – 73.
[9] A. Tversky, *Psycol. Rev.* **1977**, *84*, 327 – 352.
[10] Minimality is described as $s(a,b) \geq s(a,a) = 0$.
[11] Symmetry is described as $s(a,b) = s(b,a)$.
[12] Triangle inequality is described as $s(a,b) + s(b,c) \geq s(a,c)$.
[13] The NCI anti-AIDS database contains the structure–activity data for the compounds screened by the National Cancer Institute AIDS antiviral screening program. It can be downloaded from the public web site: http://dtp.nci.nih.gov/docs/aids/aids_data.html.
[14] The J&J corporate database contains all the chemical structures registered at Johnson & Johnson Pharmaceutical Research and Development along with their biological data. It is a private database.
[15] M. A. Flinger, J. S. Verducci, P. E. Blower, *Technometrics* **2002**, *44*, 110 – 119.
[16] MOE, Chemical Computing Group, Inc. Montreal (Canada).
[17] Daylight toolkit, Daylight Information Systems, Inc. Santa Fe, NM (USA).

Received: June 28, 2006
Revised: August 29, 2006
Published online on December 19, 2006

