DOI: 10.1002/cmdc.200700050

Apparent Asymmetry in Fingerprint Similarity Searching is a Direct Consequence of Differences in Bit Densities and Molecular Size

Yuan Wang, Hanna Eckert, and Jürgen Bajorath*[a]

Recently, systematic similarity calculations using Tversky coefficients have suggested that putting higher weight on bit settings of active reference molecules (templates) than database compounds increases hit rates in similarity searching using 2D fingerprints. These findings have been interpreted as evidence for "asymmetry" in chemical similarity searching. We have thoroughly analyzed this phenomenon and demonstrate that apparent asymmetry in similarity search calculations is a direct consequence of differences in fingerprint bit densities, which often correlate with differences in molecular size. Accordingly, a size-independent fingerprint with constant bit density does not produce asymmetrical search results. For Tversky similarity calculations, differences in fingerprint bit densities between active and inactive compounds determine which weighting factors produce high hit rates.

Introduction

Fingerprints are bit string representations of molecular structures and properties used for similarity searching.[1,2] These search calculations are conceptually based on the similar property principle: similar molecules are thought to have similar biological activity.[3] In similarity searching, known active compounds are used as templates to search databases for novel hits. In this context, the evaluation of molecular similarity critically depends on the application of similarity measures for quantitative bit string comparison.[1] A variety of similarity metrics are being used for this purpose including the popular Tanimoto coefficient[1] and the Tversky coefficients.[4] As further described below, a unique feature of Tversky coefficients is the ability to put variable weights on the bit settings of molecules that are compared. By contrast, most similarity measures put equal weight on template and database compounds. Thus, these measures are symmetrical in nature, which means that the results of pairwise molecular comparisons are order-independent. Principal and statistical limitations associated with the use of similarity coefficients have been noted previously[5,6] and an elaborate analysis of different similarity measures and their strengths and weaknesses has been presented.[7]

In a recent communication in this journal, Chen and Brown have investigated the behavior of Tversky coefficients in large-scale similarity search calculations using three different 2D fingerprints and found that putting increasingly high weight on the bit string representations of template compounds produced higher hit rates than calculations using a symmetrical coefficient with equal weights on template and database compounds.[8] Chen and Brown interpreted their findings as "the first evidence of the presence of asymmetry in chemical similarity measures by an empirical study of two large databases".[8] The study by Chen and Brown represents an important advance because it highlights possible complications of molecular similarity assessment that are often not appreciated and enables further analyses of the observed effects.

We have explored potential reasons for these interesting observations concerning Tversky similarity calculations and present the results of our analysis herein.

Results and Discussion

We begin our analysis with principal considerations about Tversky coefficients, a class of similarity coefficients with adjustable relative weights. For two molecules being compared and represented by fingerprint bit strings *A* and *B*, Tversky coefficients (*Tv*) are defined as follows:

$$Tv(A,B,\alpha) = \frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c} \quad \text{with } \alpha \text{ in } [0,1] \tag{1}$$

Here, *a* represents the number of bits set on in *A*, *b* the number of bits set on in *B*, and *c* the number of bits set to 1 in both bit strings. The α parameter varies between zero and one and determines the relative weight of uniquely set bits. For $\alpha = 0.5$ equal weights are put on both molecules (and the

[a] *Y. Wang,[+] H. Eckert,[+] Prof. Dr. J. Bajorath*
Department of Life Science Informatics
Bonn-Aachen International Center for Information Technology
Rheinische Friedrich-Wilhelms-Universität Bonn
Dahlmannstr. 2, 53113 Bonn (Germany)
Fax: (+49) 228-2699-341
E-mail: bajorath@bit.uni-bonn.de

[+] *These authors share first authorship.*

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

Tversky coefficient becomes the symmetrical Dice coefficient[1]), whereas for $\alpha > 0.5$ or $\alpha < 0.5$ more weight is put on bits that are exclusively set on in *A* and *B*, respectively. If molecules *A* and *B* are compared and their bit string representations have exactly the same number of bits set on, Tversky coefficients are symmetrical, which means that comparing *A* with *B* and *B* with *A* produces the same value. If the bit densities of *A* and *B* differ, the comparison becomes order-dependent for $\alpha \neq 0.5$ and the corresponding Tversky coefficients are asymmetrical.

On the basis of its formula, we determine Tversky similarities from relative differences in bit settings generated by a substructure-encoding fingerprint for hypothetical molecules *A*, B_1, B_2, and B_3 under systematic variation of α. The corresponding bit numbers are *a*, b_1, b_2, and b_3, respectively. Characteristic features of Tversky similarity can be best rationalized when studying examples that produce large variations in similarity values. We found this to be the case when comparing a test molecule with a sub- and superstructure and, in addition, another molecule having the same fingerprint bit density. In our example, molecule *A* sets 50 of 100 hypothetical fingerprint bits to one. Molecule B_1 is a substructure of *A* having 25 fewer bits set on, B_2 is another molecule that—like *A*—has also 50 bits set on but only 37 in common with it, and B_3 is a superstructure of *A* having 25 more bits set to one. Comparison of *A* and B_1 leads to a similarity value of 1.0 for $\alpha = 0$, comparison of *A* and B_2 to 0.74 for all α values, and *A* and B_3 to 1.0 for $\alpha = 1$. Thus, for α values approaching zero or one Tversky similarity calculations become akin to substructure searching. For α values close to one, compounds achieve high *Tv* values if they contain the query molecule as a substructure. In contrast, for α values approaching zero, compounds obtain high *Tv* values if they themselves are substructures of the query.

Figure 1 shows the similarity curves for comparisons of *A* with B_1, B_2, and B_3, respectively. With the exception of the *A* versus B_2 comparison, convex curves are obtained whose gradients strongly depend on the differences between *a* and b_i. Assuming $c \neq 0$, for $a > b_1$ *Tv* values are monotonously decreasing and for $a < b_3$ they are monotonously increasing. Figure 1 also shows the difference in similarity values for comparison of molecules *A* with B_1 and B_3, respectively, when α is set to 0.5 and *Tv* becomes a symmetrical coefficient. This reflects a general bit density-dependence of the Tversky similarity measure.

We go a step further and evaluate potential consequences of these general *Tv* characteristics for similarity searching. In addition to differences in specific bit settings, overall differences in bit densities also lead to a separation of molecules depending on α parameter values. For example, if active molecules have a comparable bit density but on average a higher bit density than inactive ones, the $a > b_1$ case applies for the comparison of active molecules against inactive molecules. As a consequence, if we increase α, similarity values decrease for inactive database molecules but are mostly unaffected for active molecules (case $a = b_2$, see Figure 1) leading to a deselection of inactive compounds. By contrast, if bit strings of active molecules have similar bit density but systematically lower bit densities than inactive molecules, the $a < b_3$ case applies and, according to Figure 1, lowering α will lead to a deselection of inactive molecules.

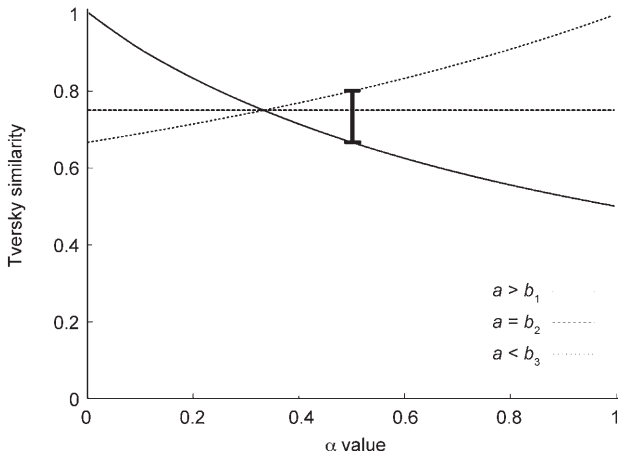

Figure 1. Tversky coefficient. Reported are Tversky similarity values, $Tv(A,B_i, \alpha)$, for a template molecule *A* against three different database compounds B_i (or hypothetical fingerprints with *a* and b_i bits set to one, respectively) as a function of the weighting parameter α. We study three cases: $a > b_1$ (fewer bits are set on in B_1 than in *A*), $a = b_2$ (the same number of bits are set on in both compounds), and $a < b_3$ (more bits are set on in B_3). The differences $a - b_1$ and $b_3 - a$ are set to be equal. The black bar marks the difference in the two similarity values of B_1 and B_3 for $\alpha = 0.5$ (symmetrical Tversky coefficient). The difference reflects the general bit density dependence of the Tversky similarity measure.

Figure 1 also reveals another general characteristic of the Tversky coefficient. As discussed above, in its symmetrical version ($\alpha = 0.5$), it assigns higher similarity values to molecules that have more bits set on than to molecules with fewer bits, even if their distance to an active reference compound *A* is the same in "bit string space". For example, as mentioned above, molecules B_1 and B_3 both deviate in exactly 25 bit positions from *A*. However, comparison of *A* and B_3 results in a significantly higher similarity value than the comparison of *A* and B_1. This is due to the fact that calculation of Tversky similarity takes only bits set on (that is, to "1") into account, which also applies to the Tanimoto and other coefficients often used in similarity searching.[1] This property is often referred to as the size effect[5] because larger molecules tend to set more bits on than smaller ones and thus often achieve higher similarity values.

Molecular complexity determines fingerprint bit density and usually correlates with molecular size. Exceptions include, for example, polymers where fragment-based fingerprints would only account for the presence of a monomer, but not the occurrence of multiple copies. However, we can generally assume that notable differences in molecular size are reflected by corresponding differences in bit density. As can be seen in Figure 1, the size effect referred to above and the corresponding differences in bit density affect symmetrical *Tv* calculations because when compared to *A*, molecule B_1 produces a lower similarity value than B_3. However, under variation of the **α** parameter, when Tversky coefficients become asymmetrical, there is an additional effect. For $\alpha > 0.5$, *Tv* values for comparison of a reference molecule with a larger compound further increase,

Table 1. Compound set characteristics.[a]

Code	Designation	numCpds	numHA	MACCS 1-Bits	TGD 1-Bits	PDR-FP 1-Bits
BEN	Benzodiazepine Agonists	57	25.6 (4.4)	51.1 (8.0)	56.2 (15.3)	93.0 (0.0)
CAT	Cathepsin Inhibitors	90	32.3 (7.9)	50.2 (12.4)	87.5 (29.1)	93.0 (0.0)
HH2	Histamine H2 Antagonists	41	27.6 (6.9)	55.6 (11.5)	91.4 (22.4)	93.0 (0.0)
NNI	Neuronal Injury Inhibitors	50	14.0 (1.8)	33.7 (9.6)	25.3 (9.4)	93.0 (0.0)
TNF	TNF-alpha Release Inhibitors	65	31.0 (8.2)	52.7 (12.9)	82.6 (28.4)	93.0 (0.0)
NCI	NCI Anti-AIDS database	42687	25.2 (12.1)	42.7 (13.6)	55.3 (32.9)	93.0 (0.0)

[a] Five activity classes and the NCI database were used for our statistical analysis. Reported are the number of compounds (numCpds), average number of nonhydrogen (or heavy) atoms (numHA), and average bit settings for three different 2D fingerprints: "MACCS 1-Bits", "TGD 1-Bits", and "PDR-FP 1-Bits" stand for average number of bits that are set on (to one) in these fingerprints for the different compound sets. Standard deviations for the different values are given in parenthesis.

whereas with a smaller compound the corresponding values further decrease. For $\alpha < 0.5$, these effects are reversed. Thus, on the basis of these considerations, molecular size effects are thought to systematically affect calculations of Tversky similarity.

To complement our theoretical considerations, we next performed test calculations on different compound data sets that are summarized in Table 1. For five activity classes and the NCI background database, we calculated the average number of nonhydrogen atoms as a measure of molecular size. We also determined for each compound set the average number of bits set on in three different fingerprints; MACCS, TGD, and PDR-FP (see Experimental Procedures). The results are reported in Table 1. For our activity classes, average numbers of nonhydrogen atoms ranged from 14.0 to 32.3 and for the NCI database, the average number was 25.2. Activity class NNI was assembled to consist of on average much smaller molecules than the other classes and had significantly lower bit density for MACCS and TGD. For PDR-FP, bit densities did not vary because this fingerprint was designed to have a constant number of bits set on independent of molecular complexity and size.[9]

We then calculated pairwise Tversky similarities for compounds within each activity class and also between activity classes and NCI compounds, both under systematic variation of α parameter values. The results are shown in Figure 2. For MACCS and TGD, average similarity values within each activity class formed symmetrical curves with a minimum at $\alpha = 0.5$. This is the case because for each pair of active molecules A_1 and A_2, both values $Tv(A_1,A_2, a)$ and $Tv(A_2,A_1, a)$ contribute to the overall average value. By contrast, average *Tv* values for activity classes against NCI compounds did not follow symmetrical curves but were monotonously decreasing for classes BEN, CAT, HH2, and TNF but monotonously increasing for NNI. For the three classes with fingerprint bit densities higher than NCI, standard deviations of bit densities were very similar (Table 1). As expected, for NNI, standard deviations were overall smallest (Table 1). These results were perfectly in accord with our expectations. As average bit densities were lower for NCI than BEN, CAT, HH2, and TNF compounds (Table 1), similarity values decreased for increasing α values and NCI molecules were selected, which corresponds to the $a > b_3$ case in Figure 1. In contrast, NNI had a lower average bit density than NCI leading to increasing similarity values when α increased and preferential selection of NCI compounds, which corresponds to the $a < b_1$ case in Figure 1. As can be seen in Figure 2b, by far the smallest differences between similarity values for variation of α were observed for BEN relative to the NCI database when using the TGD fingerprint. This was a consequence of the fact that BEN and NCI compounds produced nearly the same average bit density (56.2 versus 55.3, Table 1).

For PDR-FP, average similarities formed no monotonously increasing or decreasing curves but horizontal lines. This was because PDR-FP has consistently 93 bits set on for each molecule and, therefore, *Tv* becomes completely independent of the α parameter. This is obvious if we transform/reduce the Tversky formula accordingly:

$$\begin{aligned} & Tv(A,B,\alpha) = \frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c} \\ \Leftrightarrow\ & Tv(A,B,\alpha) = \frac{c}{\alpha(a-b)+b} \\ \xrightarrow{a=b}\ & Tv(A,B,\alpha) = \frac{c}{b} \end{aligned} \quad (2)$$

As can be seen in Figure 2, when average similarity values were calculated, maximal differences and lowest similarity values between activity classes and NCI compounds for fingerprints MACCS and TGD were achieved for $\alpha = 1$ (BEN, CAT, HH2, TNF) or $\alpha = 0$ (NNI).

In similarity searching, hit rates depend on differences between the distributions of 1) pairwise intraclass similarity values and 2) similarity values for active versus database compounds. As an example, distributions for activity class HH2 (intraclass) and HH2 versus NCI (interclass) are shown in Figure 3. Until now, we have only considered average similarity values. However, for the comparison of similarity value distributions, we also need to take standard deviations into account. First, the larger the difference between average similarity values is, the further the distributions are apart. Second, the smaller the standard deviations are, the narrower the distributions become. Both effects minimize the overlap and increase hit rates. In light of its relevance, we have defined a simple measure that approximates the overlap of two similarity distributions (see Figure 3). Given two distributions of intraclass similarities (*AC*) and similarities between active and database molecules (*DB*), we define the overlap (*OV*) as:

$$OV = (\mu_{DB} + \sigma_{DB}) - (\mu_{AC} - \sigma_{AC}). \quad (3)$$

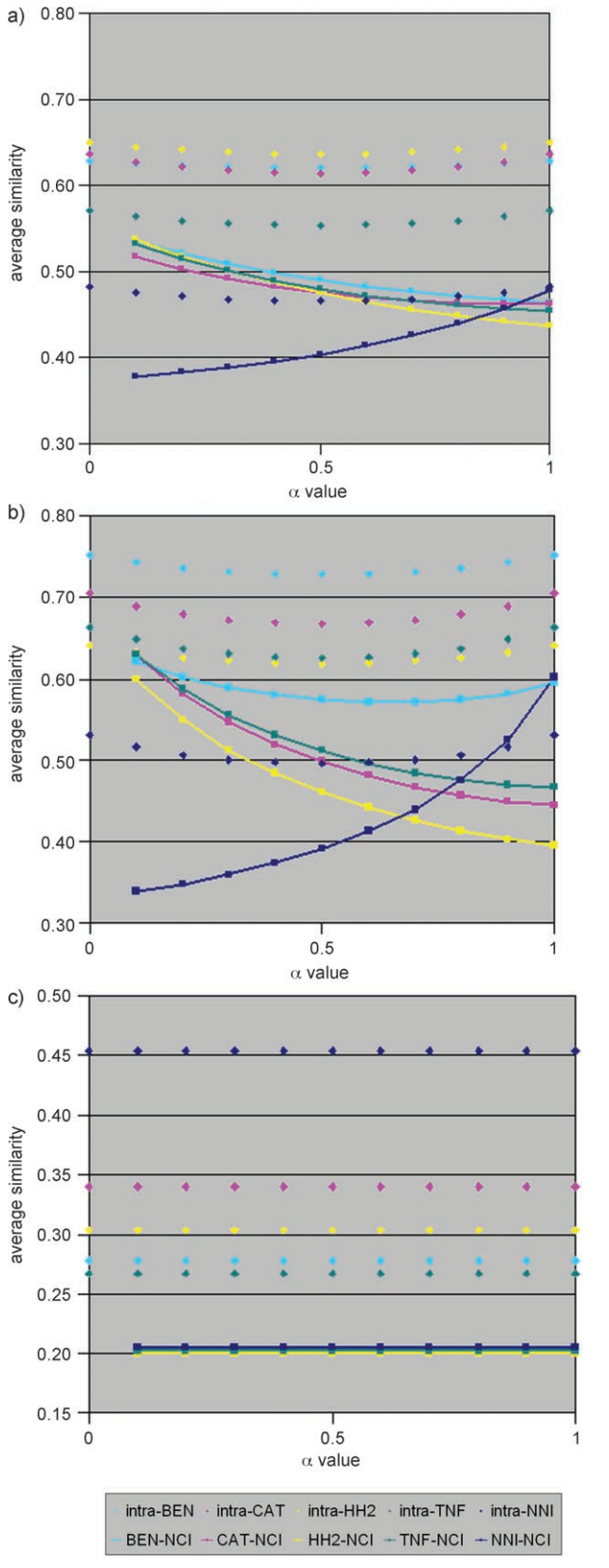

Figure 2. Statistical properties of Tversky similarity. For a) MACCS, b) TGD, and c) PDR-FP, average pairwise Tversky similarities were determined as a function of the α parameter of the Tversky coefficient within each activity class (intraclass similarity) and between activity classes and the NCI database (interclass similarity). Classes are designated according to Table 1.

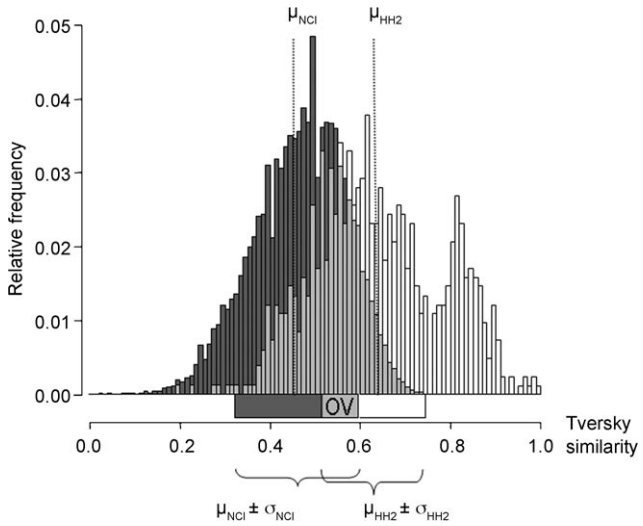

Figure 3. Overlap of Tversky similarity distributions. Value distributions for pairwise Tversky similarities ($\alpha = 0.5$) within activity class HH2 (white) and between HH2 and the NCI database (dark gray) are shown. The position of the average value (μ_{HH2} or μ_{NCI}) for each distribution is indicated by a dotted line. The intervals $[\mu_{NCI} \pm \sigma_{NCI}]$ and $[\mu_{HH2} \pm \sigma_{HH2}]$ are represented by a dark gray and white box, respectively. The area "*OV*" (light gray) represents the overlap of the intervals, as discussed in the text.

Here, μ_{AC} and μ_{DB} are mean values and σ_{AC} and σ_{DB} standard deviations of the two distributions. The more the intraclass similarity distribution *AC* shifts to the right side of the interclass distribution *DB*, the more the *OV* value decreases. The *OV* value could become negative ($(\mu_{DB} + \sigma_{DB}) < (\mu_{AC} - \sigma_{AC})$), which would provide an ideal situation for similarity searching. By contrast, when *AC* shifts to the left side of *DB*, the *OV* value increases which makes a separation of active and database compounds more difficult.

Plotting *OV* as a function of the α parameter (Figure 4), we can determine α values that minimize the overlap between the distributions and are thus preferred for similarity searching. These α values (approximated using a step-size of 0.1) are reported in Table 2. For MACCS and TGD, optimal α values were greater than 0.5 for activity classes CAT, HH2, and TNF, and smaller than 0.5 for NNI. For BEN, optimal α values were 0.6 for MACCS and 0.5 for TGD whose average bit densities were nearly identical for BEN and NCI. For PDR-FP, *OV* was constant because of its constant bit density and the results of search calculations were independent of α values. Taken together, these results confirmed that differences in fingerprint bit densities determine parameter settings for optimal Tversky similarity calculations.

We next determined if differences in bit densities also influenced the results presented by Chen and Brown.[8] In addition to a proprietary corporate compound repository, these investigators also analyzed the NCI anti-AIDS data set. We applied the same filtering procedure reported by Chen and Brown and removed compounds having a molecular weight of less then 60 Da or more than 600 Da. The resulting compound set consisted of 38265 confirmed inactives and 1097 confirmed ac-

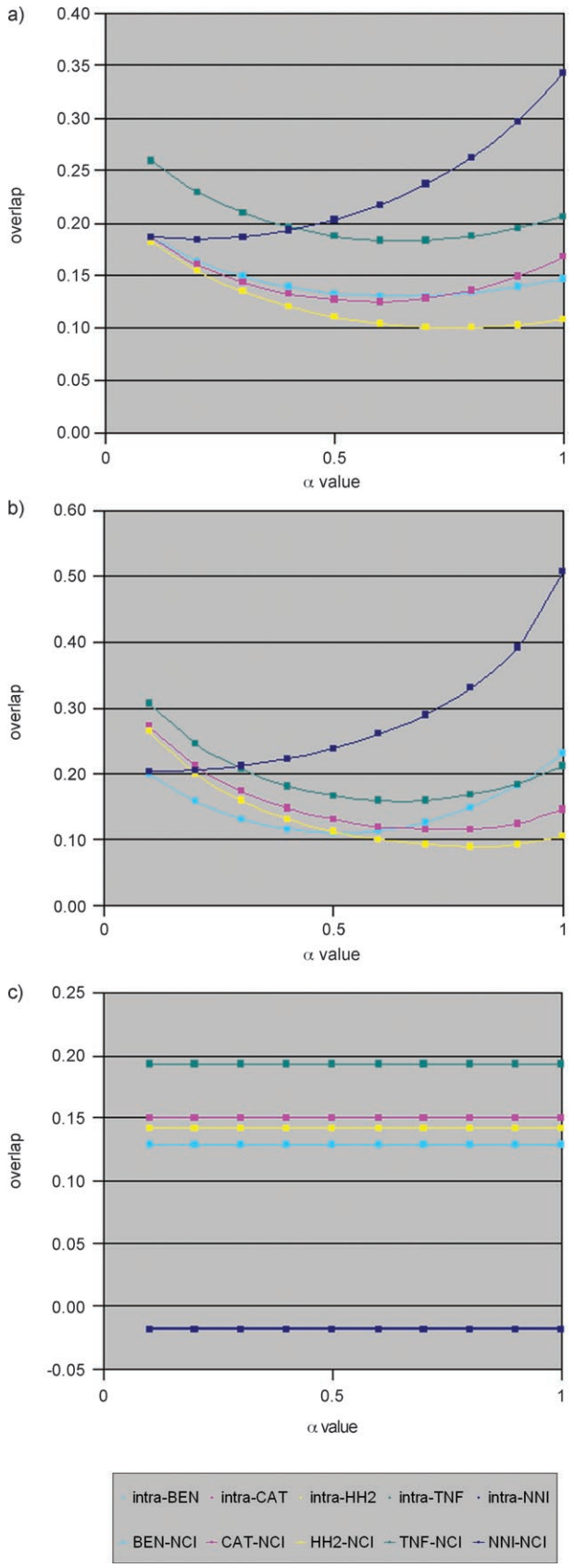

Figure 4. Statistical properties of distribution overlap. The overlap *OV* between intraclass and interclass Tversky similarity value distributions is shown as a function of the α parameter. The representation is according to Figure 2.

Table 2. Optimal α values.[a]

	MACCS	TGD	PDR-FP
BEN	0.6	0.5	–
CAT	0.6	0.7	–
HH2	0.8	0.6	–
NNI	0.2	0.1	–
TNF	0.6	0.8	–

[a] α values producing minimal overlap between intraclass and class-NCI Tversky similarity value distributions are shown as determined by graphical analysis of Figure 4. PDR-FP calculations are independent of α values because of its constant bit density. Therefore the overlap is also constant (see Figure 4c).

tives including moderately active compounds (see Table 3). These numbers were very similar but not identical to those reported by Chen and Brown, which we attribute to the use of slightly different (updated) versions of NCI. We then calculated fingerprint bit densities for active and inactive NCI compounds (Table 3). For MACCS and TGD, on average five more bits were set on in active than in inactive compounds, which rationalizes the preference for α values greater than 0.5 and explains the slight asymmetry observed in the hit-rate maps of Chen and Brown that were produced under variation of α.[8]

Similarity coefficients other than the Tversky coefficient are known to have varying degrees of size dependence.[5,10,11] For example, some coefficients preferentially detect compounds of different size and differences in fingerprint bit densities are generally found to affect compound retrieval[10]. Furthermore, the Tanimoto coefficient has different preferences for molecular size ranges in similarity calculations (selection on the basis of high values) and diversity calculations (low values), which can be attributed to size-dependent differences in fingerprint bit densities.[11] The study of Chen and Brown went beyond the analysis of characteristics of symmetrical similarity coefficients and provided evidence for improved compound recall when Tversky similarity calculations were carried out in an asymmetrical manner. We have been able to demonstrate that the apparent preference for asymmetrical Tversky coefficients is a direct consequence of systematic differences in fingerprint bit density between reference and database compounds.

For bit densities that correlate with molecular size, we need to distinguish three principal cases for the assessment of Tversky similarity. 1) If active compounds are on average larger than database compounds, $\alpha > 0.5$ produces the highest hit rates. 2) If active compounds are smaller than database compounds (such as for class NNI in our analysis), $\alpha < 0.5$ gives the highest hit rates. 3) If there are no differences in bit densities and size, Tversky similarity calculations are independent of the α parameter and always symmetrical. Are there consequences for similarity searching? Sets of active molecules available for similarity search calculations are typically optimized leads or drug candidates taken from the scientific or patent literature. These compounds tend to be larger than average database molecules. It is therefore not surprising that many activity classes used in benchmark calculations produce high hit rates

Table 3. Bit statistics of active and inactive NCI compounds.[a]

Data set	numCpds	numHA	MACCS 1-Bits	TGD 1-Bits	PDR-FP 1-Bits
Confirmed actives	255	25.5	53.7	65.4	93.0
Confirmed actives and confirmed moderate actives	1097	24.4	47.1	55.0	93.0
Confirmed inactives	38265	23.1	42.2	50.5	93.0

[a] Reported are statistics for active and inactive database molecules in the filtered NCI database. Abbreviations are used as defined in the legend of Table 1.

for α values greater than 0.5. However, in practical similarity search applications, we aim to identify novel hits that are typically smaller than available templates and still need to be optimized[12] and thus α values smaller than 0.5 would be most relevant.

Conclusions

Application of the Tversky similarity measure makes it possible to calculate molecular fingerprint similarity in a symmetrical and asymmetrical fashion. However, similarity calculations have asymmetrical characteristics only when fingerprints have different bit density. For a fingerprint design with constant bit density such as PDR-FP, *Tv* calculations are always symmetrical and independent of α parameter settings. For conventional 2D fingerprints such as MACCS, bit density is usually much influenced by molecular size. Our analysis has uncovered a direct relationship between fingerprint bit densities and asymmetry of Tversky similarity calculations and demonstrated that differences in bit densities determine preferred *Tv* parameter settings.

Experimental Section

Compound activity classes were extracted from the Molecular Drug Data Report[13] (MDDR) such that each compound in each set contained a unique cyclic carbon skeleton.[14] These scaffolds were generated using an in-house Perl script. Scaffold-based compound selection was carried out to avoid potential bias from similarity calculations on series of analogues. In the assembly of our compound sets we also monitored the molecular weight distribution, as discussed in the text. As background database for similarity searching, we used the publicly available NCI anti-AIDS database[15] (NCI) that was also used by Chen and Brown.[8] We also applied similar 2D fingerprints including MACCS structural keys[16] (166 bit positions) and TGD[17] (420 bits), a 2D two-point pharmacophore-type fingerprint that encodes distances between feature pairs similar to atom-pair descriptors.[18] In addition, we used another fingerprint recently developed in our laboratory, termed PDR-FP (500 bits), because this has the unique feature that it produces a constant bit density (93/500 bits) for test molecules irrespective of their size.[9]

Keywords: chemoinformatics · fingerprints · molecular similarity · Tversky coefficients · virtual screening

[1] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.
[2] H. Eckert, J. Bajorath, *Drug Discovery Today* **2007**, *12*, 225 – 233.
[3] M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York, **1990**.
[4] A. Tversky, *Psycol. Rev.* **1977**, *84*, 327 – 352.
[5] D. R. Flower, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379 – 386.
[6] J. W. Godden, L. Xue, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163 – 166.
[7] G. M. Maggiora, V. Shanmugasundaram, *Methods Mol. Biol.* **2004**, *275*, 1 – 50.
[8] X. Chen, F. K. Brown, *ChemMedChem* **2007**, *2*, 180 – 182.
[9] H. Eckert, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 2515 – 2526.
[10] N. Salim, J. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435 – 442.
[11] J. Holliday, N. Salim, M. Whittle, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819 – 828.
[12] M. M. Hann, T. I. Oprea, *Curr. Opin. Chem. Biol.* **2004**, *8*, 255 – 263.
[13] Molecular Drug Data Report (MDDR), MDL Elsevier, San Leandro, CA (USA) **2005** (http://www.mdl.com).
[14] Y. L. Xu, M. Johnson, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912 – 926.
[15] The publicly available NCI anti-AIDS database contains structure–activity data for the compounds screened by the National Cancer Institute AIDS antiviral screening program, **1999** (http://dtp.nci.nih.gov/docs/aids/aids_data.html).
[16] MACCS structural keys, MDL Elsevier, San Leandro, CA (USA) **2005** (http://www.mdl.com).
[17] The TGD fingerprint is available in the Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, QC (Canada) **2005** (http://www.chemcomp.com).
[18] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64 – 73.

Received: March 12, 2007
Revised: April 23, 2007
Published online on May 15, 2007

