Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching

Yuan Wang and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received August 20, 2007

Differences in molecular complexity and size are known to bias the evaluation of fingerprint similarity. For example, complex molecules tend to produce fingerprints with higher bit density than simple ones, which often leads to artificially high similarity values in search calculations. We introduce here a variant of the Tversky coefficient that makes it possible to modulate or eliminate molecular complexity effects when evaluating fingerprint similarity. This has enabled us to study in detail the role of molecular complexity in similarity searching and the relationship between reference and active database compounds. Balancing complexity effects leads to constant distributions of similarity values for reference and database molecules, independent of how compound contributions are weighted. When searching for active compounds with varying complexity, hit rates can be optimized by modulating complexity effects, rather than eliminating them, and adjusting relative compound weights. For reference molecules and active database compounds having different complexity, preferred parameter settings are identified.

INTRODUCTION

Similarity searching using fingerprint representations of molecules is a widely applied approach for chemical database mining.[1,2] It has a long history[1] and continues to be an active area of research.[3] Although many different designs have been introduced,[3] fingerprints have in common that they are bit string encodings of structural features and/or calculated molecular properties. In similarity searching, fingerprints are calculated for known active reference molecules and systematically compared to those of database compounds. Fingerprint overlap is quantified as a measure of molecular similarity using similarity coefficients. For bit string comparisons, many alternative similarity coefficients have been introduced[4,5] including the Tanimoto coefficient[4] (Tc), the currently most popular one. The Tversky coefficient[6] (Tv) is another similarity metric that makes it possible to weight the contributions of bit settings of reference and database molecules.

Molecular complexity or size effects are known to bias similarity searching using fingerprints.[7] In addition, intrinsic statistical preferences for certain Tc values in fingerprint comparisons have also been noted.[8] Differences in molecular complexity result in relative differences in fingerprint bit density, regardless of specific molecular features; the more complex test molecules are, the higher are corresponding bit densities. Such effects lead to increasing statistical probabilities of bit matches when fingerprints are compared and produce artificially high similarity values, which often favors the recognition of complex and large molecules.[7] Moreover, differences in complexity between reference and database molecules can systematically affect the outcome of similarity searching. For example, apparent asymmetry in search calculations on large databases using the Tversky coefficient[9] was shown to be a direct consequence of differences in molecular complexity and resulting fingerprint bit densities.[10]

Limitations of fingerprint comparisons that are associated with relative differences in bit densities could in principle be overcome in two ways: either by designing fingerprints that have constant bit density, regardless of the nature of test molecules, or by introducing similarity metrics that are independent of bit densities. The first fingerprint having constant bit density has recently been developed.[11] Furthermore, a modified version of the Tanimoto coefficient has been reported that can be applied to balance discrepancies in bit settings.[12] Here we introduce a bit density-independent variant of the Tversky coefficient that makes it possible to systematically change the relative contributions of bits that are set on or off in similarity calculations. We thoroughly characterize the behavior of this coefficient in similarity searching for compounds having different degrees of complexity and analyze the relationship between complexity, similarity values, and hit rates.

METHODOLOGY

The Tversky coefficient[6] is defined as

$$\mathrm{Tv}(A,B,\alpha) = \frac{c}{\alpha(a-c)+(1-\alpha)(b-c)+c} = \frac{c}{\alpha(a-b)+b}$$

where a and b are the number of bits set on ("1" bits) in molecular fingerprints A and B, respectively, and c is the number of bits shared by A and B. For the purpose of our discussion, we regard A as a known active reference molecule and B as a database compound. The factor α weights the contribution of reference molecule A: the larger α becomes, the more weight is put on the bit settings of A and the less

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci700314x CCC: $40.75
Published on Web 12/15/2007

on database molecule *B*. For comparison, the Tanimoto coefficient[4] is defined as

$$\mathrm{Tc}(A,B) = \frac{c}{a + b - c}$$

where *a*, *b*, and *c* are used according to Tv. In contrast to Tversky similarity, the calculation of Tanimoto similarity does not permit weighting relative contributions of reference and database molecules.

Calculation of Tanimoto similarity is affected by systematic differences in the number of bits that are set on.[4,7,13] In general, "1" bits are sparsely set in fingerprints and "0" bits occur more frequently. The more sparsely bits are set on, the smaller Tc values generally become.[7] Furthermore, when Tc calculations were used to guide the selection of diverse compound subsets from libraries, selected molecules often displayed the tendency to be smaller than average database compounds.[14] These observations have prompted Fligner et al.[12] to introduce a modified version of the Tanimoto coefficient (MTc) that takes all bit positions into account (i.e., set on or off):

$$\mathrm{MTc}(\beta) = \beta\mathrm{Tc}_1 + (1 - \beta)\mathrm{Tc}_0$$

In this formulation, Tc_1 and Tc_0 are Tanimoto coefficients calculated for bits set on and off, respectively. The factor β must be empirically determined to adjust bit density effects. Using this modified coefficient, Fligner et al. were able to correct the prevalence of small compounds in diverse subsets taken from the NCI database.[12]

The relationship between "1" bits in two fingerprints *A* and *B* also determines the complexity dependence of Tversky similarity calculations.[10] That is, if a reference molecule has more bits set on than the database compounds, similarity values tend to descend with increasing α. By contrast, if a reference molecule has fewer bits set on, similarity values tend to ascend with increasing α. Corresponding relationships between "0" bits in fingerprints also systematically change similarity values when α increases, but the directions are reversed compared to "1" bits.[10] Thus, for Tv calculations, it is immediately apparent that taking both "1" and "0" bits into account provides a principal possibility to eliminate the influence of complexity or size effects. A form of the Tversky coefficient accounting for bits that are set off can be written as follows

$$\mathrm{Tv}'(A,B,\alpha) = \frac{c'}{\alpha(a' - c') + (1 - \alpha)(b' - c') + c'} = \frac{c'}{\alpha(a' - b') + b'}$$

where a' and b' denote the number of "0" bits in *A* and *B*, respectively, and c' is the number of "0" bits common to both. Using a weighted combination of Tv and Tv′ it is possible to balance different densities of "1" and "0" bits in fingerprints such that neither "1" nor "0" bits dominate similarity evaluation

$$\mathrm{weighted\_Tv}(A,B,\alpha,\beta) = \mathrm{wTv} = \beta\frac{c}{\alpha(a - b) + b} + (1 - \beta)\frac{c'}{\alpha(a' - b') + b'}$$

where β is defined as the weight on "1" bits, i.e., the larger β becomes, the more weight is put on "1"s and the less on "0"s; for $\beta = 1$, wTv = Tv and for $\beta = 0$, wTv = Tv′. The above equation can be further transformed:

$$\mathrm{wTv} = \beta\left(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'}\right) + \frac{c'}{\alpha(a' - b') + b'}$$

In this formulation, the term

$$\left(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'}\right)$$

can be viewed as a coefficient of β. When it is greater than 0, the linear wTv(β) function monotonously increases. By contrast, when the coefficient is negative, the function monotonously decreases. The characteristics of this coefficient are determined by the value of α and the intrinsic bit settings of the fingerprints that are compared. The bivariate function wTv(a,β) is expected to have a complex value distribution surface for different (α, β) combinations, and systematic variation of the α and β parameters best describes this similarity metric. However, some general characteristics can be deduced by comparing cases where reference molecules and active database compounds have significant differences in bit density and where bit densities are similar. When all other parameters in the wTv equation shown above remain constant and the reference molecules have fewer bits set on than potential hits, i.e., $a < b$, then the term

$$\frac{c}{\alpha(a - b) + b}$$

increases due to the decrease of the denominator. If $a < b$, it also follows that $a' > b'$ (because a' and b' are complementary to *a* and *b*). This reduces the term

$$\frac{c'}{\alpha(a' - b') + b'}$$

and, as a result, the term

$$\beta\left(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'}\right)$$

increases wTv values relative to the situation where bit densities are similar. Increasing α and β values will further amplify this trend, which also favors the detection of hits. By contrast, when reference molecules have more bits set on than potential hits, i.e., $a > b$, the term

$$\frac{c}{\alpha(a - b) + b}$$

decreases and the term

$$\frac{c'}{\alpha(a' - b') + b'}$$

increases, thereby reducing

$$\beta\left(\frac{c}{\alpha(a - b) + b} - \frac{c'}{\alpha(a' - b') + b'}\right)$$

and the resulting wTv values. The larger the difference between *a* and *b* is, the more difficult it becomes to achieve high wTv values for reference molecules and active database compounds. In fact, the term

$$\beta\left(\frac{c}{\alpha(a-b)+b}-\frac{c'}{\alpha(a'-b')+b'}\right)$$

could potentially become negative, which would significantly reduce wTv values for potential hits and make it very difficult to distinguish them from other database compounds. Thus, differences in complexity between reference and active database molecules might significantly complicate similarity evaluation and present difficult fingerprint search situations, as further analyzed and discussed below.

TEST CALCULATIONS

Similarity Profiles for Defined β Parameter Values. We first studied the effects of predefined β settings under systematic variation of α. Calculations were carried out on five compound classes assembled from the Molecular Drug Data Report (MDDR).[15] These classes included benzodiazepines (abbreviated BEN; 57 compounds), cathepsin inhibitors (CAT; 90), vasopressin antagonists (VAS; 109), neuronal injury inhibitors (NNI; 50), and TNF-α release inhibitors (TNF; 65). With the exception of VAS, these activity classes were previously used in Tv calculations.[10] They were designed to produce fingerprints with different average bit densities. VAS was newly assembled from the MDDR and had by far the highest average bit density among the classes we studied. As a background database, the NCI anti-AIDS compound collection (42 687 molecules)[16] was chosen because the same database was used in previous analyses of Tv similarity calculations that revealed strong complexity-dependence.[9,10] For similarity calculations, subsets of 20 compounds were selected from each activity class (except for CAT where subsets of 20−80 compounds were generated). Fingerprint bit densities of our activity classes significantly differed. Each active compound was used as an individual reference molecule and searched against the background database. For each class, average pairwise wTv similarity values were determined for α values ranging from 0 to 1 and constant β values of 0, 0.5, and 1, respectively. Calculations were carried out using two fingerprints; MACCS,[17] consisting of the set of 166 publicly available MDL structural keys (i.e., 166 bit positions) and, as a control, MP-MFP,[18] a hybrid fingerprint consisting of 110 MACCS keys and 61 binary-encode property descriptors (171 bits). Despite their comparable size and similar design, these fingerprints displayed different bit patterns for the compounds studied here, due to the presence of property descriptors in MP-MFP.

Similarity Searching for Simple or Complex Active Molecules. We next investigated the role of varying bit densities in similarity search calculations under systematic variation of α and β. Therefore, a set of 1,214 tyrosine kinase inhibitors (TKI) was assembled from the MDDR and divided into four subsets with increasing average MACCS fingerprint "1" bit density (from TKI01 to TKI04). The lowest- (TKI01) and highest-complexity (TKI04) subsets were used as reference sets in separate calculations where the remaining three subsets were added to the background database as potential hits. For each reference compound, search calculations were carried out under systematic variation of α and β, the top scoring 100 or 500 database compounds were selected, and hit rates were calculated and averaged for each subset, thus producing set-specific HR(α,β) values. For example, HR(0.3, 0.6) reports the hit rate calculated for wTv(α=0.3, β=0.6) used as the similarity coefficient.

Similarity Searching for Molecules Matching the Complexity of Database Compounds. The average MACCS "1" bit density of the background database (25.7%) was taken as a reference point to search for molecules that closely matched this density (i.e. hits with complexity comparable to an average database compound). For two of our activity classes (TKI and TNF), sets of compounds were assembled from the MDDR having bit densities very similar to the background database (TKI: 250 compounds, average bit density 25.2%; TNF: 250, 25.8%). These sets were added to the background database as potential hits. Then other sets of 50 compounds having average bit densities smaller than, comparable to, or larger than the background database were used as reference molecules (TKI: average bit densities of 18.7%, 25.2%, and 39.2%; TNF: 19.1%, 25.5%, and 34.4%). For all reference compounds, wTv similarity calculations were carried out under systematic variation of α and β, as described above, and set-specific HR(α,β) values were calculated for the top scoring 100 database compounds. The calculations were repeated applying Tanimoto similarity.

RESULTS AND DISCUSSION

Weighted Tversky Similarity. The introduction of a bit density-responsive variant of the Tversky coefficient has enabled us to systematically analyze the influence of differences in molecular complexity on similarity evaluation. The fingerprint density of "1" bits is used as a measure of molecular complexity, which often correlates with molecular size. However, molecules of increasing size do not always lead to higher fingerprint bit density, for example, when features such as ring systems are duplicated for which bits are already set. Weighted Tversky calculations provide the opportunity to study, through systematic variation of the α and β parameters, complexity effects in the context of variable weights on the bit settings of reference and database molecules. The interplay between these parameters ultimately determines similarity assessment and the outcome of fingerprint search calculations.

Complexity-Independent Similarity Calculations. The similarity profiles in Figure 1 report for five activity classes average database similarity for given β and systematically changing α values. For these compound classes, MACCS "1" bit densities range from 15−46%. Thus, "1" bits are sparsely set and "0" bits dominate the fingerprint bit settings, which is typically observed for fingerprints, regardless of their design. For $\beta = 0$, all weight is put on the "0" bits and for $\beta = 1$ all weight on the "1" bits. For $\beta = 0.5$, "0" and "1" bits are equally weighted. Thus, β settings of 0 or 1 emphasize complexity effects, whereas 0.5 eliminates them from similarity evaluation. For α values ranging from 0 to 1, increasing weight is put on the bit settings of reference molecules; $\alpha = 0.5$ equally weights reference and database molecules. Thus, wTv values calculated with $\alpha = 0.5$ and $\beta = 1$ are proportional to conventional Tanimoto similarity.

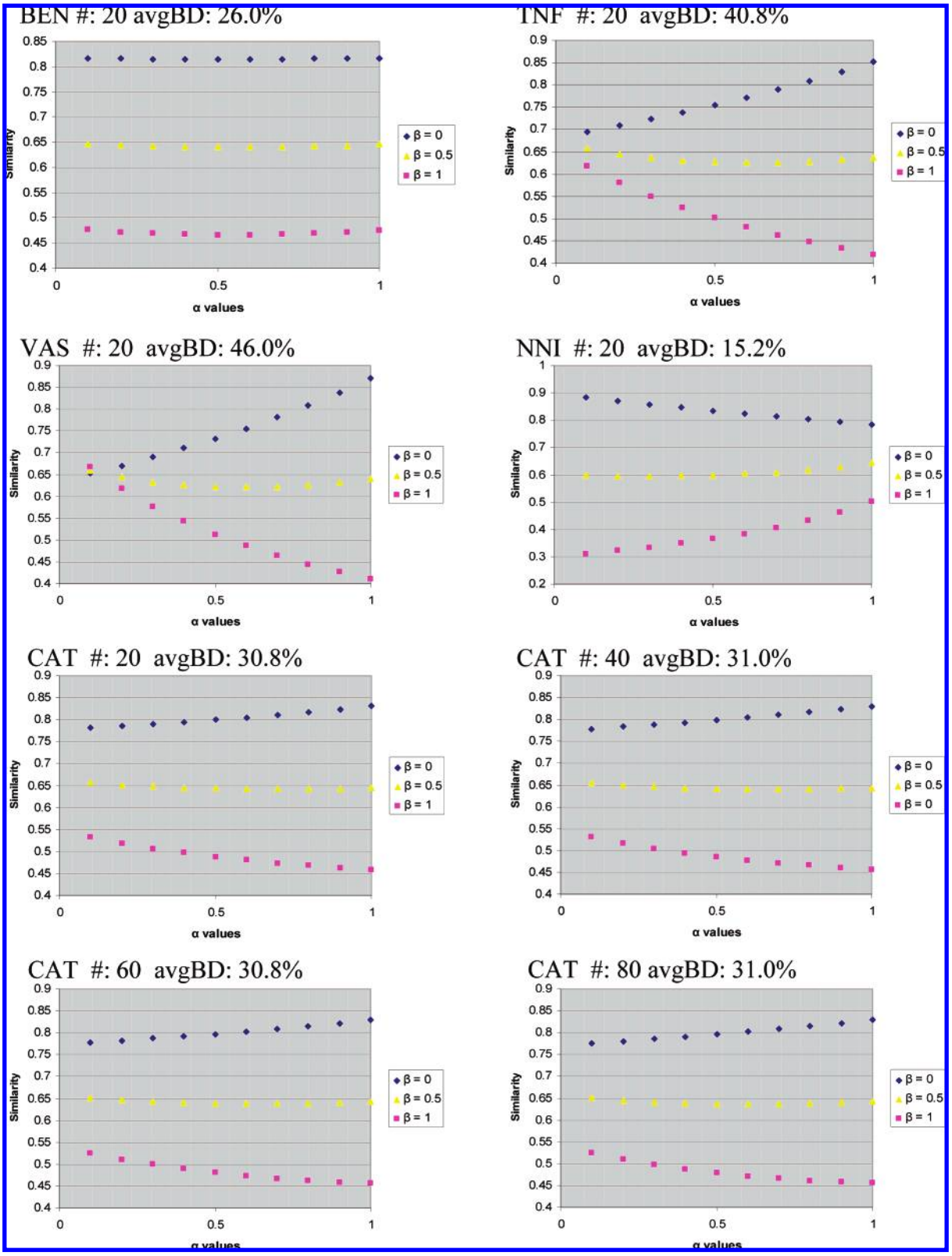


Figure 1. Similarity profiles. For five activity classes, average weighted Tversky similarity to background database compounds is calculated using the MACCS fingerprint. For each class, three curves are recorded for systematic variation of α and β values of 1 (i.e., complexity-dependent calculations over-weighting "1" bits), 0.5 (complexity-independent), and 0 (over-weighting "0" bits). Average fingerprint bit density ("avgBD") is reported, and the number ("#") of reference molecules is provided. The average bit density of the background database is 25.7%. CAT calculations are shown for four differently sized sets of reference molecules.

The profiles in Figure 1 illustrate the influence of complexity effects. As can be seen asymmetric similarity curves were obtained for activity classes whose bit densities differed from the database average. When bit densities of active molecules were higher than the database, the curves were monotonously increasing for $\beta = 0$ and decreasing for $\beta = 1$. When bit densities of active molecules were lower, these trends were reversed. The CAT profiles show that the similarity curves did not depend on the size of the reference set. Only BEN produced similarity values that were es-

sentially constant over the entire α range because its bit density was very similar to background database. When β was set to 0.5, complexity effects were balanced, and the similarity values were largely constant over the α range. Although BEN matched the bit density of the database, curves for β settings of 1 and 0 illustrate the consequences of sparsely set "1" bits in the MACCS fingerprint (bit density of 26%). At the ($\alpha = 0.5$, $\beta = 1$) reference point, the average similarity of 0.47 was artificially low; when complexity effects were balanced, i.e. ($\alpha = 0.5$, $\beta = 0.5$), the average similarity was 0.64. Fingerprints of all activity classes and database compounds contained more "0" than "1" bits, and thus similarity values for $\beta = 0$ were always higher than $\beta = 1$. Balanced average similarity relative to the database was ~0.65 for four activity classes and 0.6 for one (NNI). Thus, as one should expect, the average similarity calculated for a large number of database compounds was comparable for different activity classes when complexity no longer influenced the calculations. Balanced similarity values were fingerprint-dependent. In control calculations with MP-MFP (that is similar to, yet distinct from MACCS), average similarity for $\beta = 0.5$ was 0.74−0.75 for all classes.

Taken together, the data in Figure 1 illustrate the influence of complexity effects on similarity calculations and show that wTv calculations with $\beta = 0.5$ produce essentially constant similarity values that are independent of relative weights on reference and database molecules. Thus, in this case, database search calculations on active molecules are no longer biased by artificially increasing or decreasing similarity values.

Searching for Active Molecules Having Different Complexity. Next we carried out systematic wTv search calculations for tyrosine kinase inhibitors having increasing (Figure 2A) or decreasing (Figure 2B) bit density relative to the reference set. Hit rates for the top-scoring 100 or 500 database compounds were determined under systematic variation of the α and β parameters. Retrieval of active molecules and determination of hit rates present challenges that go beyond the similarity evaluation presented in Figure 1. This is the case because the detection of molecules having similar activity requires successfully distinguishing potential hits from average database compounds. It means that specific bit patterns must be detected that are only shared by active molecules.

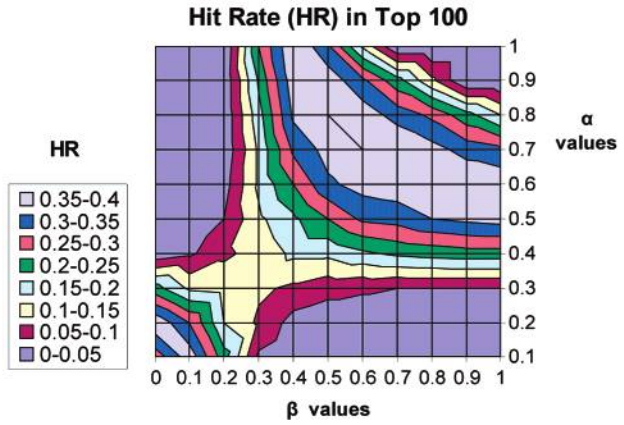
For low-complexity reference set TKI01, top hit rates between 25% and 45% were obtained with MACCS for selection sets of 100 database compounds. For high-complexity reference set TKI04, hit rates were generally lower (10−20%). In both cases, we observed that multiple (α, β) combinations produced preferred hit rates. However, top hit rates were generally not observed at the ($\alpha = 0.5$, $\beta = 1$) reference point for conventional similarity assessment. In fact, when bit densities of reference molecules and hits were different, similarity calculations using these parameter settings generally failed. However, top hit rates were typically also not produced by the ($\alpha = 0.5$, $\beta = 0.5$) parameter settings, i.e., when complexity effects were balanced ($\beta = 0.5$) and equal weight was put on the bit settings of reference and database molecules ($\alpha = 0.5$). In calculations with reference molecules and potential hits having similar bit density (top panel in Figure 2A, bottom in 2B), different (α, β) combinations produced top hit rates. When bit densities of reference molecules, potential hits, and database compounds were comparable, complexity effects only played a minor role. However, as discussed above, "0" bits dominated all fingerprint settings, and, therefore, increasing weight on shared "1" bits (i.e., increasing β) often improved hit rates in these cases. The top panel in Figure 2A and the bottom panel in Figure 2B also show an apparent approximate symmetry of hit rates along the ($\alpha = \beta$) diagonal because complementary combinations of (α, β) values produce equivalent (high or low) hit rates.

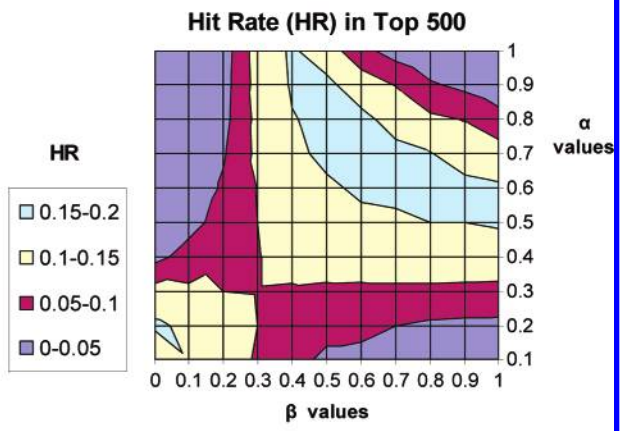
Importantly, when the complexity of reference molecules and potential hits differed, clear preferences for (α, β) combinations were observed. If the bit density of reference molecules was lower than that of potential hits (reference set TKI01, Figure 2A), combinations of high α and high β values produced the best hit rates. By contrast, if the bit density of reference molecules was higher than that of potential hits (reference set TKI04, Figure 2B) combinations of high α and low β values were preferred. In both cases, these parameter combinations increased wTv values for potential hits, which can be deduced from the wTv formula. Thus, these results are generally expected for reference molecules and hits having different fingerprint bit density. In these cases, modulating complexity effects, rather than eliminating them, and putting high weights on the bit settings of reference molecules optimized retrieval of active compounds.

Relevance for Virtual Screening. We next analyzed search calculations when potential hits closely matched the bit density of the background database and reference molecules of different complexity were used. The two instances where reference molecules have bit densities higher than or comparable to the database typically apply to practical virtual screening situations. This is the case because reference molecules for virtual screening are often optimized leads or drug candidates (having high complexity) or, alternatively, hits taken from experimental screening campaigns (with complexity often comparable to the database). Figure 3 reports calculations for two activity classes, TKI and TNF. The results reveal trends similar to those seen in Figure 2. Again, modulating complexity effects through variation of α and β resulted in the best hit rates. Table 1 reports that wTv calculations produced overall better hit rates than control calculations using standard Tanimoto similarity. Figure 4 shows examples of reference molecules of varying bit density and corresponding hits. This figure also illustrates that the density of "1" fingerprint bits provides a meaningful measure of molecular complexity. The results in Figure 3 make it possible to distinguish between three search situations. Calculations with reference compounds having lower bit density than the database are less relevant for virtual screening than the other two cases. Here combinations of high α and high β values were preferred, as discussed above. By contrast, when the bit densities of reference molecules, database compounds, and hits were comparable, many (α, β) combinations produced top hit rates. However, when reference molecules were more complex than hits and the background database, which is highly relevant for virtual screening, hit rates were much lower. Despite these very low hit rates that made the evaluation of parameter combinations difficult, there was also a preference for high α and low β values, at least in the case of TNF. Clearly, the case where

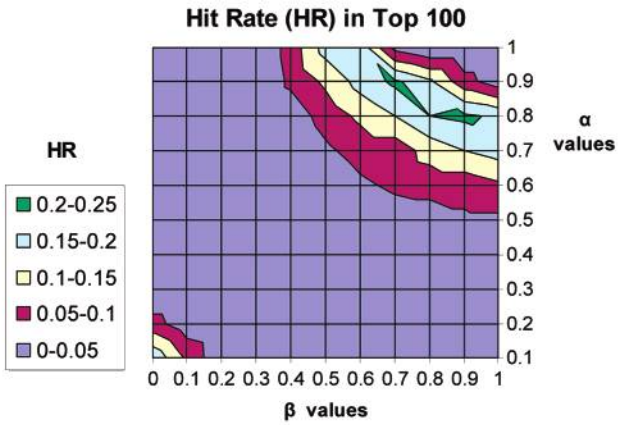
A Reference compounds: TKI01 #: 300 avgBD: 18.8%

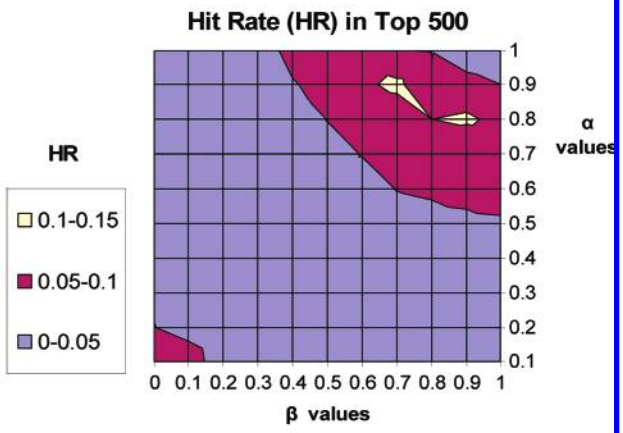
Potential hits: TKI02 #: 300 avgBD: 25.2%



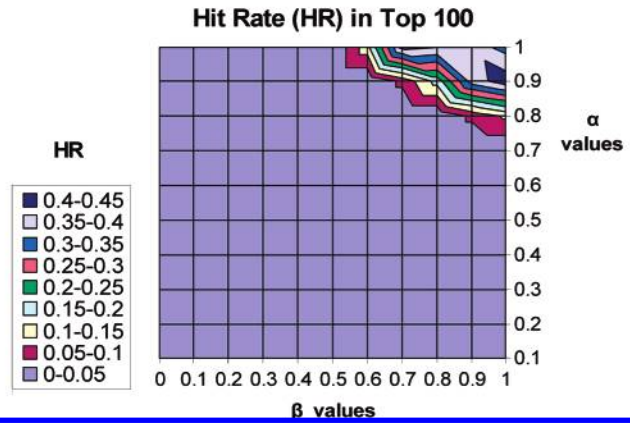


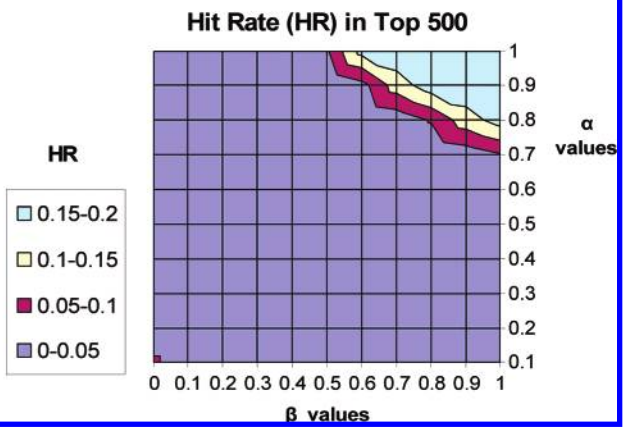
Potential hits: TKI03 #: 300 avgBD: 31.0%





Potential hits: TKI04 #: 314 avgBD: 39.5%





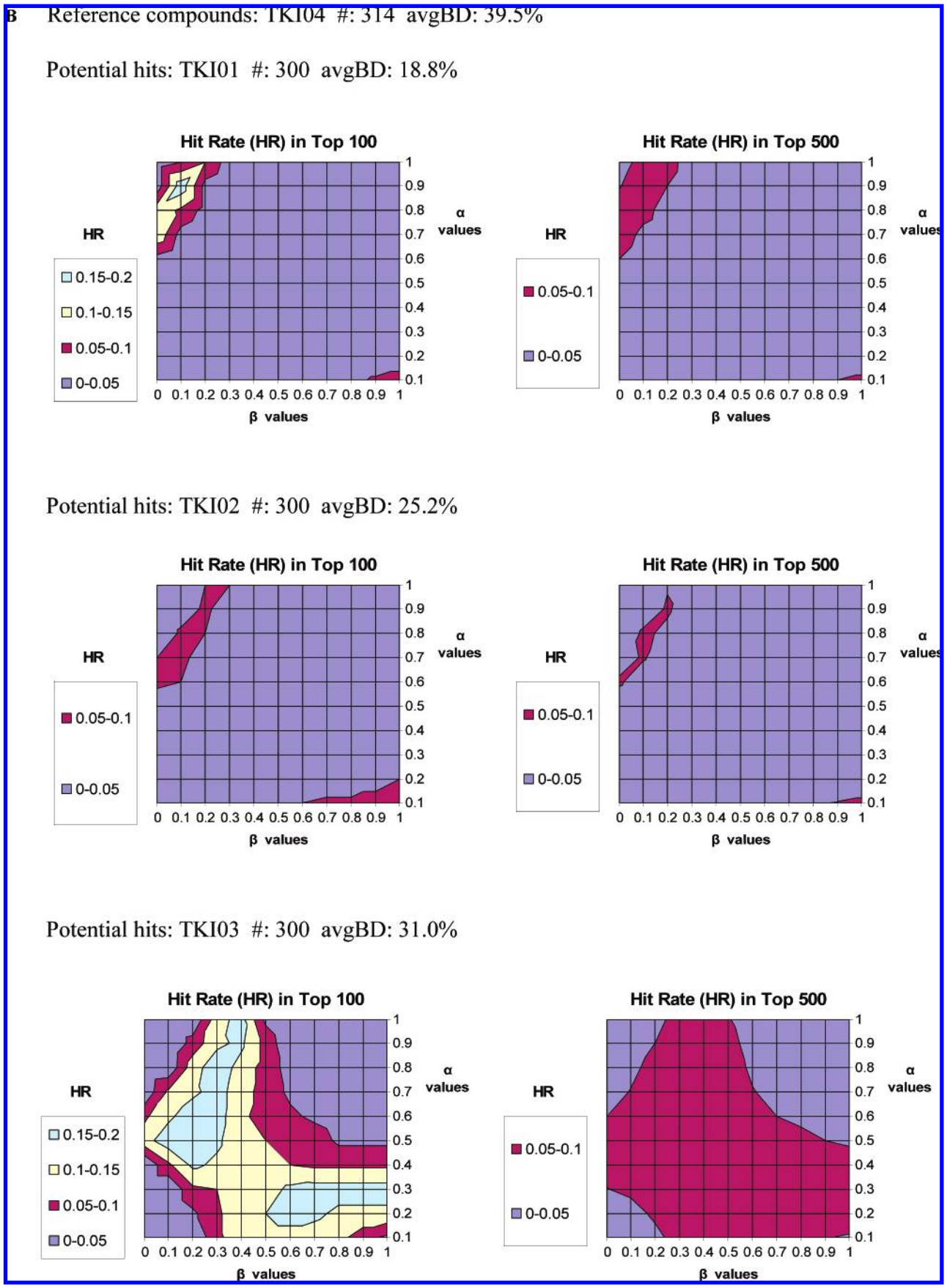


Figure 2. Searching for molecules of varying bit density. Similarity search trials are reported for 1214 TKI compounds divided into four subsets of increasing bit density. (A) Searching for TKI02, TKI03, and TKI04 using the low-avgBD set TKI01 as reference molecules. (B) Searching for TKI01, TKI02, and TKI03 using the high-avgBD set TKI04 as the reference. The numbers of reference compounds and potential hits are given. Weighted Tversky similarity is calculated using MACCS, and hit rates are reported for the top-scoring 100 (left) or 500 (right) database compounds. Hit rate maps are color-coded as indicated.

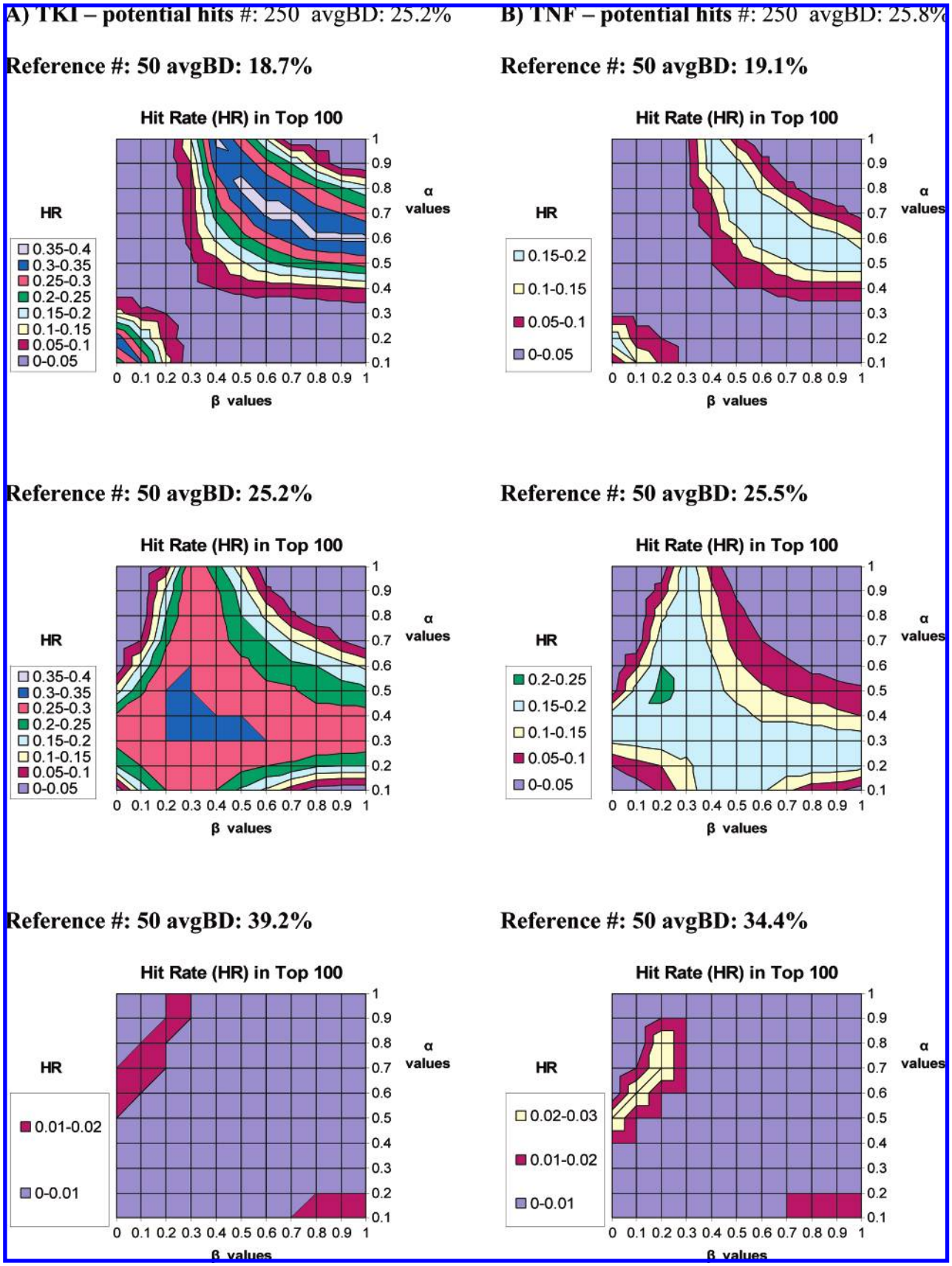


Figure 3. Searching for molecules having average database bit density. Similarity search calculations are reported for TKI and TNF. Potential hits closely match the bit density of database compounds, but the bit density of the reference molecules varies. Weighted Tversky similarity is calculated using MACCS, and hit rates are reported for the top 100 database compounds. Results are presented according to Figure 2.

reference molecules had higher complexity than potential hits presented the most challenging search scenario (where evaluation of standard Tanimoto similarity failed). These findings are well in accord with principal expectations

Table 1. Hit Rates[a]

class	hits avgBD (%)	Ref. avgBD (%)	Tc hit rate (%)	Tc_Hits avgBD (%)	wTv hit rate (%)	wTv_Hits avgBD (%)
TKI	25.2	18.7	23	24.5	36	24.7
		25.2	28	25.1	30	24.8
		39.2	0		1	25.3
TNF	25.8	19.1	20	25.6	19	25.7
		25.5	8	27.4	21	25.4
		34.4	0		3	25.7

[a] The best hit rates for selection of the top 100 database compounds are reported for TKI and TNF search calculations when potential hits closely match the MACCS bit density of the background database (25.7%). For each activity class, three sets of reference molecules with increasing bit density are used. "Ref." stands for reference molecules, "avgBD" stands for average bit density, and "Tc_Hits" and "wTv_Hits" stand for hits identified on the basis of Tanimoto and weighted Tversky similarity, respectively.

presented at the end of the Methodology section. Thus, the trends observed here should generally apply to wTv calculations and related similarity metrics.

CONCLUDING REMARKS

Fingerprint search performance is determined by chosen molecular representations, intrinsic features of fingerprint descriptors, chosen search strategies, and the way fingerprint similarity is quantified. Similarity evaluation depends on the similarity measures that are used. Molecular complexity effects and sparsely set "1" bits are known to bias fingerprint similarity calculations. When discussing aspects of molecular complexity in the context of similarity evaluation, it should also be considered that alternative molecular representations (for example, 2D versus 3D representations) mirror complexity in different ways. Clearly, molecular complexity is determined by multiple components. Depending on the chosen molecular representations, not all factors that contribute to complexity might be taken into account. Table 2 provides examples of complexity-relevant factors that can be accounted for at the level of 2D representations and others that require the use of 3D representations. However, regardless of which factors are ultimately considered, when using (2D or 3D) fingerprints, differences in molecular complexity and size typically lead to intrinsically different bit densities.

With the weighted Tversky coefficient, we have introduced a versatile similarity metric that makes it possible to study and balance complexity effects and differently weight contributions of reference and database compounds. The interplay between these parameters produces complex similarity value distributions that we have analyzed to study the influence of molecular complexity on fingerprint searching in detail. Balancing complexity effects leads to constant similarity values for reference and background database molecules, independent of how compound contributions are weighted. Under these conditions, no systematic errors occur in calculating the similarity of database molecules. However, taking differences in molecular complexity into account also provides opportunities to optimize the retrieval of active compounds. Accordingly, in fingerprint searching for active compounds having different complexity, modulating complexity effects, rather than eliminating them, and putting high weight on reference molecules led to the best hit rates in our analysis. Hit rate maps have revealed preferred parameter

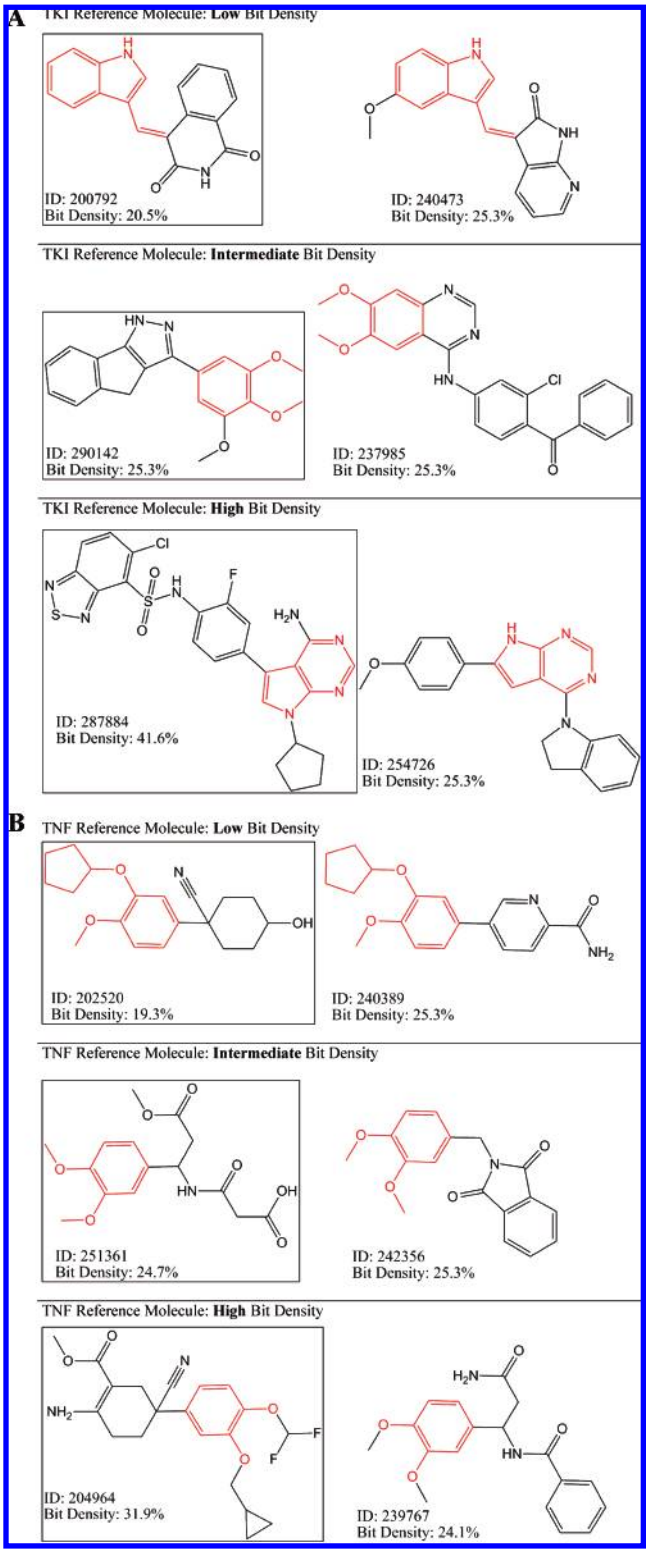


Figure 4. Structures of reference molecules and hits. Examples of TKI (A) and TNF (B) reference molecules of varying complexity are shown (boxed) together with hits identified using these reference molecules in the calculations summarized in Figure 3. MDDR IDs and bit densities are reported. Substructures shared by corresponding reference molecules and hits are colored red.

combinations for similarity searching and helped to better understand preferred characteristics of reference molecules, which has implications for virtual screening. In wTv calculations, highly complex molecules are, for principal reasons, much less suitable as references than active compounds

Table 2. Factors Related to Molecular Complexity[a]

factors	
2D	3D
element distribution	conformational entropy
H-bond acceptors/donors	electrostatic potentials
hybridization states	interatomic distance distr.
rigidity	intramolecular interactions
bond topology	stereochemistry

[a] Examples of factors are listed that contribute to molecular complexity together with the dimensionality of the molecular representation that is required to capture or deduce them.

having complexity comparable to the screening database. The findings reported herein are expected to trigger further analyses of similarity metrics and aid in the design of sound fingerprint search protocols.

REFERENCES AND NOTES

(1) Willett, P. Similarity-based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(2) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(3) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225−233.
(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(5) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. *Methods Mol. Biol.* **2004**, *275*, 1−50.
(6) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327−352.
(7) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(8) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.
(9) Chen, X.; Brown, F. K. Asymmetry of Chemical Similarity. *ChemMedChem* **2007**, *2*, 180−182.
(10) Wang, Y.; Eckert, H.; Bajorath, J. Apparent Asymmetry in Fingerprint Similarity Searching is a Direct Consequence of Differences in Bit Densities and Molecular Size. *ChemMedChem* **2007**, *2*, 1037−1042.
(11) Eckert, H.; Bajorath, J. Design and Evaluation of a Novel Class-directed 2D Fingerprint to Search for Structurally Diverse Active Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2515−2526.
(12) Fligner, M.; Verducci, J.; Blower, P. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.
(13) Salim, N.; Holliday, J.; Willet, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.
(14) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspect. Drug. Discovery Des.* **1997**, *7/8*, 65−84.
(15) *Molecular Drug Data Report (MDDR), version 2005.2*; MDL Elsevier: San Leandro, CA, 2005.
(16) The publicly available NCI anti-AIDS database contains structural and activity data for compounds screened by the AIDS antiviral screening program of the National Cancer Institute, 1999. http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed Feb 2007).
(17) *MACCS structural keys*; MDL Elsevier: San Leandro, CA, 2002.
(18) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1151−1157.

CI700314X