Development of a Compound Class-Directed Similarity Coefficient That Accounts for Molecular Complexity Effects in Fingerprint Searching

Yuan Wang and Jürgen Bajorath*

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

Received March 20, 2009

In chemoinformatics, fingerprints are defined as bit string representations of molecular structure and properties. The evaluation of fingerprint similarity, that is, quantification of fingerprint overlap, is known to be biased by differences in molecular complexity and size. For example, similarity searching using optimized lead compounds that are typically more complex and larger than average database compounds often leads to the artificial selection of large molecules. A modified version of the Tversky coefficient has been introduced to balance such complexity effects. In addition, compound class-directed fingerprint bit position-dependent weight vectors have been designed to aid in the assessment of Tanimoto similarity. We show that by merging weight vectors with the modified Tversky coefficient, a class-directed similarity metric is obtained that effectively balances complexity effects and further improves the accuracy of fingerprint search calculations.

INTRODUCTION

Similarity searching using molecular fingerprint representations is one of the most widely applied techniques for chemical database mining.[1,2] Typical molecular fingerprints are bit string encodings of structural features or calculated physicochemical properties of small molecules.[3,4] Different types of fingerprints have been introduced that encode, for example, structural fragments, topological descriptors, or 2D/3D pharmacophore patterns.[3] Furthermore, a variety of similarity metrics are available for fingerprint comparison and the quantification of fingerprint overlap[4] including, for example, the popular Tanimoto coefficient (Tc) or the Tversky coefficient[5] (Tv), that makes it possible to weight the contributions of bit settings of reference and database molecules.

Despite the relative simplicity of many fingerprint designs, in particular, 2D fingerprints, these descriptors and search tools have been shown to be surprisingly successful in many similarity search applications.[1,3] However, a known conundrum of fingerprint search calculations is their vulnerability to molecular complexity effects.[6] Independent of specific features that fingerprints encode, their bit density generally increases with increasing topological complexity or size of test molecules. Such effects often lead to artificially high similarity values because fingerprints with high bit densities have high statistical probabilities to match bit positions in other fingerprints.[6] As a consequence, complex and large molecules are preferentially detected in such calculations and the result of similarity searching are often systematically affected by bit density differences between reference and database molecules. For example, in a previous study Tversky similarity calculations showed apparent asymmetric behavior because optimal search performance was achieved by assigning higher weights on bit settings of reference than database compounds.[7] However, this phenomenon was shown to be a direct consequence of different levels of complexity of reference and database molecules and the resulting differences in fingerprint bit densities.[8] For Tversky similarity calculations, such biasing effects could be corrected by introducing the weighted Tversky coefficient (wTv), which made it possible to set relative weights on "1" and "0" bits and thereby balance complexity differences between reference and database molecules.[9] In general, fingerprint searching with chemically optimized reference compounds that were more complex than average database molecules made it most difficult to identify novel hits.[9]

In addition to balancing complexity effects, emphasizing compound class-specific bit patterns in similarity calculations has been shown to improve fingerprint search performance.[10−14] For example, by systematic silencing of bit positions, that is, converting individual "1" bits to "0", the contribution of each fingerprint bit to the search performance can be evaluated and a class-specific similarity metric, the bit position-weighted Tanimoto coefficient (bwTc), can be derived.[14] A bit position is assigned a high weight in the bwTc similarity comparison if its silencing causes a reduction in the recall of active compounds; the larger the reduction, the higher the weight. In *k*-NN nearest neighbor searching[15] using multiple reference compounds, bwTc produced higher compound recall than conventional Tc calculations.[16] In *k*-NN calculations, the final similarity score of a database molecule is the average of the similarity values of the *k* highest-scoring reference compounds.

Different similarity coefficients have also been systematically evaluated in fingerprint search calculations utilizing compound reference sets of varying complexity and the best-performing coefficient for each complexity level has been determined.[16] When reference and database compound had comparable complexity, Tanimoto similarity calculations were found to be preferred. However, when reference

* To whom correspondence should be addressed. Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/ci900108d CCC: $40.75
Published on Web 05/29/2009

molecules were more complex than database compounds, the Forbes or Simple Match coefficient performed best.[16]

In this study, we introduce a similarity metric that simultaneously balances complexity effects and emphasizes compound class-specific bit settings during fingerprint searching. This class-directed similarity coefficient is generated by combining the wTv and bwTc functions. The resulting "weighted Tversky coefficient with class-specific bit weighting" or wbwTv represents a parametric account of similarity. In systematic search calculations utilizing compound reference sets of increasing complexity, wbwTv search calculations outperformed its parental methods and other similarity metrics.

METHODOLOGY

Bit Position-Weighted Tanimoto Coefficient. First we introduce the conventional form of the Tanimoto coefficient. Given a known active reference molecule A and a database compound B, the Tanimoto coefficient[4] for binary vectors is defined as

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

where a and b are the number of bits set on ("1" bits) in molecular fingerprints A and B, respectively, and c the number of bits shared by A and B. This function only accounts for the sum of "1" bits, that is, bits that are set off are not taken into account in similarity calculations.

The Tanimoto coefficient can also be applied to nonbinary vectors. Thus, when using two molecular bit vectors of length N, $\mathbf{A} = (a_1, a_2, \ldots, a_N)$ and $\mathbf{B} = (b_1, b_2, \ldots, b_N)$, Tc can be represented as

$$\mathrm{Tc}(\mathbf{A},\mathbf{B}) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i)} \tag{2}$$

In this formulation, a_i and b_i are variables representing the ith bit position in fingerprint **A** and **B**, respectively, and $a_i b_i$ represents their product.

Furthermore, variable weights can be added to each individual bit position by calculating the product of the Tc and a weight vector **W**. Formally, this weight vector is a vector of N elements, $\mathbf{W} = (w_1, w_2, \ldots, w_N)$, which represents derived weights on the N bit positions of a fingerprint. Calculating the product of the weight vector and the general Tanimoto coefficient yields a *bit position-dependent* Tc, which we term bwTc and which is defined as[14]

$$\mathrm{bwTc}(\mathbf{A},\mathbf{B},\mathbf{W}) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i) w_i} \tag{3}$$

In this case, **W** is derived from training calculations on compound data sets where all "1" bit positions in a fingerprint are individually set off (i.e., to 0) prior to similarity searching. This procedure is termed *bit silencing*.[14] For silencing of each bit position, the resulting difference in compound recall rates is calculated and compared to the one obtained using the unmodified fingerprint. Thus, given N recall rates when silencing is applied to N individual bit positions, (hr_1, hr_2,..., hr_N) and hr_0, the corresponding recall rate for the unmodified fingerprint, the weight on the ith bit, w_i, is defined as

$$w_i = (1 + (\mathrm{hr}_0 - \mathrm{hr}_i) \times \mathrm{sf}) \times 100\% \tag{4}$$

where sf is an empirical scaling factor to further amplify rate changes (usually set to 100).[14] Expressing w_i in percentage scale is an arbitrary choice and does not change the similarity values.

Weighted Tversky Coefficient. Given a known active reference molecule A and a database compound B, the Tversky coefficient[5] is defined as

$$\begin{aligned}\mathrm{Tv}(\mathrm{A},\mathrm{B},\alpha) &= \frac{c}{\alpha(a-c) + (1-\alpha)(b-c) + c} \\ &= \frac{c}{\alpha(a-b) + b}\end{aligned} \tag{5}$$

where a, b, and c correspond to the Tc formulation in eq 1. 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 less on database molecule B. Analogously to Tc, here only "1" bit positions are taken into consideration.

To also account for "0" bit positions, an alternative form of the Tversky coefficient can be also defined that accounts for bit positions that are set off (i.e., "0" bits):

$$\begin{aligned}\mathrm{Tv}'(\mathrm{A},\mathrm{B},\alpha) &= \frac{c'}{\alpha(a'-c') + (1-\alpha)(b'-c') + c'} \\ &= \frac{c'}{\alpha(a'-b') + b'}\end{aligned} \tag{6}$$

where a' and b' denote the number of "0" bits in A and B, respectively, and c' the number of "0" bits common to both.

Furthermore, we can combine Tv and Tv′ and introduce a weighting parameter β to balance the relative contributions of "1" and "0" bits

$$\begin{aligned}\mathrm{weighted\_Tv}(\mathrm{A},\mathrm{B},\alpha,\beta) = \mathrm{wTv} = \beta\frac{c}{\alpha(a-b)+b} + \\ (1-\beta)\frac{c'}{\alpha(a'-b')+b'}\end{aligned} \tag{7}$$

where β is defined as the weight on "1" bits, that is, the larger β becomes, the more weight is put on "1"s (and less on "0"s) and vice versa. Thus, applying this modified form of Tversky similarity, "1" bit positions no longer intrinsically dominate the similarity calculation, which can be utilized to balance complexity effects arising from differences in fingerprint bit densities.[9]

Merging Bit Weight Vectors and Tversky Coefficients. The Tversky coefficient can also be generalized in analogy to Tc, as described above. Hence, given two molecular bit vectors of length N, $\mathbf{A} = (a_1, a_2, \ldots, a_N)$ and $\mathbf{B} = (b_1, b_2, \ldots, b_N)$, the general form of Tv becomes

$$\mathrm{Tv}(\mathbf{A},\mathbf{B},\alpha) = \frac{\sum_{i=1}^{N} a_i b_i}{\sum_{i=1}^{N} [\alpha(a_i^2 - b_i^2) + b_i^2]} \quad (8)$$

where a_i, b_i, and $a_i b_i$ are defined according to eq 2.

We can also incorporate the weight vector, $\mathbf{W} = (w_1, w_2, ..., w_N)$, into the general Tversky coefficient to obtain a bit position-weighted Tv, or bwTv, which is defined as

$$\mathrm{bwTv}(\mathbf{A},\mathbf{B},\mathbf{W},\alpha) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} [\alpha(a_i^2 - b_i^2) + b_i^2] w_i} \quad (9)$$

Accordingly, the alternative form of this coefficient accounting for "0" bits is represented as

$$\mathrm{bwTv}'(\mathbf{A},\mathbf{B},\mathbf{W},\alpha) = \frac{\sum_{i=1}^{N} a_i' b_i' w_i}{\sum_{i=1}^{N} [\alpha(a_i'^2 - b_i'^2) + b_i'^2] w_i} \quad (10)$$

where a_i' and b_i' are the complements of the ith bit element (i.e., $1 - a_i$ and $1 - b_i$, respectively) in fingerprint **A** and **B**. A weighted linear combination of eq 9 and eq 10 incorporating the β parameter then is

$$\mathrm{wbwTv}(\mathbf{A},\mathbf{B},\mathbf{W},\alpha) = \beta \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} [\alpha(a_i^2 - b_i^2) + b_i^2] w_i} + (1-\beta) \frac{\sum_{i=1}^{N} a_i' b_i' w_i}{\sum_{i=1}^{N} [\alpha(a_i'^2 - b_i'^2) + b_i'^2] w_i} \quad (11)$$

It follows that this similarity metric integrates three weighting schemes: (i) relative weights on "1" bit settings of reference and database compounds, (ii) relative weights on "1" and "0" bit positions, and (iii) compound class-specific weights on "1" bits. Thus, it is designed to balance differences in complexity between reference and database molecules and emphasize compound class-specific bit patterns in similarity calculations. In Figure 1, the design and calculation scheme of wbwTv is illustrated.

Control calculations were also carried out using the Simple Match and Forbes coefficient, as suggested for complex reference molecules.[16] If a and b represent the number of "1" bits in fingerprints A and B, respectively, c represents the number of "1" bits common to both A and B, d the number of "0" bits shared by A and B, and n the total number of bit positions comprising the fingerprint, the Forbes coefficient (For) is defined as

$$\mathrm{For}(\mathrm{A},\mathrm{B}) = \frac{nc}{ab} \quad (12)$$

and the simple match coefficient (SM) as

$$\mathrm{SM}(\mathrm{A},\mathrm{B}) = \frac{c+d}{n} \quad (13)$$

Compound Benchmark Systems and Test Calculations. For training and similarity searching, three sets of database compounds were used including a randomly collected set of 5000 ZINC[17] compounds (previously utilized in bwTc calculations[14]), the NCI anti-AIDS database,[18] containing 42687 compounds (previously used in wTv calculations[9]), and another randomly selected set of 50000 ZINC compounds that approximately matched the size of the NCI database. These screening databases were named ZINC5000, NCI, and ZINC50000, respectively. The ZINC5000 screening set was used to derive bit weight vectors, as described below, and evaluate systematic parameter variations in wTv and wbwTv calculations.

For bit silencing and systematic similarity search calculations, 10 compound activity classes were assembled from the MDDR,[19] as summarized in Table 1. To ensure that active compounds had properties comparable to the screening sets, they were filtered applying the rules of the ZINC database,[17] that is, maximum molecular weight of 600 Da, logP values between −2 and +6, no more than 18 rotatable bonds, and between one and 10 hydrogen bond donors and acceptors. Furthermore, in order to avoid potential bias through inclusion of series of analogs, only active molecules with distinct core structures were extracted from the MDDR, which was accomplished by applying a scaffold analysis algorithm.[20] The so prepared activity classes contained between 139 and 645 compounds. For all active and database compounds, MACCS fingerprints[21] consisting of 166 bit positions were calculated.

From each activity class, a subset of potential database hits of varying size (ranging from 10−100, Table 1) was selected having a MACCS bit density comparable to the screening database compounds, i.e. an average bit density of 0.22 (ZINC) to 0.25 (NCI). These subsets of active molecules having comparable complexity to screening set compounds served as active database compounds (ADC) for similarity searching. The bit density requirements limited the number of active compounds that could be selected as ADC. The remaining active molecules were utilized as training compounds for bit silencing and the derivation of the weight vectors. From each activity class training set, 10 different subsets of 20 compounds each were randomly selected and the remaining training compounds were added to ZINC5000 to derive the bit weight vector. Therefore, for each of the 10 reference sets, 166 bit silencing calculations were carried out (i.e., one for each bit position) in combination with 20-NN ranking, which equally takes contributions of all reference molecules into account, and hit rates were calculated for the top-ranked 100 database molecules. From these hit rates, ten individual weight vectors were calculated for each reference set with sf = 100 and the activity class-specific weight vector for each class was derived by averaging these reference set vectors.

Next, active reference compounds with different levels of complexity were selected for each activity class training set,

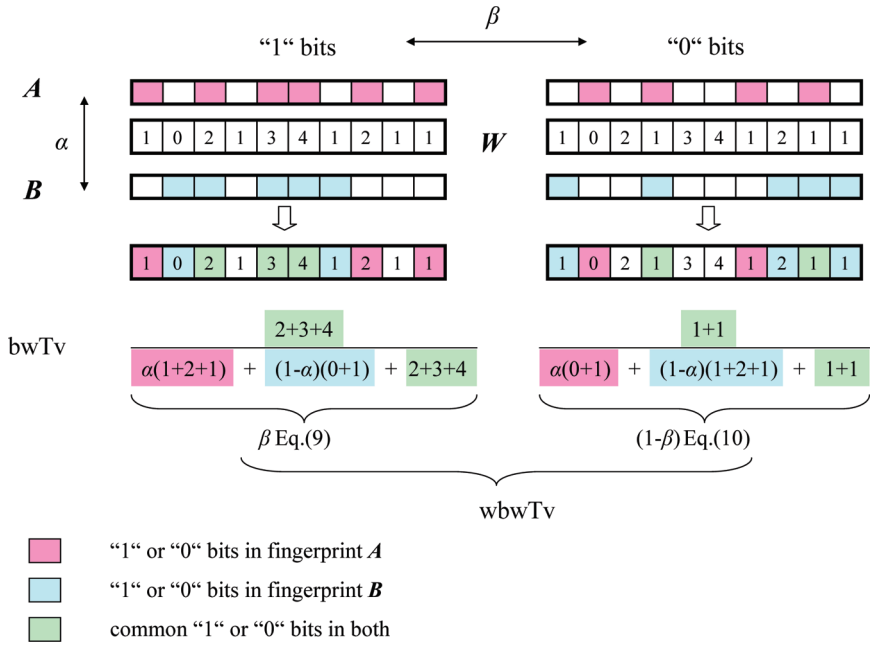

Figure 1. Calculation of wbwTv. Two hypothetical fingerprints **A** and **B** consisting of 10 bits are compared with wbwTv using the bit position-dependent weight vector **W** that assigns compound class-specific weights to "1" bits. The two parameters α and β modulate the relative weights on reference vs database compounds and on "1" vs "0" bits, respectively. The variables a, b, and c in Tv calculations (eq 5) are replaced with summation of weighted terms as described in eq 9. In addition, in Tv′ (eq 6), a', b' and c' are modified according to eq 10. For example, the number of "1" bits shared by the two fingerprints (c) is 3 in conventional calculations, whereas weighted calculations produce the value 9 (highlighted in green). The weighted linear combination of eq 9 and eq 10 yields the final wbwTv similarity value.

that is, 20 compounds with highest bit density, 20 having average bit density, and 20 with lowest bit density. These different reference sets for similarity searching were named level I (high complexity), II (moderate complexity), and III (low complexity). Level III reference compounds were comparable in complexity (i.e., bit density) to screening set compounds or slightly more complex. MACCS bit densities are reported in Table 1. These sets were used as the reference sets to search for ADC of the corresponding activity class. Exemplary structures of reference and screening set compounds and ADC are shown in Figure 2.

Similarity search calculations using six similarity metrics (Tc, bwTc, wTv, wbwTv, Forbes, Simple Matching) were carried out combined with 20-NN ranking in ZINC50000 and NCI. Compound recovery rates (i.e., the percentage of correctly identified ADC relative to the total number of ADC) were calculated for the top-ranked 100 database compounds. In wTv and wbwTv test calculations, the α and β parameters were systematically and independently varied between 0 to 1 in increments of 0.1. For the resulting 121 combinations, the top recovery rate of each calculation was determined. Hence, parameter variation was not involved in the training process to derive the weight vector.

Upon publication of this study, the composition of our especially designed compound benchmark systems including reference set levels I−III can be freely obtained for other studies on molecular complexity and activity class characteristic features in similarity searching via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

RESULTS AND DISCUSSION

In our analysis, we address the questions how similarity metrics controlling molecular complexity effects and emphasizing compound class-specific fingerprint features might be combined and what the consequences of using such similarity metrics might be for compound recall in fingerprint searching. We used multiple compound reference sets having different complexity and screening databases of different composition to thoroughly investigate differences in search performance of alternative similarity coefficients.

Complexity Effects and Tanimoto Similarity. The influence of varying molecular complexity on MACCS Tanimoto similarity calculations is evident in Table 2. For all compound classes and screening databases, compound recall of Tc calculations systematically decreased with increasing fingerprint bit density of reference compounds. For the least complex reference molecules (complexity level III), active compounds were detected in standard search calculations for seven of 10 classes in the ZINC (Table 2A), and all 10 classes in the NCI database (Table 2B). By contrast, for the most complex reference compounds (level I), Tc calculations consistently failed in ZINC (Table 2A) and for all but one class in NCI (Table 2B). Thus, in the presence of significant complexity effects, standard MACCS Tc calculations essentially failed to recover any active compounds. Using complexity level II reference molecules, active compounds were also only detected for two and three classes, respectively.

Bit Position-Dependent Tanimoto Similarity. Adding compound class-specific weights to bit positions (bwTc) only marginally improved the search performance for levels I and II. For level of III (where complexity effects are essentially absent), bwTc calculations produced moderate increases in compound recall for seven of 10 classes in Table 2A and six in 2B (i.e., 3%−10%, with one exception). Thus, complexity effects severely limited the influence of compound class weight vectors and the search performance of bwTc calculations.

Table 1. Activity Classes and Complexity Levels[a]

class		total	ADC	average ref bit density	description
ACE	I	245	30	0.37	angiotensin-converting
	II			0.32	enzyme inhibitors
	III			0.28	
ADR	I	320	70	0.41	aldose reductase
	II			0.33	inhibitors
	III			0.28	
CAM	I	143	10	0.42	cell adhesion molecule
	II			0.36	antagonists
	III			0.29	
CLG	I	166	20	0.40	collagenase inhibitors
	II			0.35	
	III			0.28	
FXA	I	645	40	0.50	factor Xa inhibitors
	II			0.38	
	III			0.28	
MM1	I	198	20	0.38	muscarinic M1 agonists
	II			0.33	
	III			0.27	
PA2	I	184	100	0.37	phospholipase A2
	II			0.36	inhibitors
	III			0.31	
PAF	I	248	50	0.43	platelet-activating factor
	II			0.35	antagonists
	III			0.28	
PKC	I	199	70	0.41	protein kinase C
	II			0.35	inhibitors
	III			0.30	
SST	I	139	40	0.38	squalene synthetase
	II			0.35	inhibitors
	III			0.28	
average	I			0.41	
	II			0.35	
	III			0.28	

[a] Compound activity classes are reported, and average MACCS bit densities for reference sets having different levels of complexity (I: high complexity, II: moderate complexity, III: low complexity). "total" gives the total number of compounds extracted from the MDDR as training, reference, and active database compounds. "ADC" gives the number of active database compounds (potential hits) that were available to match the average bit density of screening data sets. "average ref bit density" reports the average bit density of reference molecules having different levels of complexity.

Forbes and Simple Match Coefficients. For the most complex reference molecules, Forbes calculations detected active compounds in five (Table 2A) and three cases (Table 2B), where both Tc and bwTc calculations failed, whereas Simple Match calculations did not produce notable increases. However, Forbes calculations also frequently failed to detect active compounds on the basis of complex reference molecules and showed lower performance than Tc, bwTc, or Simple Match for level III reference molecules. For low-complexity reference compounds, the performance of the Simple Match coefficient was comparable to Tc and bwTc in ZINC (Table 2A) but was higher for seven of 10 classes in NCI (Table 2B).

Weighted Tversky Similarity. Different from Tc, bwTc, Forbes, or Simple Match, the bit position-independent weighted Tversky coefficient (wTv) balances complexity effects by modulating relative contributions of "1" and "0" bit positions. In this case, a systematic increase in compound recovery rates was found in both screening databases. For level I and level II reference compounds, wTv calculations failed in only three (ZINC) and two (NCI) instances, respectively, to recover active compounds and recall rates

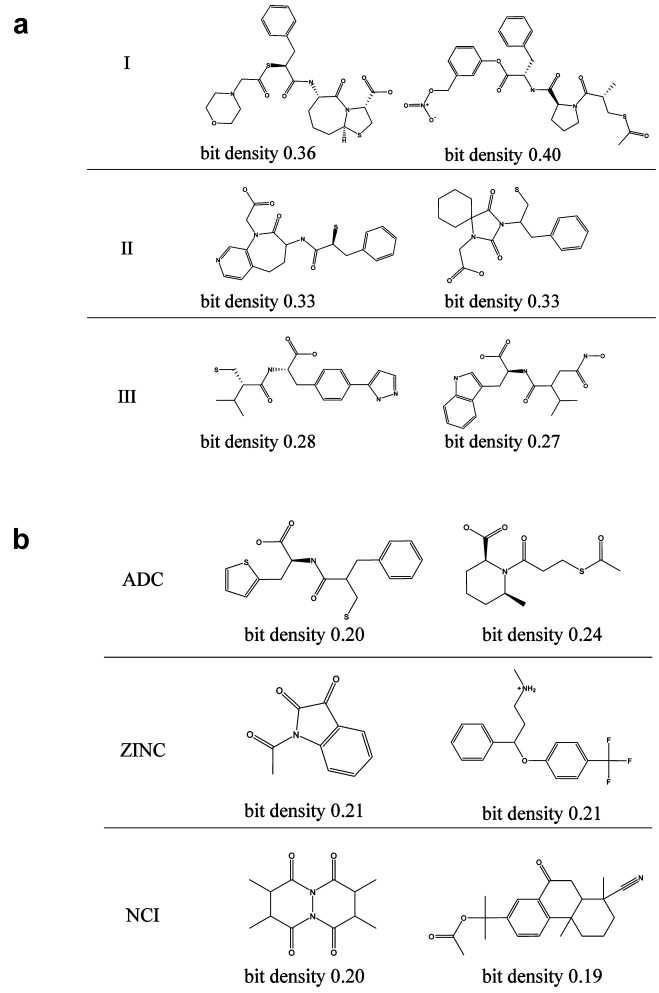

Figure 2. Exemplary compounds. For activity class ACE, examples are shown of (a) reference molecules of different complexity (level I, II, and III) and (b) active database compounds (ADC) and screening database molecules from ZINC and NCI having comparable complexity.

of up to 27% (level I) and 40% (level II) were obtained. Here, we also observed the general trend that recovery rates often increased from level I to level III. For the least complex reference molecules, wTv calculations produced average hit rates over 10 classes of ~23% in ZINC and ~27% in NCI. Thus, directly addressing complexity effects at the level of similarity calculations clearly improved the search results.

Weighted Tversky Similarity with Class-Specific Bit Weight Vectors. By designing compound reference sets with varying complexity, we have created search situations where conventional Tanimoto similarity calculations consistently fail. The results discussed above mirror the crucial role of complexity effects that were only effectively balanced in wTv calculations. In the presence of complexity effects, bwTc calculations that emphasized compound class-specific bit patterns also failed to produce significant compound recall. With wbwTv, we introduce a similarity coefficient that combines the complexity-balancing potential of wTv calculations with class-specific bit weight vectors. When applying wbwTv, we observed consistent improvements in recovery rates over all complexity levels. Top recovery rates were obtained in 18 of 30 cases (i.e., of three calculations per activity class) in Table 2A and in 19 cases in Table 2B. Thus, despite differences in compound compositions, results ob-

Table 2. Similarity Searching Using Different Similarity Coefficients[a]

A

class		Tc	bwTc	max wTv	max wbwTv	Forbes	Simple Match	class		Tc	bwTc	max wTv	max wbwTv	Forbes	Simple Match
ACE	I	0	0	**27**	23	17	0	PA2	I	0	2	4	**8**	0	3
	II	3	3	**33**	30	23	10		II	0	0	**3**	**3**	0	**3**
	III	57	60	77	**83**	33	77		III	**3**	**3**	**3**	**3**	2	**3**
ADR	I	0	0	**6**	0	3	0	PAF	I	0	0	0	0	0	0
	II	0	0	**9**	1	3	6		II	0	0	0	**2**	0	0
	III	4	7	**10**	9	4	7		III	2	6	**16**	12	2	**16**
CAM	I	0	0	0	0	0	0	PKC	I	0	0	**7**	**7**	**7**	0
	II	0	**20**	**20**	**20**	**20**	0		II	0	4	20	**29**	13	1
	III	0	**30**	20	**30**	0	20		III	4	10	20	**26**	11	10
CLG	I	0	5	15	**25**	0	0	SST	I	0	0	18	**23**	8	0
	II	0	15	30	**40**	25	0		II	5	3	20	**25**	20	20
	III	40	30	**45**	40	20	40		III	20	23	25	**28**	20	23
FXA	I	0	0	0	0	0	0	average	I	0	1	9	**10**	4	0
	II	0	0	0	0	0	0		II	1	5	14	**16**	11	4
	III	0	3	8	**25**	0	8		III	13	17	23	**27**	10	21
MM1	I	0	0	**15**	10	10	0								
	II	0	0	**5**	**5**	5	0								
	III	0	0	**10**	**10**	10	10								

B

class		Tc	bwTc	max wTv	max wbwTv	Forbes	Simple Match	class		Tc	bwTc	max wTv	max wbwTv	Forbes	Simple Match
ACE	I	0	0	27	**30**	23	3	PA2	I	3	3	11	**12**	0	7
	II	3	3	**40**	30	27	10		II	3	2	**8**	7	0	3
	III	57	60	**83**	**83**	47	**83**		III	3	3	5	**10**	2	3
ADR	I	0	0	**6**	3	3	0	PAF	I	0	0	0	0	0	0
	II	0	0	**10**	9	1	7		II	0	0	0	**2**	0	0
	III	6	26	11	**23**	4	10		III	4	10	**20**	12	0	16
CAM	I	0	0	0	0	0	0	PKC	I	0	0	7	**10**	7	0
	II	0	20	20	**30**	20	0		II	0	6	17	**19**	10	6
	III	20	30	**40**	**40**	20	**40**		III	4	10	16	**21**	6	13
CLG	I	0	5	15	**25**	5	0	SST	I	0	0	18	**23**	0	0
	II	0	10	**40**	**40**	30	0		II	10	3	20	**23**	10	20
	III	40	10	**45**	40	35	**45**		III	20	23	23	**25**	20	23
FXA	I	0	0	0	0	0	0	average	I	0	1	8	**10**	4	1
	II	0	0	0	0	0	0		II	2	4	**16**	**16**	10	5
	III	5	5	15	**28**	0	8		III	17	18	27	**30**	14	26
MM1	I	0	0	0	0	0	0								
	II	0	0	**5**	**5**	**5**	0								
	III	10	0	15	**20**	5	15								

C

class		20-NN Tc	1-NN Tc	centroid Tc	20-NN Forbes	1-NN Forbes	centroid Forbes	class		20-NN Tc	1-NN Tc	centroid Tc	20-NN Forbes	1-NN Forbes	centroid Forbes
ACE	I	0	0	0	23	3	23	PA2	I	3	3	3	0	3	0
	II	3	0	3	27	3	27		II	3	4	3	0	2	0
	III	57	60	73	47	40	47		III	3	8	3	2	4	2
ADR	I	0	0	0	3	7	3	PAF	I	0	0	0	0	0	0
	II	0	1	4	1	4	1		II	0	0	0	0	0	0
	III	6	21	9	4	9	4		III	4	26	8	0	14	0
CAM	I	0	0	0	0	0	0	PKC	I	0	0	0	7	9	7
	II	0	0	0	20	20	20		II	0	1	0	10	9	10
	III	20	30	20	20	20	20		III	4	14	7	6	7	6
CLG	I	0	0	0	5	0	5	SST	I	0	8	0	0	3	0
	II	0	0	0	30	10	25		II	10	20	20	10	20	5
	III	40	45	40	35	25	35		III	20	35	23	20	23	20
FXA	I	0	0	0	0	0	0	average	I	0	1	0	4	3	4
	II	0	0	0	0	0	0		II	2	3	3	10	7	9
	III	5	10	5	0	23	0		III	17	26	20	14	17	14
MM1	I	0	0	0	0	5	0								
	II	0	0	0	5	5	5								
	III	10	15	15	5	10	5								

[a] Average recovery rates (in %) are reported for MACCS search calculations using different similarity coefficients and the (A) ZINC50000 and (B) NCI screening databases. For each class, I, II, and III report the results for reference sets of varying complexity, according to Table 1. In each row, the best-performing similarity coefficient is boldface. In section C, search calculations using two similarity coefficients, Tc, and Forbes, and three data fusion techniques, 20-NN, 1-NN and centroid, are compared for the NCI database.

tained for the ZINC and NCI screening databases were overall similar in this case. In many instances, wbwTv calculations produced recall rates of ~20% or more when other similarity coefficients (in particular, Tc) completely failed. However, wbwTv calculations were not always successful. For example, for classes CAM, FXA, or PAF, level I reference molecules presented an intractable search problem for any of the similarity coefficients. In one case, ADR level I, wTv calculations detected a few active compounds (recovery rate 6%), but wbwTv essentially failed. In another case, PAF level II, the opposite occurred. With these minor exceptions, a clear trend was observed: when

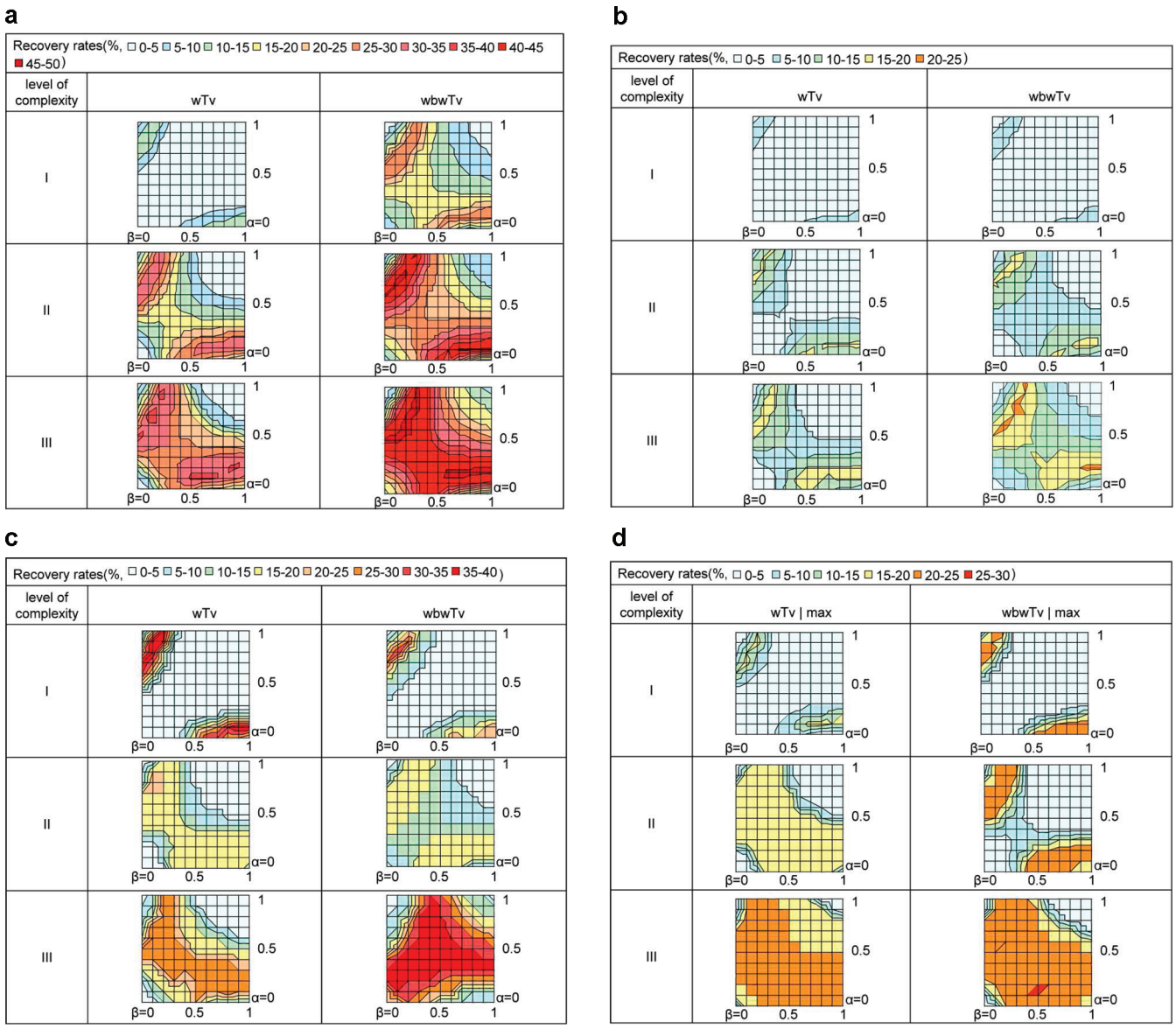

Figure 3. Recovery rate landscapes. Shown are maps reporting search results for wTv and wbwTv calculations under systematic parameter variation using reference sets of different complexity: (a) Class PKC/ZINC5000, (b) PKC/NCI, (c) MM1/ZINC5000, and (d) SST/NCI.

wTv was not capable of detecting active compounds, wbwTv was not either. However, when wTv calculations succeeded, an increase in recovery rates was often observed when wbwTv was applied, although the relative search performance varied in a compound class-dependent manner. For the total of 60 test calculations reported in Table 2, wTv and wbwTv recovery rates were the same in 19 cases and wTv and wbwTv performed best in 14 and 27 cases, respectively. Thus, taken together, these findings indicated that simultaneous balancing of complexity effects and emphasizing of class-specific bit settings yielded overall best performance in these difficult similarity search test cases.

Comparison with Data Fusion Approaches. Data fusion approaches have often been shown to further improve similarity search performance when using multiple reference molecules.[15,22,23] For example, *k* nearest neighbor (*k*-NN) calculations average the *k* highest similarity values calculated for reference molecules to produce the final similarity score of a database compound. In 1-NN calculations, only the highest similarity value is taken.[15,22] By contrast, centroid calculations, another data fusion approach, first generate an average vector from all reference compounds and then compare individual fingerprints of database compounds with this average fingerprint (often applying the general form of the Tanimoto coefficient shown in eq 2). In comparative studies, 1-NN calculations often produced highest compound recall rates among data fusion techniques and other fingerprint search strategies.[22,23]

We have also compared different data fusion techniques to wbwTv calculations. Table 2C reports the results for 20-NN, 1-NN, and centroid calculations and the Tc and Forbes similarity metrics on our compound test sets and the NCI database. Nearest neighbor calculations produced better results than centroid searches, but were overall inferior to wbwTv calculations, as reported in Table 2B, especially when reference compounds of high complexity were used. 1-NN Tc calculations moderately increased the search performance of 20-NN calculations by 1% to 9% for reference sets I−III, but recovery rates of wbwTv were 10% to 14% higher. A similar trend was observed for the Forbes coefficient. Overall, there were only two instances where 1-NN Tc performed better than wbwTv or wTv (PAF set III

and SST set III) and two where 1-NN Forbes performed better (ADR set I and MM1 set I), but the differences were marginal. It follows that data fusion techniques were not capable to effectively balance molecular complexity effects, as expected. By contrast, balancing complexity effects through wbwTv led to overall highest search performance.

Recovery Rate Landscapes. We have also compared recovery rate distributions for the overall preferred wTv and wbwTv coefficients under systematic variation of the α and β parameters. Representative examples are shown in Figure 3. In these recovery rate landscapes, regions color-coded in red represent parameter combinations producing high recovery rates. For PKC screening in ZINC, shown in Figure 3A, areas of high recovery rates were larger for wbwTv than for wTv. A similar trend was observed for PKC in the NCI, although recovery rates were in this case lower for both coefficients (Figure 3B). Equivalent observations were also made for MMI in ZINC (Figure 3C) and SST in NCI (Figure 3D). The recovery rate landscapes also reveal trends for preferred α, β parameter settings. For complexity level I, combinations of low α and high β or vice versa generally produced highest recovery rates, although search performance was low in these cases. Going from complexity level I to II and III combinations of increasingly larger α and β value ranges produced highest rates, while search performance was increasing.

In general, we found that wbwTv calculations produced larger areas of high recovery rates than wTv calculations and smaller areas where calculation produced only low recovery of active compounds (light blue in Figure 3). This means that wbwTv search calculations were less sensitive to α, β parameter settings than wTv calculations (i.e., more wbwTv parameter combinations produced high compound recall). Therefore, taking bit position-specific information into account made wbwTv search calculations more stable over all complexity levels, in addition to achieving net increases in recovery rates.

CONCLUDING REMARKS

Fingerprint-based searching generally depends on the similarity measures that are applied. Differences in topological complexity or size between reference and database compound lead to differences in bit density and often bias similarity calculations. We have introduced a complex similarity metric that is based on the Tversky formalism and simultaneously balances complexity effects and emphasizes class-specific bit settings. In systematic similarity searching over different compound classes and complexity levels, the wbwTv coefficient often produced significant recall in cases where standard Tanimoto similarity calculations failed and further improved the performance of the weighed Tversky coefficient that was previously introduced. Moreover, compared to the Forbes and Simple Match coefficients, which have been shown to be particularly suitable for searching with complex reference molecules, wbwTv achieved consistently higher recovery rates over all reference set complexity levels. In addition to practical similarity applications, wbwTv calculations can be utilized to study the relationship between molecular complexity and compound class characteristic features and further explore basic aspect of molecular similarity measures.

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) Tversky, A. Features of Similarity. *Psycol. Rev.* **1977**, *84*, 327–352.
(6) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Comput. Sci.* **1998**, *38*, 379–386.
(7) Chen, X.; Brown, F. K. Asymmetry of Chemical Similarity. *ChemMedChem* **2007**, *2*, 180–182.
(8) 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.
(9) Wang, Y.; Bajorath, J. Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching. *J. Chem. Inf. Model.* **2008**, *48*, 75–84.
(10) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm to Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
(11) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746–753.
(12) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
(13) Williams, C. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. *Mol. Div.* **2006**, *10*, 311–332.
(14) Wang, Y.; Bajorath, J. Bit Silencing in Fingerprints Enables the Derivation of Compound Class-Directed Similarity Metrics. *J. Chem. Inf. Model.* **2008**, *48*, 1754–1759.
(15) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Mmetrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(16) Chen, J.; Holliday, J.; Bradshaw, J. A Machine Learning Approach to Weighting Schemes in the Data Fusion of Similarity Coefficients. *J. Chem. Inf. Model.* **2009**, *49*, 185–194.
(17) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(18) 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. http://dtp.nci.nih.gov/docs/aids/aids_data.html (accessed February 1, 2007).
(19) *MDL Drug Data Report (MDDR)*, version 2005.2; Symyx Software: San Ramon, CA, 2005.
(20) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97–102.
(21) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(22) Hert, J.; Willett, P.; Wilton, D. J. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(23) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2006**, *2*, 208–217.

CI900108D