Bit Silencing in Fingerprints Enables the Derivation of Compound Class-Directed Similarity Metrics

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 June 16, 2008

Fingerprints are molecular bit string representations and are among the most popular descriptors for similarity searching. In key-type fingerprints, each bit position monitors the presence or absence of a prespecified chemical or structural feature. In contrast to hashed fingerprints, this keyed design makes it possible to evaluate individual bit positions and the associated structural features during similarity searching. Bit silencing is introduced as a systematic approach to assess the contribution of each bit in a fingerprint to similarity search performance. From the resulting bit contribution profile, a bit position-dependent weight vector is derived that determines the relative weight of each bit on the basis of its individual contribution. By merging this weight vector with the Tanimoto coefficient, compound class-directed similarity metrics are obtained that further increase fingerprint search calculations compared to conventional calculations of Tanimoto similarity.

INTRODUCTION

Similarity searching using fingerprint representations of molecules is one of the most widely applied approaches for chemical database mining.[1,2] Despite significant differences in their design,[3] all types of fingerprints are bit string encodings of structural features and/or calculated molecular properties.[4] A variety of similarity metrics are available for quantitatively comparing fingerprint overlap between reference and database molecules,[4] and the Tanimoto coefficient (Tc) continues to be the most popular one.[1,4] Different from many other virtual screening techniques that require multiple reference compounds,[3] fingerprint searching can also be carried out for individual reference molecules. If multiple reference compounds are available, fingerprint averaging techniques or data fusion of multiple Tc values can be applied,[5,6] which typically further increases search performance relative to calculations using single reference molecules.[1,3] Among fingerprint search strategies for multiple reference compounds, nearest neighbor data fusion has become rather popular.[5,6] In *k*-nearest neighbor (*k*-NN) searching, the similarity of a database molecule against each available reference compound is considered, and the values for *k* top-scoring reference compounds are averaged to yield the final similarity score.[5]

MACCS structural keys[7] have long been a prototype of keyed fingerprint designs. The publicly available version of MACCS consists of 166 structural fragments.[7] If a fragment is present in a molecule for which MACCS is calculated, the corresponding bit position is set to "1"; if the fragment is absent, it is set to "0". Similarity searching using this type of fingerprint assigns high similarity values to database compounds that share many structural features with active reference molecules. A possible complication of such fingerprint search calculations arises from so-called molecular complexity effects.[3,8] This means that large and topologically complex molecules generally produce fingerprints of higher bit density than less complex ones.[8] Intrinsically high bit densities result in a statistical tendency to yield increased similarity values in pairwise molecular comparisons, regardless of specific molecular features. Complexity effects can bias fingerprint searching in different ways.[8,9] For example, the use of optimized lead compounds or drug candidates as reference molecules, which are in general more complex than average database compounds, has been shown to represent a particularly difficult search scenario for the identification of novel hits.[9] However, complexity effects can be balanced by use of similarity metrics that equally weight the contributions of "1" and "0" bits in keyed fingerprints[9] (whereas conventional similarity coefficient only take "1" bits into account) or by designing fingerprints that have a constant bit density, regardless of molecular complexity or size.[10]

In addition to keyed fingerprints, other popular fingerprint designs include hashed fingerprints[11] and feature ensembles.[12] In hashed fingerprints, different molecular features or connectivity pathways are "folded" and mapped to overlapping bit segments. Feature ensembles typically represent layered atom environments in molecules that are recorded as individual strings. In contrast to keyed or hashed fingerprints that are of predefined length, feature ensemble methods generate varying numbers of strings for different test molecules.

For key-type fingerprints, chemical interpretation of bit settings is readily possible.[13] For hashed designs, this is much more difficult. For example, by comparing keyed fingerprints of active and inactive compounds or by calculating bit frequency-based fingerprint profiles for different compound sets, bit patterns can be determined that are characteristic of a series of active molecules.[13−15] This information can be utilized, for example, to generate consensus fingerprints that

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

10.1021/ci8002045 CCC: $40.75
Published on Web 08/13/2008

emphasize structural commonalities within sets of active compounds[14] or assign scores to structural keys based on their frequency of occurrence.[15] High-frequency or consensus bits in compound activity classes can also be scaled or weighted in reference fingerprints in order to further improve similarity search performance.[16−18]

While bit patterns in keyed fingerprints have been studied in order to identify consensus bit settings,[14−16] the contribution of individual bit positions to similarity search performance has not yet been systematically analyzed. Here we introduce bit silencing as an approach to determine the contribution of each bit position in a keyed fingerprint. However, the bit silencing technique can in principle also be applied to hashed fingerprint designs or pharmacophore fingerprints. For a given fingerprint and compound activity class, bit silencing makes it possible to derive a bit position-dependent weighting scheme that can then be used to modify similarity metrics in a compound class-specific manner. We design a bit position-dependent weighted variant of the Tanimoto coefficient and systematically analyze its search performance for a variety of compound activity classes. In many instances, the use of the compound class-directed Tc variant is found to increase hit rates of conventional search calculations.

METHODOLOGY

Bit Silencing. Each individual bit position in a keyed fingerprint is systematically set to "0" for all reference compounds prior to similarity searching. Bit silencing as introduced herein differs from modification of fingerprints through removal of individual bit positions. Thus, in bit silencing, the length of fingerprints is kept constant. For a fingerprint with N bits, a total of N search calculations are carried out with variable settings on *(N-1)* bits, except for the silenced bit that is constantly set to "0" and does not contribute to the search. Bit settings in fingerprints might frequently be correlated. It should be noted that silencing of one of several correlated bit positions always affects the Tc calculations and compound ranking. Therefore, correlated bit positions cannot have compensatory effects in bit silencing. We have subjected MACCS keys to the bit silencing procedure. Thus, in each test case, as described below, hit rates were calculated for 166 silencing calculations and recorded in a bit position-dependent hit rate profile.

Weight Vector. From the hit rate profile, a *bit position-dependent weight vector* is calculated on the basis of weights that are assigned to each bit position according to the effects of silencing. If silencing of a bit position leads to a reduction in search performance, the bit makes a positive contribution and is emphasized. By contrast, if silencing of a bit increases search performance, it negatively contributes and is de-emphasized. If silencing has no effect, the bit makes no contribution and is not weighted. Accordingly, the weight vector can be derived as follows: If hr_O is the hit rate obtained with the unmodified fingerprint and (hr_1, hr_2,..., hr_N) are N silenced hit rate values, the weight on the i-th bit, w_i, is defined as

$$w_i = (1 + (hr_O - hr_i) \cdot sf) \cdot 100\%$$

where sf is a predefined scale factor reflecting the magnitude of change observed in the hit rate profile. The higher sf is, the more sensitive the weight vector becomes to fluctuation in hit rates as a consequence of silencing. For example, if sf is set to 100 and silencing of the i-th bit reduces the hit rate by 3%, then $w_i = (1 + 3\% \cdot 100) \cdot 100\% = 400\%$, which means the corresponding bit is scaled 4-fold relative to the original 100% weight because of its positive contribution. With $sf = 200$ and a 3% reduction in hit rate, the value of w_i becomes 700%. By contrast, if silencing of a bit leads to a 2% increase in the hit rate and $sf = 200$, then weight on this bit position becomes −300%, which corresponds to 3-fold negative scaling. The bit position-dependent weight vector W consists of the weights of all N bit positions and mirrors the significance of each individual bit. The calculation of W is fingerprint- and compound class-dependent and influenced by the composition of the reference set.

Weighted Tanimoto Similarity. The weight vector makes it possible to generate a bit position-dependent weighted Tanimoto coefficient. Given two molecular bit vectors of length N, $\boldsymbol{A} = (a_1, a_2,..., a_N)$ and $\boldsymbol{B} = (b_1, b_2,..., b_N)$, the general form of Tc[4] is

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

In this formulation, a_i and b_i are binary variables representing the *i-th* bit in fingerprints A and B, respectively, and $a_i b_i$ represents their product. We add variable weights to each individual bit position corresponding to the results of silencing by calculating the product of the Tc and weight vector W. Thus, given a vector of N elements, $W = (w_1, w_2,..., w_N)$, representing the weights on the N bits of the fingerprint, the bit position-dependent Tc, *bw_Tc*, is defined as

$$bw\_Tc(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{W}) = \frac{\sum_{i=1}^{N} a_i b_i w_i}{\sum_{i=1}^{N} (a_i^2 + b_i^2 - a_i b_i) w_i}$$

The calculation of *bw_Tc* is illustrated in Figure 1. Two hypothetical fingerprints with 10 bits are compared using the conventional Tc and *bw_Tc*. For the latter we used a hypothetical weight vector represented in percentage format. Because negative values are permitted for the weight vector's elements, as discussed above, *bw_Tc* similarity values can also become negative. Thus, compared to Tc-based ranking, larger value ranges and differences between similarity values are possible in *bw_Tc* calculations.

Test Calculations. For bit silencing and systematic similarity search calculations, 21 activity classes were assembled from the MDDR.[19] As a background set, 5000 compounds were randomly taken from ZINC.[20] To ensure that active compounds had properties comparable to the background set, they were filtered applying the ZINC filter rules;[20] i.e., 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. For each activity class, two distinct sets of compounds were taken from the MDDR, a training set for bit silencing and derivation of the weight vectors and a hit set for

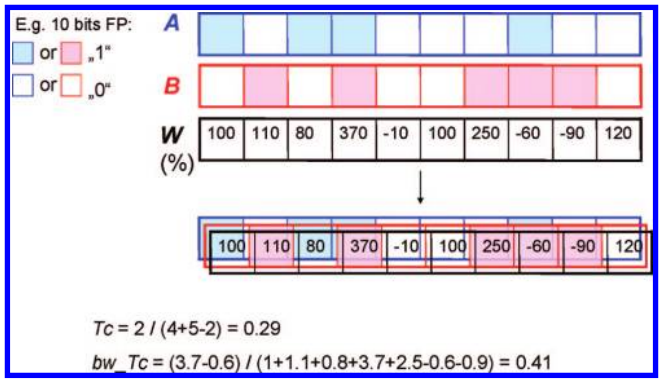

Figure 1. Calculation of the bit position-dependent weighted Tc. Two hypothetical fingerprints consisting of 10 bits each are compared using Tc and *bw_Tc*. The latter value is calculated on the basis of a hypothetical weight vector. In this calculation, the numerator contains the sum of the weights over all bits set to "1" common to A and B and the denominator the sum of the weights on the bits that are set to "1" in either A or B. Positive weights are added, and the absolute values of negative weights are subtracted. In this example, the two hypothetical molecules become more similar when *bw_Tc* is calculated because they share a bit position that makes a significant contribution to search performance, having a relative weight of 370%.

Table 1. Activity Classes

class	training compounds	potential hits	description
ACE	215	30	angiotensin-converting enzyme inhibitor
ADR	250	70	aldose reductase inhibitor
CAM	133	10	cell adhesion molecule antagonist
CLG	146	20	collagenase inhibitor
COX2	122	40	cyclooxygenase-2 inhibitor
COX	102	140	cyclooxygenase inhibitor
ELA	112	10	elastase inhibitor
FXA	605	40	factor Xa inhibitor
HIV	148	50	HIV-1 protease inhibitor
LKT	181	120	leukotriene antagonist
LPO	138	70	lipid peroxidation inhibitor
MM1	178	20	muscarinic M1 agonist
NEP	196	60	neutral endopeptidase inhibitor
PA2	84	100	phospholipase A2 inhibitor
PAF	198	50	platelet-activating factor antagonist
PDV	327	10	phosphodiesterase V inhibitor
PKC	129	70	protein kinase C inhibitor
RTI	177	100	reverse transcriptase inhibitor
SST	99	40	squalene synthetase inhibitor
TKI	253	250	tyrosine-specific protein kinase inhibitor
TNF	185	50	tumor necrosis factor inhibitor

"Training compounds" were used for bit silencing calculations and the derivation of the class-specific bit position-dependent weight vectors and "potential hits" for similarity searching using MACCS Tc and *bw_Tc* calculations. Training and potential hit sets were distinct in each case.

similarity searching using MACCS Tc and *bw_Tc* calculations. The activity classes and composition of training and hit sets are reported in Table 1. The number of training compounds ranged from 84 to 605, and the number of potential hits ranged from 10 to 250. Compound numbers differed in each case because only active molecules with distinct core structures were extracted from the MDDR (in order to avoid potential bias through inclusion of large series of analogs). This was accomplished by applying a scaffold analysis algorithm.[21]

From each training set, 10 different reference subsets of 20 compounds were randomly selected, and the remaining compounds were added to the background molecules for deriving the bit silencing hit rate profile. For each reference set, 166 bit silencing calculations were carried out in combination with 20-NN ranking (to equally take contributions of all reference molecules into account), and hit rates were calculated for the top-ranked 100 database molecules. From these hit rates, weight vectors were obtained for each reference set, and the activity class-specific weight vector was derived by averaging these 10 vectors. The calculations were carried out using three different scale factors (50, 100, and 200) that produced weight vectors of different composition.

The so derived class-specific weight vectors were then used to compare *bw_Tc* calculations with standard MACCS Tc similarity searching and MACCS bit scaling calculations. Therefore, hit sets for each activity class were added to the ZINC background database. Search calculations were carried out as described above for bit silencing. The reference compounds for these search calculations were taken from the training sets. In each case, hit and compound recovery rates were determined for the top-ranked 100 database compounds. Hit rates report the percentage of correctly identified active compounds within database selection sets of predefined size (e.g., 100 molecules), while recovery rates give the percentage of correctly identified active compounds relative to the total number of active molecules available in the database.

RESULTS AND DISCUSSION

The bit silencing technique is introduced in order to systematically evaluate the contribution of individual fingerprint bit positions to similarity search performance and, in addition, enable the derivation of compound class-directed similarity metrics. Bit silencing initially produces a hit rate profile that monitors the change in hit rate as a consequence of modifying each bit position. This profile is then converted into a class-specific bit weight vector that can be combined with similarity metrics such as the Tanimoto coefficient. The approach is evaluated here using MACCS as a prototypic keyed fingerprint.

Effects of Bit Silencing. Our systematic silencing calculations have revealed that individual MACCS bit positions consistently affect the recognition of active compounds. For each of the 21 activity classes, a number of bits were identified whose silencing either increased or reduced hit rates. However, in each case, significant numbers of bits were also identified whose presence or absence had no influence on compound recognition. For example, for one of the reference sets of COX training compounds, MACCS Tc calculations produced a hit rate of 23%. Silencing of 17 of 166 bits reduced this hit rate by 1% to 4%, whereas silencing of 55 other bits resulted in higher hit rates of between 24% and 35%. Thus, silencing of individual bits led to increases in the hit rate of up to 12%, which represents a significant improvement of search performance. In this case, silencing of the remaining 94 bit positions did not change the hit rate. Many of these were "0" bits. These findings illustrate that individual "1" bits can significantly compromise the ability to detect active compounds, which is a noteworthy finding.

Figure 2 shows the bit position-dependent weight distribution of COX calculated with a scale factor of 100 averaged over 10 individual trials. Here 71 bits obtained weights

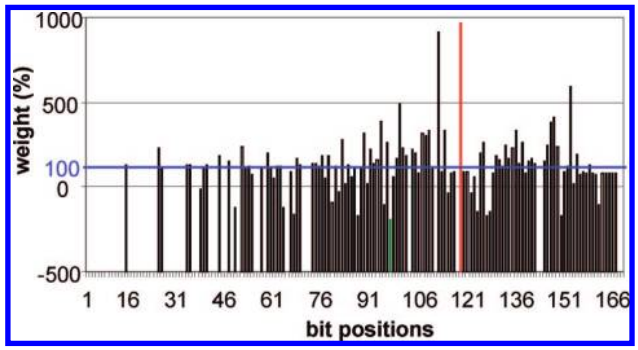

Figure 2. Derivation of a weight vector. Shown is the average bit position-dependent weight distribution of activity class COX generated using a scale factor of 100. Weights of bit positions that increased or decreased the hit rate during silencing are displayed and bits whose silencing did not affect the hit rate (and thus obtained weights of 100%) omitted for clarity. Bit positions with maximum weight (positive scaling due to decrease in hit rate) and minimum weight (negative scaling due to increase in hit rate) are shown in red and green, respectively.

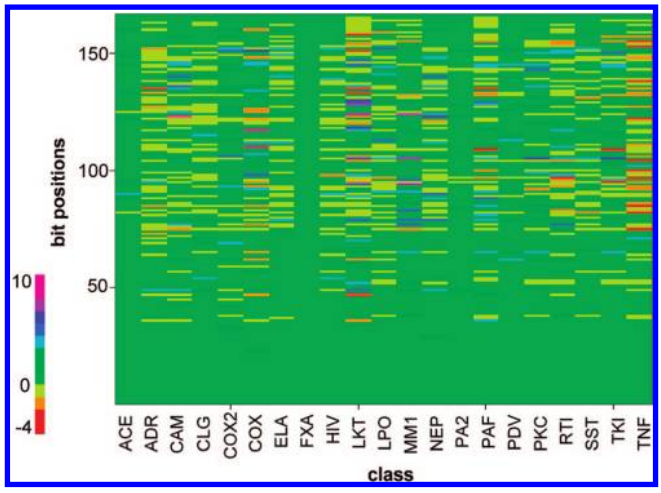

Figure 3. Heat map of bit weight vectors. Average bit weight vectors of the 21 activity classes are represented as a heat map. The different color distributions show that the weights on bit positions are largely class-specific.

greater than 100%, with a maximum of 970%, or 9.7-fold scaling, because silencing of these bits reduced search performance. By contrast, 45 bits had weights of less than 100%, with a minimum of −190%, because their silencing increased hit rate. Silencing of the remaining 50 bits did not affect hit rates (obtaining weights of 100%). These results also illustrate that only subsets of fingerprint bits determine search performance. For COX, nearly one-third of MACCS bit positions did not detectably contribute. Weight vectors of all activity classes are compared in Figure 3. These weight vectors significantly differ in bit positions having highest and lowest weights and are thus class-specific. It is not possible to select MACCS bit positions that are generally associated with different biological activities.

Compound Class-Directed Similarity Coefficients. Because the different effects of bit silencing described above were consistently observed for all 21 activity classes, the derivation of class-directed bit-position dependent similarity metrics was thought to be a promising approach of general relevance. Therefore, we calculated average weight vectors for each class, applied them in systematic compound class-directed *bw_Tc* search calculations, and compared the results with conventional MACCS Tc calculations. Table 2 reports

Table 2. Similarity Search Results[a]

	Tc		*bw_Tc*, *sf* = 50		*bw_Tc*, *sf* = 100		*bw_Tc*, *sf* = 200	
class	HR	RR	HR	RR	HR	RR	HR	RR
ACE	7	23	6	20	8	27	9	30
ADR	6	9	10	14	11	16	6	9
CAM	0	0	4	40	4	40	4	40
CLG	6	30	8	40	8	40	9	45
COX2	5	13	4	10	3	8	3	8
COX	9	6	21	15	20	14	15	11
ELA	0	0	1	10	1	10	2	20
FXA	0	0	0	0	1	3	2	5
HIV	5	10	9	18	9	18	9	18
LKT	6	5	34	28	44	37	39	33
LPO	0	0	6	9	12	17	20	29
MM1	0	0	2	10	2	10	0	0
NEP	24	40	39	65	37	62	34	57
PA2	12	12	12	12	12	12	12	12
PAF	0	0	3	6	5	10	4	8
PDV	0	0	1	10	1	10	1	10
PKC	4	6	15	21	13	19	10	14
RTI	1	1	4	4	6	6	11	11
SST	8	20	10	25	11	28	4	10
TKI	5	2	25	10	40	16	53	21
TNF	0	0	11	22	8	16	1	2
average	5	8	11	19	12	20	12	19

[a] Hit rates (HR) and recovery rates (RR) are reported (in %) for 21 activity classes using conventional Tc and *bw_Tc* calculations with different scale factors (*sf*).

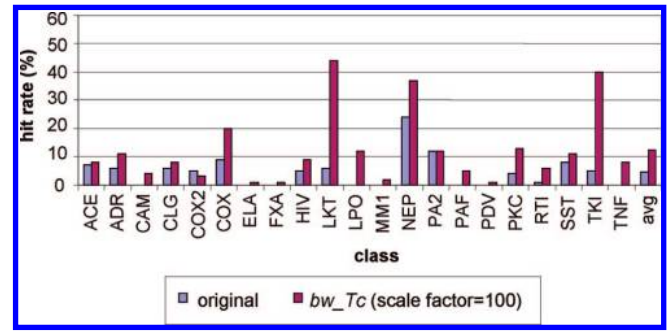

Figure 4. Hit rate comparison. Hit rates for 21 activity classes and the overall average ("avg" are reported for Tc (blue) and *bw_Tc* (red)). In *bw_Tc* calculations, a scale factor of 100 was applied.

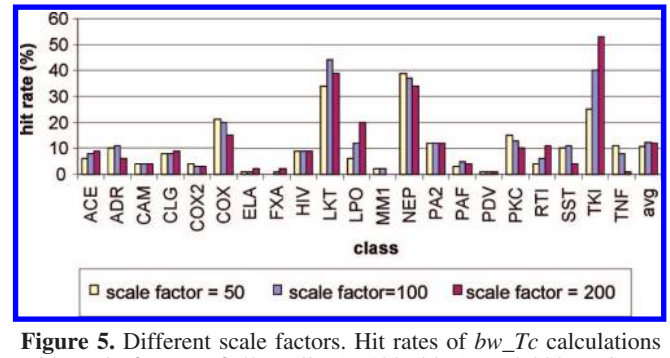

Figure 5. Different scale factors. Hit rates of *bw_Tc* calculations with scale factors of 50 (yellow), 100 (blue), and 200 (red) are reported.

the hit and recovery rates for all test calculations, and Figure 4 shows a graphical comparison of hit rates for Tc and *bw_Tc* calculations using a scale factor of 100. In Figure 5, *bw_Tc* control calculations using different scale factors are reported.

The results in Table 2 and Figure 4 show that the application of *bw_Tc* generally increased hit and recovery rates of conventional MACCS Tc calculations. COX2 was

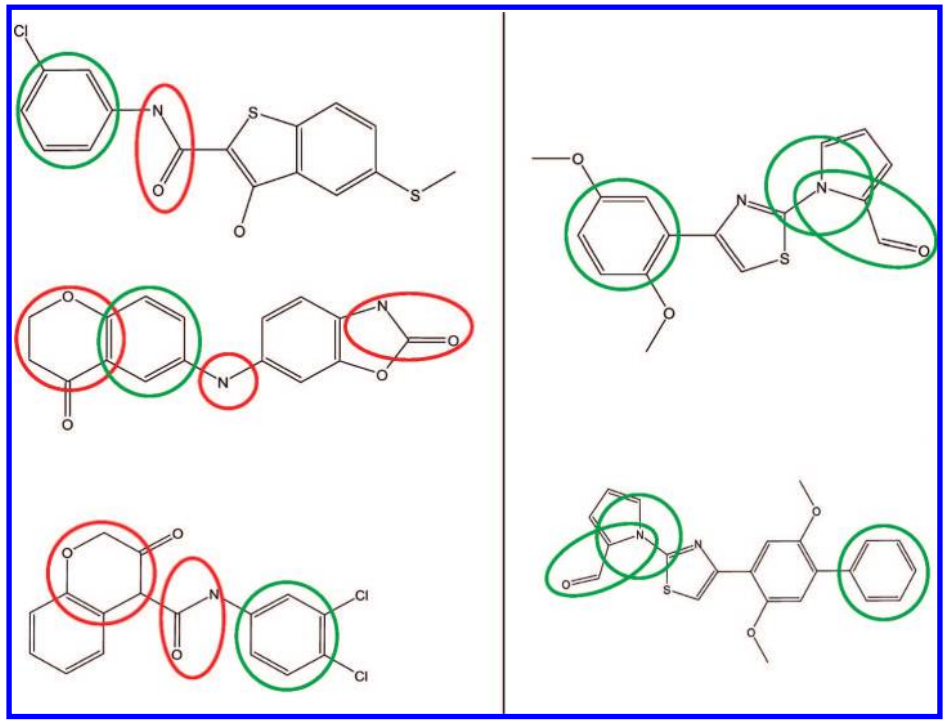

Figure 6. Substructures of COX inhibitors with high and low weights. Shown on the left are examples of COX inhibitors that were correctly identified using the *bw_Tc* metric but not conventional Tc calculations. On the right, ZINC compounds are shown that were found in COX compound selection sets obtained on the basis of Tc calculations but were deselected when the *bw_Tc* metric was applied. Substructures having high and low *bw_Tc* weights are highlighted in red and green, respectively.

the only one of 21 classes for which Tc calculations produced higher rates. The average hit rate over all activity classes increased from 5% for Tc to 12% for *bw_Tc* calculations, and the average recovery rate increased from 8% to 20%. For most classes, applying increasingly large scale factors for the generation of weight vectors did not substantially affect *bw_Tc* search results, as illustrated in Figure 5, i.e. a scale factor of 50 essentially produced results comparable to those obtained with scale factors of 100 or 200. We also carried out test calculations with scale factors of 400 and 800 and generally observed reduced hit and recovery rates under these drastic amplification conditions.

Depending on the activity class, the magnitude of hit rate improvements achieved in *bw_Tc* calculations differed. For eight classes, Tc calculations failed to identify active compounds, but in all of these cases, *bw_Tc* calculations correctly recognized active molecules and achieved hit rates of up to 20% and recovery rates of up to 40% (Table 2). For six of the classes where Tc calculations succeeded, *bw_Tc* hit rate improvements ranged from 5% and 10%, and for six other classes improvements of more than 10% were observed. In some cases, these effects were very significant. For example, for LKT and TKI, Tc calculations produced hit rates of 5% or 6% hit rate, but *bw_Tc* calculations increased these rates to 40% or more (Table 2). Because our compound sets were assembled to contain only inhibitors with unique core structures, increasing hit rates in *bw_Tc* calculations also suggest an increase in the potential of recognizing structurally diverse compounds compared to standard similarity searching. Taken together, these results indicate that compound class-directed evaluation of fingerprint similarity provides a promising alternative to conventional similarity search protocols.

Chemical Interpretation of Bit Significance. Analysis of substructures corresponding to bits obtaining high or low weights in *bw_Tc* calculations makes it possible to interpret the results in a chemically intuitive manner. For example, as illustrated in Figure 6, substructures might be identified that are responsible for the detection of active compounds. COX inhibitors that were correctly identified using *bw_Tc* but not conventional Tc calculations are compared to ZINC compounds that were detected using Tc calculations but deselected by *bw_Tc*. A benzene moiety shared by all compounds is assigned a low *bw_Tc* weight. By contrast, two MACCS keys accounting for an "aliphatic six-membered ring containing a heteroatom" and a "N-X-O" unit detect an oxane substructure and an amide bond, respectively, that occur in the COX inhibitors but not in the ZINC compounds. These substructures are assigned high weights and help to distinguish the COX inhibitors from background database compounds.

Bit Silencing, Consensus Bits, and Bit Scaling. Consensus bits that are mostly or always set on in fingerprints of compounds having similar activity can be scaled in order to emphasize these bit positions during similarity searching.[14,15] We have also determined MACCS consensus bit positions for the activity classes studied here. Silencing of these bit positions revealed that consensus bits were generally not among the most significant bit positions for MACCS search performance. Thus, scaling of these bit positions does not emphasize the most critical bits for each activity class, although scaling calculations were also found to increase recall of active compounds using MACCS keys.[16] Thus, the silencing method should have the principal advantage over consensus bit scaling that the most important bit positions are identified. This conclusion was tested by carrying out systematic bit scaling calculations using MACCS. Consensus bit positions (bits set on in all reference compounds) were identified for each activity class, and systematic similarity search calculations were carried out on the same compound

Table 3. Comparison with Fingerprint Scaling Method[a]

	Tc		*bw_Tc*, *sf* = 100		FP_scaling, *sf* = 3.0	
class	HR	RR	HR	RR	HR	RR
ACE	7	23	8	27	7	23
ADR	6	9	11	16	7	10
CAM	0	0	4	40	0	0
CLG	6	30	8	40	6	30
COX2	5	13	3	8	5	13
COX	9	6	20	14	11	8
ELA	0	0	1	10	0	0
FXA	0	0	1	3	0	0
HIV	5	10	9	18	6	12
LKT	6	5	44	37	6	5
LPO	0	0	12	17	0	0
MM1	0	0	2	10	0	0
NEP	24	40	37	62	24	40
PA2	12	12	12	12	12	12
PAF	0	0	5	10	0	0
PDV	0	0	1	10	0	0
PKC	4	6	13	19	4	6
RTI	1	1	6	6	1	1
SST	8	20	11	28	8	20
TKI	5	2	40	16	5	2
TNF	0	0	8	16	0	0
average	5	8	12	20	5	9

[a] Hit rates (HR) and recovery rates (RR) are reported (in %) for 21 activity classes using conventional Tc, *bw_Tc* calculations with a scale factor of 100 and fingerprint scaling ("FP_scaling") of consensus bit positions with a scale factor of 3.0.

sets used for *bw_Tc* and MACCS Tc calculations applying a scaling factor of 3.0 to all consensus bits.[16] The results are reported in Table 3. Compared to conventional MACCS Tc calculations, fingerprint scaling was found to increase hit and recovery rates for three compound classes. However, *bw_Tc* calculations performed better than fingerprint scaling in 20 of 21 cases (except COX2), consistent with our expectations.

CONCLUDING REMARKS

Previous analyses of bit settings in keyed fingerprints have largely focused on identifying bit positions that are set on with high frequency in compounds having similar activity and attempted to emphasize such positions, for example, through fingerprint scaling or calculation of consensus fingerprints for activity classes. The bit silencing technique, as introduced herein, makes it possible to systematically evaluate positive or negative contributions of all bit positions in keyed fingerprints to similarity searching. Silencing calculations on a large number of activity classes consistently revealed differential contributions of MACCS bit positions. In many instances, individual bit settings were found to substantially increase or decrease search performance. On the basis of these observations, bit position-dependent weight vectors were derived that account for positive or negative contributions of bits and used to modify the Tanimoto coefficient in a compound class-specific manner. However, the results of bit silencing might also be utilized for fingerprint modification by assigning variable scaling factors to all bit positions. Such scaling factors can be derived from class-specific weight vectors. The application of this modified similarity metric was found to generally improve the similarity search performance of conventional Tc and fingerprint scaling calculations. The *bw_Tc* function represents the first compound class-directed similarity coefficient.

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) 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.
(6) Salim, N.; Holliday, J.; Willet, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.
(7) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(8) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Comput. Sci.* **1998**, *38*, 379–386.
(9) Wang, Y.; Bajorath, J. Balancing the Influence of Molecular Complexity on Fingerprint Similarity Searching. *J. Chem. Inf. Model.* **2008**, *48*, 75–84.
(10) 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.
(11) James, C. A.; Weininger, D. *Daylight theory manual*; Daylight Chemical Information Systems Inc.: Aliso Viejo, CA, 2008.
(12) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2008.http://accelrys.com/products/scitegic/ (accessed June 2008).
(13) Godden, J. W.; Stahura, F. L.; Xue, L.; Bajorath, J. Searching for Molecules with Similar Biological Activity: Analysis by Fingerprint Profiling. *Pac. Symp. Biocomput.* **2000**, *5*, 566–575.
(14) 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.
(15) Williams, C. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. *Mol. Diversity* **2006**, *10*, 311–332.
(16) 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.
(17) 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.
(18) Morent, D.; Patterson, D. E.; Berthold, M. R. Towards Context-aware Similarity Metrics In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China 2005*; IEEE Systems, Man and Cybernetics Technical Committee on Cybernetics, Eds.; IEEE Press: Piscataway, NJ, pp 5596−5598.
(19) *Molecular Drug Data Report (MDDR), version 2005. 2*; Symyx Software: San Ramon, CA, 2005.
(20) 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.
(21) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97–102.

CI8002045