Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients

Jeffrey W. Godden,[†] Ling Xue,[†] and Jürgen Bajorath*[,†,‡]

New Chemical Entities, Inc., 18804 North Creek Parkway South, Bothell, Washington 98011, and
Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received September 16, 1999

A combinatorial method was developed to calculate complete distributions of the Tanimoto coefficient (Tc) for binary fingerprint (FP) representations of specified length, regardless of the chemical parameters they reflect. Theoretical Tc distributions were calculated for FPs consisting of up to 67 bit positions which revealed significant statistical preferences of certain Tc values. Calculation of Tc distributions in a large compound database using different FPs mirrored the effects identified by our general analysis. On the basis of these findings, an average Tc is biased by statistically preferred values.

INTRODUCTION

Evaluation of molecular similarity or diversity has become an intensely studied topic in chemistry and pharmaceutical research,[1−5] at least in part spurred on by advances in computational combinatorial chemistry.[4−6] The most popular abstract expressions of molecular structure and properties for such calculations are binary bit string representations called fingerprints (FPs).[7−9] FPs capture molecular features in a binary format, for example, connectivity paths through a molecule,[8] the presence or absence of defined structural fragments,[10] or values of molecular descriptors.[9,11] They can be hashed or folded or, alternatively, keyed (where each bit position is associated with a specific fragment or descriptor value) and may significantly vary in length, from less than a hundred to several thousand bit positions.[8,9]

Molecular similarity/diversity can be assessed in conceptually different ways including a variety of algorithms and high-level descriptions of molecular structure, properties, and conformations.[4−6] Such calculations often involve comparisons of molecular fingerprints. A variety of metrics are available to quantitatively compare FPs and calculate similarity values,[12] which can be affected by problems inherent in encoding chemical properties as discrete bit string representations, as recently demonstrated by Flower.[7] Of these metrics, the Tanimoto coefficient (Tc)[7,12] for pairwise comparison of molecules is probably the most widely used estimator of molecular similarity. This coefficient is defined as Tc $= N_{ab}/(N_a + N_b - N_{ab})$, with N_a being the number of bits set on (i.e., 1) in molecule a, N_b the number of bits set on in molecule b, and N_{ab} the number of bits set on common to both molecules.

Figure 1 illustrates how fingerprints are compared and Tc values calculated. Database searches for compounds similar to a query molecule usually rely on pairwise comparisons and Tc calculations. Moreover, averaged Tc values have frequently been reported, as a first approximation, for large compound collections to estimate and compare their degree of diversity.

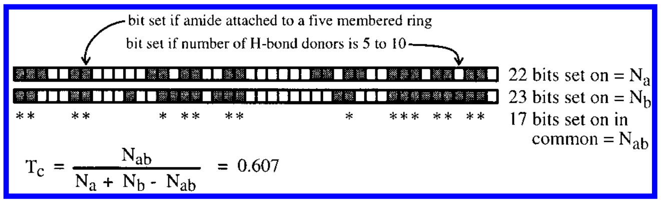
bit set if amide attached to a five membered ring
bit set if number of H-bond donors is 5 to 10
22 bits set on = N_a
23 bits set on = N_b
17 bits set on in common = N_{ab}
$T_c = \frac{N_{ab}}{N_a + N_b - N_{ab}} = 0.607$

Figure 1. Definition and calculation of the Tanimoto coefficient. Two hypothetical fingerprints are shown, with darkened squares symbolizing bits set on (i.e., 1) and light squares bits set off (i.e., 0). In this case, each bit position indicates the presence or absence of a pertinent molecular feature, for example, a structural pattern, or a numerical range of a molecular descriptor (H-bond donors are given as an example). As shown, only bits set on are enumerated and used in Tc calculations. Tc values approaching 1 indicate increasing molecular similarity (identical molecules, or molecules with identical FPs, have a Tc of 1).

We wished to determine if there are statistical effects that influence the results of Tc calculations, regardless of the FPs used and the chemical parameters they reflect. The underlying idea was that combinatorial effects might influence Tc-like comparison of binary bit patterns. We investigated the question of whether there are statistically preferred Tc values that could influence the results of calculations involving many pairwise FP comparisons.

Therefore, we have developed a general combinatorial method to systematically calculate and compare all possible combinations of bit occupancy for an FP of given length. The analysis was designed to be independent of the molecular properties and descriptors encoded in binary fingerprints. The method was applied to calculate complete distributions of Tc values for FPs consisting of up to 67 bit positions. Exhaustive calculations for longer FPs exceed our current computational capacity, due to the increasingly large combinatorial problem.

In our study, we have identified a narrow distribution of strongly preferred Tc's that is dominated by a few discrete

* To whom correspondence should be addressed at New Chemical Entities, Inc. Phone: (425) 424-7297. Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com.

[†] New Chemical Entities, Inc.

[‡] University of Washington.

10.1021/ci990316u CCC: $19.00 © 2000 American Chemical Society
Published on Web 12/02/1999

values. The generality of these findings was tested by analysis of Tc distributions in a large compound database using two fingerprints of very different composition and length. Herein we introduce our method and present the theoretical and database analyses of Tc distributions. We rationalize the results and discuss implications for molecular similarity/diversity calculations.

METHODS

To develop a method to study Tc distributions a priori, we have investigated a situation where each possible binary fingerprint (FP), however defined, is represented by one and only one compound. An analytical expression was developed to calculate the distribution of Tc values for FPs of defined length. Let Tc(i,j) be the Tc for two molecules, i and j, and L the total number of bit positions. The average Tc (avTc) of a hypothetical compound collection with all possible bit combinations is then calculated as

$$\mathrm{avTc} = \frac{\sum_{i=1}^{2^L}\sum_{j=1}^{2^L}\mathrm{Tc}(i,j) - 2^L}{(2^L)^2 - 2^L} \qquad (1)$$

This calculation for all possible FP settings and pairwise Tc-like comparison becomes a large combinatorial problem with increasing number of bit positions. However, we can also calculate the total Tc value (Tc_m) for comparison of each FP with all other FPs in our collection. In this case, avTc of the library is the sum of Tc_m values of all FPs divided by the total number of FP pairs. Moreover, if k is the number of bits set on (i.e., 1), all FPs with the same k also have the same Tc_m value because a comparison with all possible bit settings is made. Thus, all possible FP combinations can be considered for each number of bits set on. Let Tc_m(k) be the Tc_m value for FPs with k bits set on. In this case, the number of compounds with k bits set on is simply the binomial coefficient

$$\binom{L}{k} = \frac{L!}{k!(L-k)!} \qquad (2)$$

and avTc for the whole collection is

$$\mathrm{avTc} = \frac{\sum_{k=0}^{L}\binom{L}{k}\mathrm{Tc\_m}(k)}{(2^L)^2 - 2^L} \qquad (3)$$

Furthermore, suppose an FP with k bits set on is compared to an FP that differs in i bit positions. Within those i bit positions, let j be the number of bits set off (i.e., 0) in the second FP. Then, the Tc for comparison of these two FPs is

$$\mathrm{Tc} = \frac{k-j}{k+i-j} \qquad (4)$$

The number of "compounds" in our collection with such i and j values for an FP with k bits set on is

$$\binom{k}{j}\binom{L-k}{i-j} \qquad (5)$$

The value of i can vary from 1 to L, and the upper limit for j is determined by the minimal value of i and k. The lower limit for j is determined by the relationship $L - k \geq i - j$, so that $j \geq i + k - L$. Therefore, the value of Tc_m for an FP with k bits set on is obtained as follows:

$$\mathrm{Tc\_m}(k) = \sum_{i=1}^{L}\sum_{j=\max[k+i-L,0]}^{\min[k,i]}\binom{k}{j}\binom{L-k}{i-j}\frac{k-j}{k+i-j} \qquad (6)$$

By replacing Tc_m(k) in eq 3 with eq 6, the avTc value for the compound collection is

$$\mathrm{avTc} = \frac{\sum_{k=0}^{L}\binom{L}{k}\sum_{i=1}^{L}\sum_{j=\max[k+i-L,0]}^{\min[k,i]}\binom{k}{j}\binom{L-k}{i-j}\frac{k-j}{k+i-j}}{(2^L)^2 - 2^L} \qquad (7)$$

or, when the equation is rearranged

$$\mathrm{avTc} = \frac{\sum_{k=0}^{L}\sum_{i=1}^{L}\sum_{j=\max[k+i-L,0]}^{\min[k,i]}\binom{L}{k}\binom{k}{j}\binom{L-k}{i-j}\frac{k-j}{k+i-j}}{(2^L)^2 - 2^L} \qquad (8)$$

Equation 8 indicates that there are

$$\binom{L}{k}\binom{k}{j}\binom{L-k}{i-j}$$

Tc values for each

$$\frac{k-j}{k+i-j}$$

Thus, the distribution of Tc values for the FP collection is obtained by plotting all unique Tc values against the sum of this Tc. A flow chart summarizing the approach is shown in Figure 2. The analysis takes advantage of intrinsic symmetries of bit strings that are captured by calculation of Tc values and reduces the computational running time from an $O(2^n)$ order process to an $O(n^3)$ process.

A program for these calculations was generated using SVL[13] code and implemented in MOE.[14] To test the generality of the theoretical findings, random sampling of Tc values for compounds in the ACD[15] database was performed using two conceptually different FPs. A short FP was used, "SSKey_3DS", consisting of only 54 bit positions (accounting for 32 structural keys and numerical ranges of three 2D descriptors),[9] and a long FP, "ph4pd2d", a 2D pharmacophore graph distance-type expression consisting of 1024 bits.[14,16] Both FPs were generated with MOE for all 250 000 compounds in the current ACD database and manipulated using perl scripts. A program was generated to extract all FPs and save them in a compact form, randomly select pairs of different compounds, and calculate their Tc values. Another script was used to accumulate the results and produce a histogram.

RESULTS AND DISCUSSION

The combinatorial method introduced here has made it possible to calculate complete distributions of Tc values for

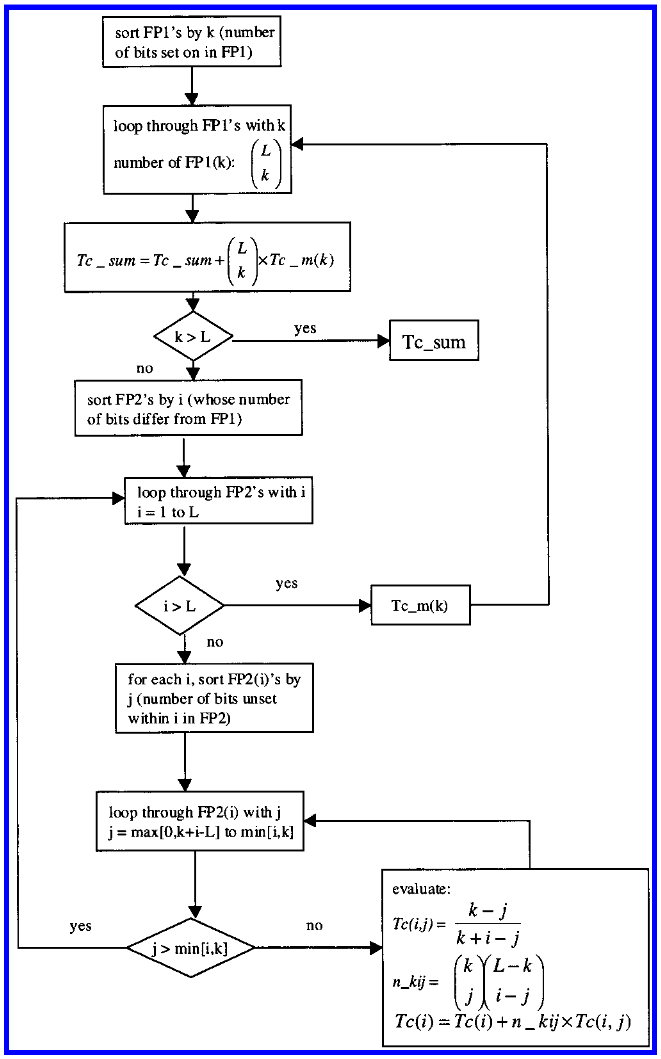
sort FP1's by k (number of bits set on in FP1)
loop through FP1's with k number of FP1(k): $\binom{L}{k}$
$Tc\_sum = Tc\_sum + \binom{L}{k} \times Tc\_m(k)$
k > L
yes
Tc_sum
no
sort FP2's by i (whose number of bits differ from FP1)
loop through FP2's with i i = 1 to L
i > L
yes
Tc_m(k)
no
for each i, sort FP2(i)'s by j (number of bits unset within i in FP2)
loop through FP2(i) with j j = max[0,k+i-L] to min[i,k]
yes
j > min[i,k]
no
evaluate:
$Tc(i,j) = \frac{k-j}{k+i-j}$
$n\_kij = \binom{k}{j}\binom{L-k}{i-j}$
$Tc(i) = Tc(i) + n\_kij \times Tc(i,j)$

Figure 2. Algorithm to systematically calculate Tanimoto coefficients for bit string representations. The diagram summarizes our method for calculating the complete distribution of Tc values for a specified fingerprint (FP). Variables are used as defined in the Methods.

FPs of specified length. An example calculation for an FP consisting of only three bit positions is shown in Figure 3 and the complete Tc distribution for an FP consisting of 54 bit positions in Figure 4. This distribution is representative of FPs that approach our current calculation limit (maximum 67 bits). The large number of occurrences (e.g., greater than 4*e*+32 for Tc = 1/3) illustrates the magnitude of the combinatorial problem. The histogram shows that the Tc distribution is discontinuous and that Tc values between ~0.15 and ~0.55 are prevalent. The Tc interval between 0.3 and 0.4 is most populated and 0.33... is by far the most frequently occurring Tc. This value also represents the average Tc for this calculation (as one may expect considering the shape of the distribution and the dominance of Tc = 1/3). The distribution also shows other discrete peaks at ratios of small integers, e.g., at Tc values of 0.25, 0.40, and 0.50. Thus, a relatively narrow range of Tc values is statistically preferred, due to general characteristics of binary FPs and Tc calculations.

Can we rationalize these observations? The fact that Tc is a ratio of relatively small integers means that some floating-point values are unattainable while others are combinatorially highly favored and thus frequently observed.

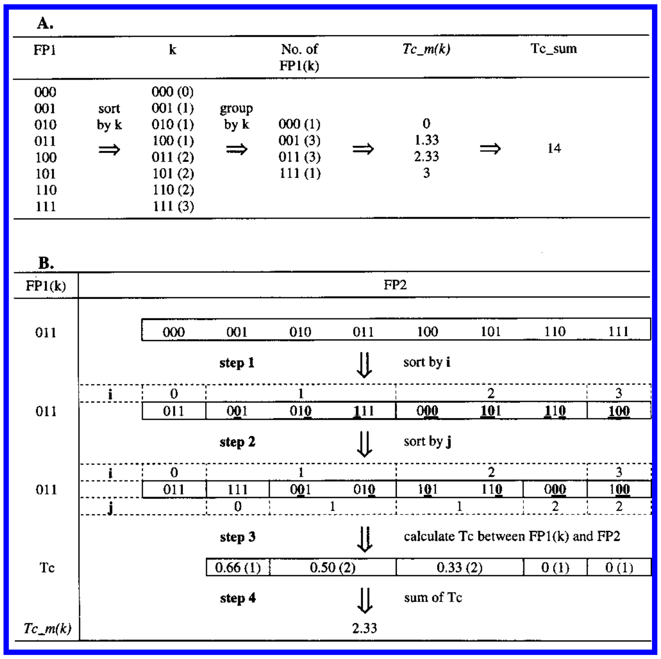

Figure 3. Example calculation for a fingerprint consisting of three bit positions. (A) All possible bit patterns are sorted and grouped by the number of bits set on. FPs representative of each group are termed FP1. For each group, Tc₋m(*k*) values are calculated and summed. In (B), the calculation of Tc₋m(*k*) values is illustrated for one of the four FPs (*k* = 2) and all possible three-bit FPs (termed FP2). Step 1: Sort FP2s by *i* (bits that differ from FP1). Step 2: Sort and group by *j* (*i* bits that are unset, i.e., equal to 0). Step 3: Calculate Tc values and their frequencies (in parentheses) for FP1 (*k*) and FP2 (*i*,*j*). Step 4: Set Tc₋m(*k*) to the sum of Tc weighted by frequencies.

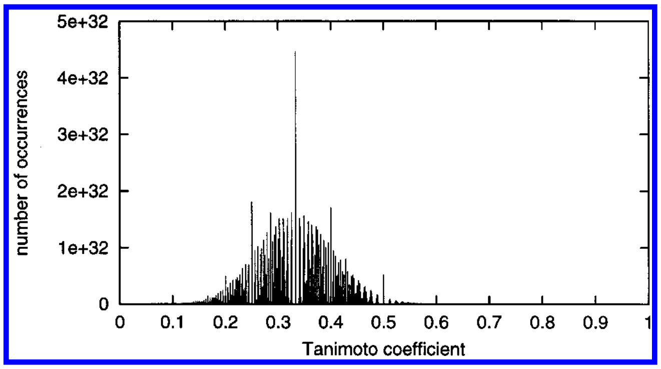

Figure 4. Distribution of Tanimoto coefficients for a fingerprint with 54 bit positions. All possible Tc values were calculated as outlined in Figures 2 and 3. A narrow range of Tc values is strongly preferred, with Tc = 1/3 dominating the distribution.

The sequence of possible Tc values is discontinuous, and a minimal increment between accessible values is the reciprocal of the number of bits. For example, for a long FP comprised of 1056 bits, there is no combination of N_a, N_b, and N_{ab} that can produce a Tc = 0.501 (±0.000 25), whereas there are a very large number of combinations (approximately 2.3*e*+181) that will produce exactly 0.500. Thus, the range of possible Tc values for any FP is a discontinuous function. A graph of the number of possible combinations that yield a particular Tc value would produce a modified version of the "ruler pattern" of rational number theory. For example, for all possible integers *a* and *b* smaller than or equal to 64, there are 21 combinations of *a*/*b* that generate exactly 1/3, only one combination that produces exactly 0.34 (17/50), and none that generate 0.341.

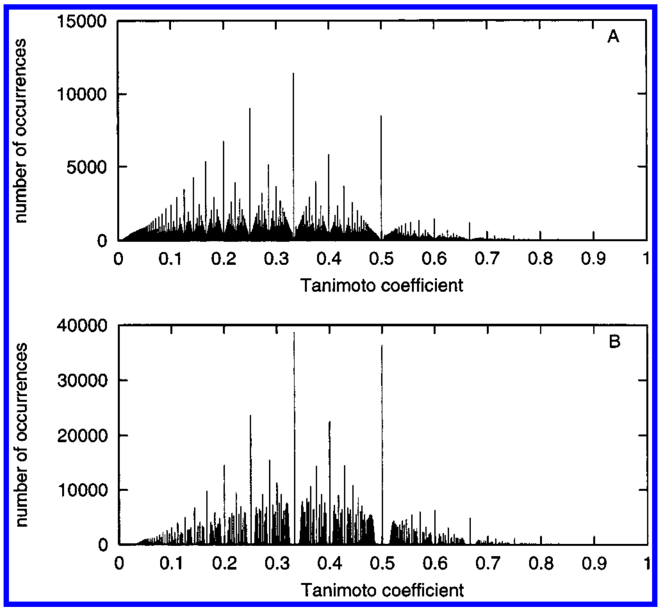

Figure 5. Distribution of Tanimoto coefficients for compounds in the ACD database. Two different fingerprints were applied: (A) "ph4pd2d" (1024 bits) and (B) "SSKey_3DS" (54 bits).

To extend our theoretical analysis, we have determined Tc distributions in a large compound collection, the ACD chemical database. Tc values for ACD molecules were calculated using two FPs that differ significantly in length. For each FP, one million pairs of (nonidentical) molecules were randomly selected, and Tc values were calculated. The average number of bits set on for the 1024 bit FP "ph4pd2d" was 51.2, and the corresponding average for the 54 bit FP "SSKey_3DS" was 16.2. The Tc distributions are shown in Figure 5. Both distributions resemble the theoretical Tc profile in Figure 4. This illustrates the influence of general combinatorial effects, regardless of the FPs used. Peaks in the Tc distributions for ACD compounds also match those observed in the general analysis, emphasizing the statistical preference of a few Tc values. The number of Tc occurrences is much larger in Figure 5B than in Figure 5A because the short FP "SSKey_3DS" has fewer possible bit combinations and thus limits the number of accessible Tc values.

What are the implications of these findings for molecular similarity/diversity calculations? The statistically preferred range of Tc values identified in Figure 4 is outside those regions that have been determined to indicate structure−function similarities of molecules, e.g., 0.7 or greater[9] or 0.85.[17] According to the Tc distributions, FP comparisons within this range of high Tc values should not be significantly influenced by chance occurrences. It should be noted that these considerations do not apply to any transformations of Tc (that would effectively change the metric). This is illustrated in Figure 6 where a theoretical distribution was calculated for the cube root of Tc (rather than Tc). As to be expected, combinatorial preferences are also observed, but in this case, the distribution has a different shape and is shifted toward higher values. However, the Tc distributions in Figures 4 and 5 have implications for the calculation of average Tc values that are frequently used to characterize or compare the level of diversity in large compound collections. Our findings suggest that such calculations are biased by statistically preferred Tc values.

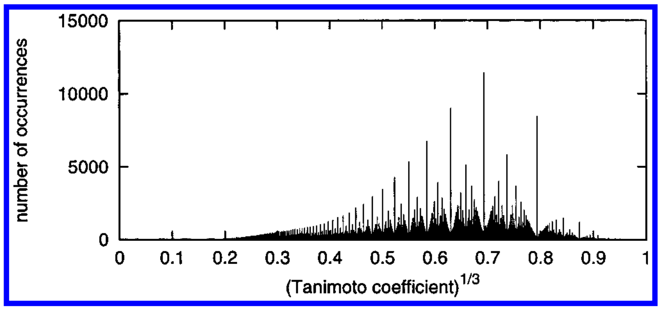

Figure 6. Distribution of modified Tanimoto coefficients for a model fingerprint with 54 bit positions. Different from Figure 4, the cube root of Tc was calculated, which caused a shift of the distribution.

CONCLUSION

The analysis presented herein was carried out to better understand principal characteristics of Tc calculations, a widely used technique to study molecular similarity/diversity. The results of our theoretical and compound database analysis revealed general preferences of a subset of Tc values, and these observations could be rationalized considering the combinatorial features inherent in Tc calculations. For example, average Tc values around 1/3 are statistically highly preferred and may often be observed for any ensemble of unrelated compounds.

REFERENCES AND NOTES

(1) Downs, G. M.; Willett, P. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* **1997**, *7*, 1−66.

(2) Dean, P. M., Ed. *Molecular similarity in drug design*; Chapman and Hall: Glasgow, 1994.

(3) Johnson, M., Maggiora, G. M., Eds. *Concepts and applications of molecular similarity*; Wiley: New York, 1990.

(4) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.

(5) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Des.* **1998**, *9*, 339−353.

(6) Mason, J. S.; Hermsmeier, M. A. Diversity assessment. *Curr. Opin. Chem. Biol.* **1999**, 342−349.

(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) James, C. A.; Weininger, D. Daylight theory manual, Daylight Chemical Information Systems, Inc. (URL: www.daylight.com), Irvine, CA, 1995.

(9) Xue, L.; Godden, J. W.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.

(10) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(11) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(12) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 983−996.

(13) Santavy, M.; Labute, P. SVL: The scientific vector language. *J. Chem. Comput. Group* (URL: www.chemcomp.com/feature/svl.htm), **1998**.

(14) MOE (Molecular Operating Environment) Chemical Computing Group, Inc. (URL: www.chemcomp.com), 1255 University St., Montreal, Quebec, Canada H3B 3X3, 1998.

(15) MDL Information Systems, Inc., 14600 Catalina St., San Leandro, CA 94577.

(16) Sheridan, R. P.; Bush, B. L. "Patty": A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.

(17) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of molecular descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

CI990316U