# Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval

S. Joshua Swamidass[†] and Pierre Baldi*[,†,‡]

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, and Department of Biological Chemistry, University of California, Irvine, Irvine, California 92697-3435

Received November 20, 2006

In many modern chemoinformatics systems, molecules are represented by long binary fingerprint vectors recording the presence or absence of particular features or substructures, such as labeled paths or trees, in the molecular graphs. These long fingerprints are often compressed to much shorter fingerprints using a simple modulo operation. As the length of the fingerprints decreases, their typical density and overlap tend to increase, and so does any similarity measure based on overlap, such as the widely used Tanimoto similarity. Here we show that this correlation between shorter fingerprints and higher similarity can be thought of as a *systematic* error introduced by the fingerprint folding algorithm and that this systematic error can be corrected mathematically. More precisely, given two molecules and their compressed fingerprints of a given length, we show how a better estimate of their uncompressed overlap, hence of their similarity, can be derived to correct for this bias. We show how the correction can be implemented not only for the Tanimoto measure but also for all other commonly used measures. Experiments on various data sets and fingerprint sizes demonstrate how, with a negligible computational overhead, the correction noticeably improves the sensitivity and specificity of chemical retrieval.

## 1. INTRODUCTION

One of the most fundamental tasks of chemoinformatics is the rapid search of large repositories of molecules containing millions of compounds, such as PubChem, ZINC,[1] or ChemDB.[2] In a typical task, given a query molecule, one is interested in retrieving all the molecules in the repository that are similar to the query. To facilitate this task, one of the most practical and widely used computer representation for molecules is the binary fingerprint or binary feature vector representation[3−7] (and references therein), whereby a molecule is represented by a binary vector recording the presence or absence of particular functional groups, features, or substructures. It is these fingerprints and the associated similarity measures,[8−11] such as the Tanimoto measure, that are used for efficiently searching large repositories.

In early chemoinformatics, systems as well as in some of the current applications, these binary feature vectors are relatively short, with typically a few dozen components associated with a small basis of more or less hand-picked features derived mostly from expert chemical knowledge. In most modern systems, however, the major trend is toward the combinatorial construction of very long feature vectors associated with, for instance, all possible labeled paths up to a certain length. The advantage of these much longer representations is two-fold: they do not rely on expert knowledge which may be incomplete or unavailable, and they can support extremely large numbers of molecules, such as those that are starting to become available in public repositories and commercial catalogs, as well as the recursively enumerable space of virtual molecules.[12]

In most modern chemoinformatics systems, these long vector representations are in turn compressed to shorter binary fingerprint vectors, of fixed or variable length, using a simple folding operation described in detail in the next section. The advantage of the compression is that it yields more compact binary representations that require less storage space and can be searched faster than the uncompressed version. The drawback of the compression, however, is that information is lost, and, therefore, when similarity between molecules is measured by similarity between their compressed representations, retrieval quality deteriorates, and increasingly so, as the length of the fingerprint is reduced. Specifically, as discussed in ref 4, as the length of the fingerprints decreases, their typical density tends to increase. As their typical density increases, their typical overlap tends to increase, and so does any similarity measure based on overlap, such as the Tanimoto similarity. Flower concludes that this relationship between density and similarity can be problematic and similarity computed on compressed chemical fingerprints is at best a relative measure of similarity, which is highly dependent on an *arbitrarily* chosen fingerprint length.

Here we show that the correlation between shorter fingerprints and higher similarity can be thought of as a *systematic* error introduced by the fingerprint folding algorithm and that this systematic error can be corrected. More precisely, given two molecules and their compressed fingerprints of a given length, we show how a better estimate of their uncompressed overlap, hence of their similarity, can be derived to correct for this bias. We show how this correction can be implemented not only for the Tanimoto

* Corresponding author e-mail: pfbaldi@ics.uci.edu.

[†] Institute for Genomics and Bioinformatics, School of Information and Computer Sciences.

[‡] Department of Biological Chemistry.

10.1021/ci600526a CCC: $37.00
Published on Web 04/20/2007

measure, but also for all other commonly used measures and how the correction improves the sensitivity and specificity of chemical retrieval.

## 2. FINGERPRINTS, COMPRESSION, AND SIMILARITY MEASURES

**2.1. Fingerprints.** We use $\mathscr{A}$ to denote a molecule. We assume that molecules are represented by feature vectors, or fingerprints, of length $N_*$, denoted by $\vec{A}_* = (A_i)$. The precise interpretation of these fingerprints is irrelevant for our purpose, but, to fix the ideas, in the binary case the reader may think of each component $A_i$ as a 0−1 variable associated with the presence or absence of a labeled path of labeled atoms and bonds present in the molecule $\mathscr{A}$. In the nonbinary case, each component may correspond to the number of features of a certain type present in a molecule. The value of $N_*$ depends on the features, for instance the depth of the paths, but a typical value could be $N_* = 2^n$ with $n$ in the 15−20 range, corresponding to a 32,768−1,048,576 range for $N_*$. In fact, while it is often practical to have $N_*$ be a power of 2, this is not necessary, and the theory and results to be presented apply equally well to other values of $N_*$. In fact, we do not even distinguish between the cases where $N_*$ corresponds to the total number of possible features of a certain type or to the total number of observed features of a certain type in a given database of molecules. Furthermore, for a typical user, the value of $N_*$ is often unknown. Thus in some theoretical derivations we will first derive formulas that contain $N_*$ and then show how one can dispense from such knowledge when $N_*$ is large.

**2.2. Fixed-Length Compression.** Long fingerprints are often compressed, or "folded'', using a simple modulo operator, down to fingerprints $\vec{A}_N$ of length $N$. The main requirement for standard $k$-fold compression to fingerprints of length $N$, as described below, is that $N_* = Nk$. Obviously, long fingerprints can always be padded with zeroes to achieve this relationship. Typically, in current chemoinformatics systems, $N = 2^9 = 512$ or $N = 2^{10} = 1024$. Ultimately, it is these short fingerprints that are used to derive similarity measures and to rapidly search large databases of molecules. In the binary case, a bit in position $j$ of the compressed fingerprint is set to 1 if and only if there is at least one bit set to 1 in position $j$ modulo $N$ in the full fingerprint of length $N_*$. Figure 1 illustrates the simple process of folding a binary fingerprint vector of size $N_* = 16$ into a compressed fingerprint vector of size $N = 4$ using a modulo operator.

While in some applications it may be possible to exploit or weigh information associated with specific components, the compression modulo $N$ is most effective only if all the bits are treated equally, so that the specific ordering of the bits becomes irrelevant. In practice, this requires applying a fixed but random permutation to all the fingerprints of length $N_*$ prior to compression or using a good hashing function to derive "randomized'' fingerprints of length $N$.

In some systems,[5] more than one bit (typically, 2, 3, or 4) are set to 1 in the compressed vector, for every bit set to 1 in the uncompressed vector. This variation is equivalent to concatenating a fixed number of replicates of the uncompressed vector before applying the random permutation and then folding this expanded, redundant vector to the appropriate size. This method uses redundancy to decrease the unsystematic error of compression at the cost of doubling, tripling, or quadrupling both encoding space and query time while still suffering (perhaps increasingly so) from the same systematic error we aim to correct.

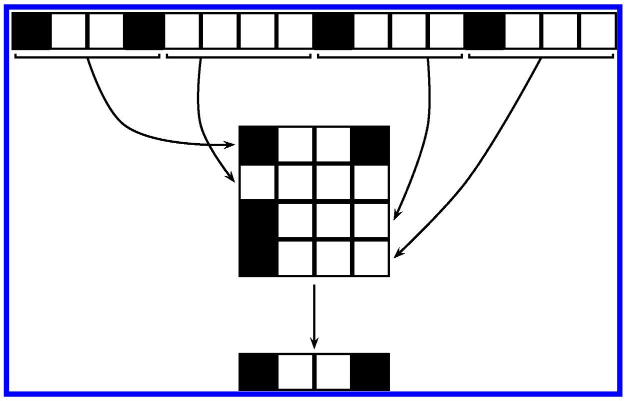

**Figure 1.** Illustration of the folding process with a binary vector of length $N_* = 16$ (1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0) folded into a binary vector of length $N = 4$ (1 0 0 1), modulo 4.

In contrast, the correction method we describe decreases systematic error, both negligibly impacting query time and without using additional space to encode fingerprints. Rather than viewing these two methods as competing solutions to the same problem, we see them as complementary solutions to different problems. Both methods are entirely compatible with each other and can be used together in the same chemoinformatics system, without modification, to simultaneously reduce systematic and unsystematic error.

For each molecule $\mathscr{A}$, we use $A_*$ to denote the corresponding number of bits set to 1 in the full fingerprint and $A_N$ to denote the number of bits set to 1 in the corresponding compressed fingerprint. We drop the index $N$ whenever the context clearly indicates that we are considering fingerprints of fixed length $N$. The corresponding count and projection (compression) operators are denoted by $c$ and $p$ so that $A_* = c(\vec{A}_*)$, $A = c(\vec{A})$, and $\vec{A}_N = p(\vec{A}_*)$. We use Greek letters to denote the bit density $\alpha_* = A_*/N_*$ and $\alpha = \alpha_N = A/N$. For binary compressed fingerprints, we let $\vec{A} \cup \vec{B}$ and $\vec{A} \cap \vec{B}$ denote the corresponding union (OR) and intersection (AND) binary fingerprints, and similarly for uncompressed fingerprints. We also denote $c(\vec{A} \cup \vec{B})$ by $A \cup B$, and $c(\vec{A} \cap \vec{B})$ by $A \cap B$, and similarly for uncompressed fingerprints.

**2.3. Variable-Length Compression.** It is also possible to use variable-length compression,[5] where the length of the encoding of a molecule depends on the molecule itself. This is done by setting a threshold $\alpha$ and repeatedly folding the fingerprint vector $\vec{A}_*$ until the density of the compressed fingerprint exceeds the value $\alpha$. In this case, $\vec{A}_*$ is represented by a compressed fingerprint of length $N(\vec{A}_*)$, where $N(\vec{A}_*)$ varies across molecules, such that $N_* = N(\vec{A}_*)k(\vec{A}_*)$ and $A_{N(\vec{A}_*)}/N(\vec{A}_*) \geq \alpha$, the latter inequality being violated if $\vec{A}_*$ is folded less than $k(\vec{A}_*)$ times. Thus the compressed fingerprint is the shortest compressed vector that satisfies the density inequality. In the following derivations, compressed fingerprints of fixed length are used by default. However, in the Appendixes, we also show how the same principles can be applied to compressed fingerprints of variable lengths.

**2.4. Similarity Measures.** There are several fingerprint similarity measures that can be used to search large repositories of molecules represented by their fingerprints. Given two uncompressed fingerprints vectors $\vec{A}_*$ and $\vec{B}_*$ these measures are usually defined in terms of the bit counts $A_*$, $B_*$, $A_* \cup B_*$, and $A_* \cap B_*$ as well as $N_*$ (see Appendixes). Here we demonstrate our methods using the most widely used similarity measure, the Tanimoto measure, which corresponds to the ratio of the number of common bits set to 1 to the total number of bits set to 1

$$T_* = (A_* \cap B_*)/(A_* \cup B_*) \qquad (1)$$

However, the same ideas can be applied to the other similarity measures found in the literature, as described in the Appendixes.

When fingerprints are compressed, it is customary to apply the Tanimoto measure to the compressed fingerprints in the form

$$T = (A \cap B)/(A \cup B) \qquad (2)$$

The fundamental point of this paper is to show that $T$ is not the best possible estimate of $T_*$—in particular, $T$ tends to overestimate $T_*$, and that a mathematical correction can be applied to derive a better similarity value and estimate of $T_*$. In turn, this correction improves chemical retrieval accuracy.

## 3. DATA

Before we proceed with the mathematical derivations, we describe the data used in the corroborating simulations. To assess our results, we use small molecules in the ChemDB database.[2] For illustration purposes, the results reported here are obtained using a large random sample of 50,000 molecules from ChemDB. Fingerprints are associated with labeled paths of length up to 8 (i.e., 9 atoms and 8 bonds). In this case, the total number of observed labeled paths is $N_* = 152{,}087$. Compression is done using a simple modulo operator. Most results are reported for fingerprints of length $N = 512$. However we have tested all values $N = 2^n$, with $5 \leq n \leq 10$, and report the corresponding results when the dependence on fingerprint length is relevant. Robust results are obtained by increasing the path length, varying $N_*$, or $N$, or varying the random sample. All fingerprints are computed using an in-house program written in Python. The algorithm used to compress these fingerprints is exactly equivalent to the algorithms described in the documentation of commercial fingerprint systems (e.g., Daylight, Avalon, and Unity). When the same modulo compression algorithm is used, the exact details of the fingerprinting schemes do not affect the analysis or application of our method.

To assess biochemical relevance, we also use the six data sets in ref 14 which correspond to six groups of diverse molecules with similar activity. The molecules of each group are known to interact with the same protein. These data sets consist of 128 chemicals which interact with Cox-2, 55 which interact with estrogen receptor, 43 which interact with gelatinase-A, 17 which interact with neuraminidase, 25 which interact with p38-MAP kinase, and 67 which interact with thrombin. All data sets are available upon request.

## 4. MATHEMATICAL CORRECTION OF SIMILARITY MEASURES

To estimate similarity between uncompressed vectors from the observed compressed vectors, one needs first to understand how to derive good estimates of $A_*$ given the compressed count $A$, that is the expected value $E(A_*|A)$. Since the relationship between $A_*$ and $A$ is nondeterministic, we need to define an underlying probabilistic model and set of assumptions.

**4.1. Estimation of $A$ Given $A_*$ and of $A_*$ Given $A$.** As in the case of random graph theory,[14] starting from a given value $A_*$, we can consider two slightly different, but asymptotically, equivalent, generative models for $\vec{A}_*$: fixed density and fixed size uniform models. In the fixed density model, bits are produced by independent identically distributed coin flips with probability $\alpha = A_*/N_*$ of success. In the fixed size model, a subset of $A_*$ components is selected uniformly among all possible subsets of $A_*$ components in the long fingerprint. The bits in the corresponding subset are set to 1, and the bits in the complement are set to 0. The fixed size model generates vectors with exactly $A_*$ bits set to 1 with a uniform distribution consistent with a random bit permutation. In the Appendix, we briefly describe how the fixed size model can be treated exactly. However, the fixed density binomial approximation is more readily tractable and leads to formula that are very effective for most practical purposes. Furthermore, both models yield the same results asymptotically.

In the binomial model, the probability of setting a given bit to 0 in the compressed $\vec{A}$ is $(1 - \alpha)^k$, where $k = N_*/N$. Therefore, the corresponding distribution $P(A|\alpha)$ for $A$ is also a binomial distribution $\mathscr{B}(N,p)$ with $p = 1 - (1 - \alpha)^k$. As a result, given $\alpha$ (or $A_*$), we can estimate $A$ by

$$E(A|\alpha) = N[1 - (1 - \alpha)^k] \qquad (3)$$

or

$$E(A|A_*) \approx N\left[1 - \left(1 - \frac{A_*}{N_*}\right)^{N_*/N}\right] \qquad (4)$$

Conversely, by inverting this relationship, given $A$ we can estimate $A_*$ by

$$E(A_*|A) \approx N_*\left[1 - \left(1 - \frac{A}{N}\right)^{N/N_*}\right] \qquad (5)$$

Equation 5 corresponds to an additional level of approximation since it depends both on the binomial approximation contained in eq 4 and the algebraic inversion of eq 4 to estimate $E(A_*|A)$, rather than using Bayes theorem and computing the corresponding expectation. As we shall see, in practice both approximations work well for our purposes.

All these expressions have nice limits when $N_*$ is large, which is important in practical applications because the exact value of $N_*$ is not always available. When $N_*$ is large, we have

$$\lim_{N_* \to \infty} [1 - (1 - \alpha)^k] = \lim_{N_* \to \infty} \left[1 - \left(1 - \frac{A_*}{N_*}\right)^{N_*/N}\right] = 1 - e^{-A_*/N} \qquad (6)$$

which does not depend on $N_*$, hence the irrelevance of its exact value. Thus given $A_*$, for $N_*$ large, $A$ is approximately binomial with $\mathscr{B}(N,1 - e^{-A_*/N})$ and

$$E(A|A_*) \approx N(1 - e^{-A_*/N}) \tag{7}$$

From the binomial and exponential series expansions, it is easy to see that the estimate in eq 7 is always smaller than the estimate in eq 4. By inverting the one-to-one function in eq 7, we also get an estimate of $A_*$ given $A$ for large $N_*$

$$E(A_*|A) \approx -N \log\left(1 - \frac{A}{N}\right) \tag{8}$$

Note that this estimate is not valid when $A$ is close to $N$, since it yields a diverging value. In fact, we must always have $A_* \le kA \le kN$. Thus for the formula to be applicable we must have $A \le N(1 - e^{kA/N})$ and $N_*$ large. When $A$ is close to $N$, we can always apply eq 5. When $A$ approaches $N$, however, the estimate of $A_*$ given by eq 5 approaches $N_*$, which in many cases is excessive. In practice, this is not problematic because in current databases containing millions of records and fingerprints of length $N = 512$ or $N = 1024$, the case $A = N$ never occurs, and eqs 4 and 5 and their asymptotic approximations given by eqs 7 and 8 give excellent estimates, as shown in Figures 2 and 3.

The binomial model for the conditional distribution $P(A|A_*)$ is thus a reasonable model that yields good first-order predictions. For a given value of $A_*$, the model also predicts the variance

$$\mathrm{Var}(A|\alpha) = N[1 - (1 - \alpha)^k](1 - \alpha)^k \tag{9}$$

with the asymptotic form

$$Var(A|A_*) \approx N(1 - e^{-A_*/N})e^{-A_*/N} \tag{10}$$

Both empirically and theoretically, it is easy to see that eqs 9 and 10 tend to overestimates the variance, because instead of holding $A_*$ fixed, they hold the *density* $A_*/N_*$ fixed and therefore looks, at a range for values of $A_*$. However, this is of little concern here since the correction formulas to be derived do not rely on any estimates of the variance.

**4.2. Estimation of $A_* \cap B_*$ from $A$, $B$, and $A \cup B$.** Most fingerprint similarity measures, such as the Tanimoto measure, are expressed in terms of $A$, $B$, $A \cup B$, and $A \cap B$. If the goal is to estimate the similarity value on the uncompressed fingerprints from the compressed values, then eqs 5, and 8, when $N_*$ is large, provide a mean for estimating the values of $A_*$, $B_*$, and $A_* \cup B_*$ from $A$, $B$, and $A \cup B$. For instance, $A_* \cup B_*$ which can be estimated directly from $A \cup B$ using eq 5 by

$$A_* \cup B_* \approx N_*\left[1 - \left(1 - \frac{A \cup B}{N}\right)^{1/k}\right] \tag{11}$$

or, for large $N_*$,

$$A_* \cup B_* \approx -N \log\left(1 - \frac{A \cup B}{N}\right) \tag{12}$$

The fundamental point, however, is that we cannot apply eqs 5 or 8 directly to recover an estimate of $A_* \cap B_*$ from $A \cap B$. Doing so would lead to overestimating $A_* \cap B_*$: some bits in $A \cap B$ are set to 1 by chance and do not correspond to a compression of bits present in the uncompressed intersection vector $\vec{A}_* \cap \vec{B}_*$. However, we can use the identity $A_* \cup B_* = A_* + B_* - (A_* \cap B_*)$ to derive better estimates of $A_* \cap B_*$ in the form

$$A_* \cap B_* \approx A_* + B_* - N_*\left[1 - \left(1 - \frac{A \cup B}{N}\right)^{1/k}\right] \tag{13}$$

$$\approx N_*\left[1 - \left(1 - \frac{A}{N}\right)^{1/k}\right] + N_*\left[1 - \left(1 - \frac{B}{N}\right)^{1/k}\right] - N_*\left[1 - \left(1 - \frac{A \cup B}{N}\right)^{1/k}\right] \tag{14}$$

$$= N_*\left[1 - \left(1 - \frac{A}{N}\right)^{1/k} - \left(1 - \frac{B}{N}\right)^{1/k} + \left(1 - \frac{A \cup B}{N}\right)^{1/k}\right] \tag{15}$$

and for large $N_*$

$$A_* \cap B_* \approx A_* + B_* + N \log\left(1 - \frac{A \cup B}{N}\right) \approx N \log \frac{\left(1 - \frac{A \cup B}{N}\right)}{\left(1 - \frac{A}{N}\right)\left(1 - \frac{B}{N}\right)} \tag{16}$$

Note that while the relationship $A_* \cap B_* = A_* + B_* - (A_* \cup B_*)$ is exact, when approximate values are substituted for each one of its terms, the right-hand side of eqs 15 and 16 can occasionally become negative. This is particularly true when $A$ and $B$ are large and $A \cap B$ is small, e.g., when A and B are highly dissimilar. In these rare cases, which are not important when the main goal is to retrieve molecules similar to the query, we set the estimate of $A_* \cap B_*$ to 0 and, correspondingly, the estimate of $A_* \cup B_*$ to $-N \log(1 - (A/N)) - N \log(1 - (B/N))$.

In summary, starting from $A$, $B$, $A \cup B$, and $N$ we can derive for large $N_*$ the following estimates:

$$A_* \approx -N \log\left(1 - \frac{A}{N}\right) \tag{17}$$

$$B_* \approx -N \log\left(1 - \frac{B}{N}\right) \tag{18}$$

$$A_* \cup B_* \approx \min\left[-N \log\left(1 - \frac{A \cup B}{N}\right), -N \log\left(1 - \frac{A}{N}\right)\left(1 - \frac{B}{N}\right)\right] \tag{19}$$

$$A_* \cap B_* \approx \max\left[N \log \frac{\left(1 - \frac{A \cup B}{N}\right)}{\left(1 - \frac{A}{N}\right)\left(1 - \frac{B}{N}\right)}, 0\right] \tag{20}$$

## 5. RESULTS

Here we conduct a number of experiments to show how the correction addresses some of the limitations described by Flower,[4] so that fingerprint similarity can be computed in a manner unbiased by the choice of fingerprint length. We first use the similarity results obtained with the uncompressed fingerprints as the gold standard and compare how well the corrected and uncorrected similarity measures

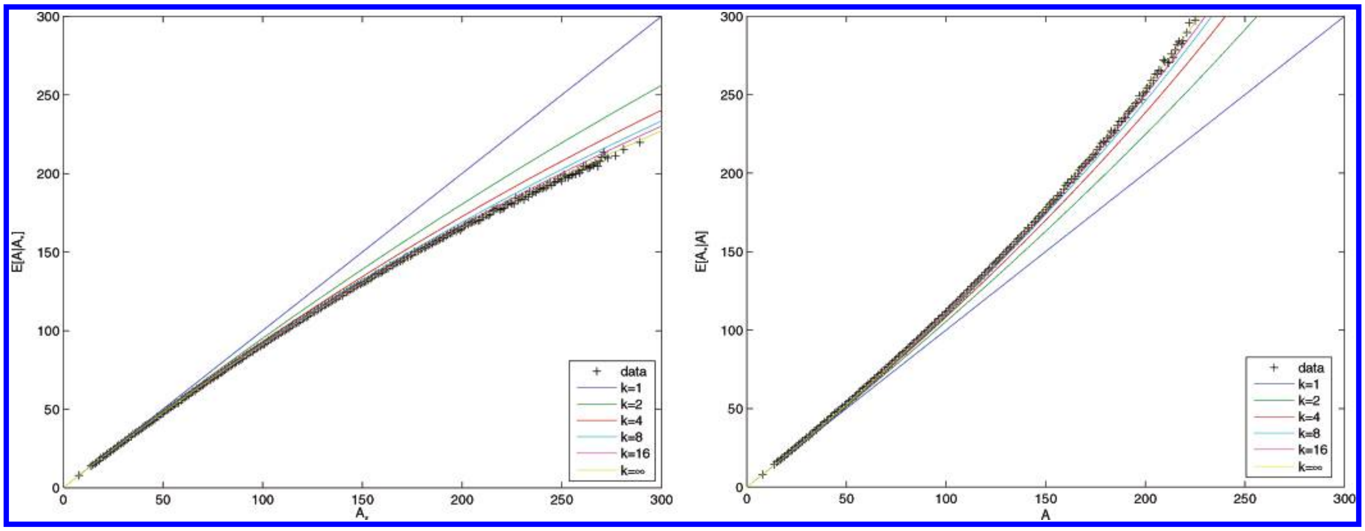

**Figure 2.** Left: Estimation of $E(A|A_*)$ using the binomial model at different values of $k$ (eq 4) and the corresponding asymptotic approximations (eq 7). The lines correspond to predictions at different values of $k$, while each data point corresponds to the empirical mean $A$ estimated at $N = 512$ from at least 15 chemicals from the ChemDB subset with the same $A_*$. In this case, $N_*$ and $k$ are large, so the asymptotic density model most accurately predicts the expectation. Right: Estimations of $E(A_*|A)$ using the binomial model at different values of $k$ (eq 5) and the corresponding asymptotic approximations (eq 8). The lines correspond with predictions at different values of $k$, while each data point corresponds with the mean $A_*$ of at least 15 chemicals from the ChemDB subset with the same $A$ measured at $N = 512$. In this case, $N_*$ and $k$ are large, so the asymptotic density model most accurately predicts the expectation.

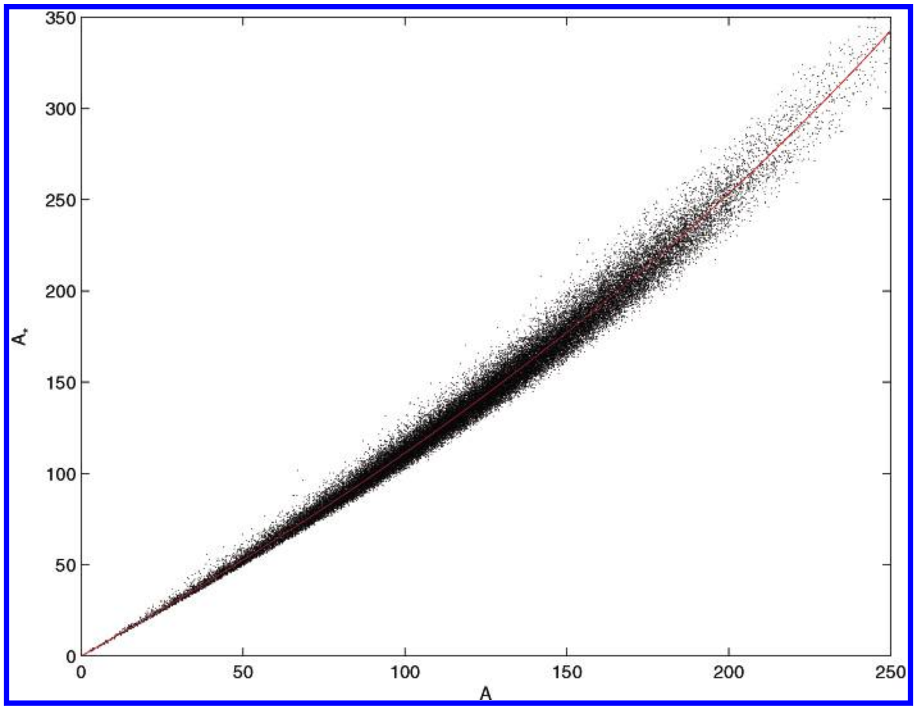

**Figure 3.** Each point represents a molecule in the ChemDB subset using binary fingerprints of length $N = 512$. Random jitter uniform over $[-0.5,0.5]$ is injected in each coordinate to improve readability. The red curve corresponds to the predicted relationship between $A$ and $A_*$ using the asymptotic approximation to the binomial model (eq 8). The predicted values correspond closely to the expected empirical values.

applied to the compressed fingerprints approximate this standard. We use the notation: FP = False Positives, TP = True Positives, FN = False Negatives, and TN = True Negatives. Performance is assessed in terms of standard information retrieval performance measures, such as Precision $P = TP/(TP + FP)$, Recall $R = TP/(TP + FN)$, area under the ROC curve (AUC) which captures the tradeoff between false positive rates and true positive rates, and the F measure[15] which is the harmonic mean of Precision and Recall $F = 2PR/(P + R)$.

Figure 4 illustrates the effect of the correction by comparing the behavior of the corrected and uncorrected Tanimoto similarity measures using random uniform fingerprint vectors satisfying $A_* = 200$, $B_* = 200$, and $A_* \cap B_* = 100$ for various compression lengths $N$. In this case, the true (uncompressed) Tanimoto similarity is held constant and equal to 1/3. For small values of $N$ the corrected Tanimoto underestimates the true value, and the uncorrected Tanimoto overestimates the true value. However, the corrected Tanimoto is significantly closer, and converges faster, to the true value. The corrected

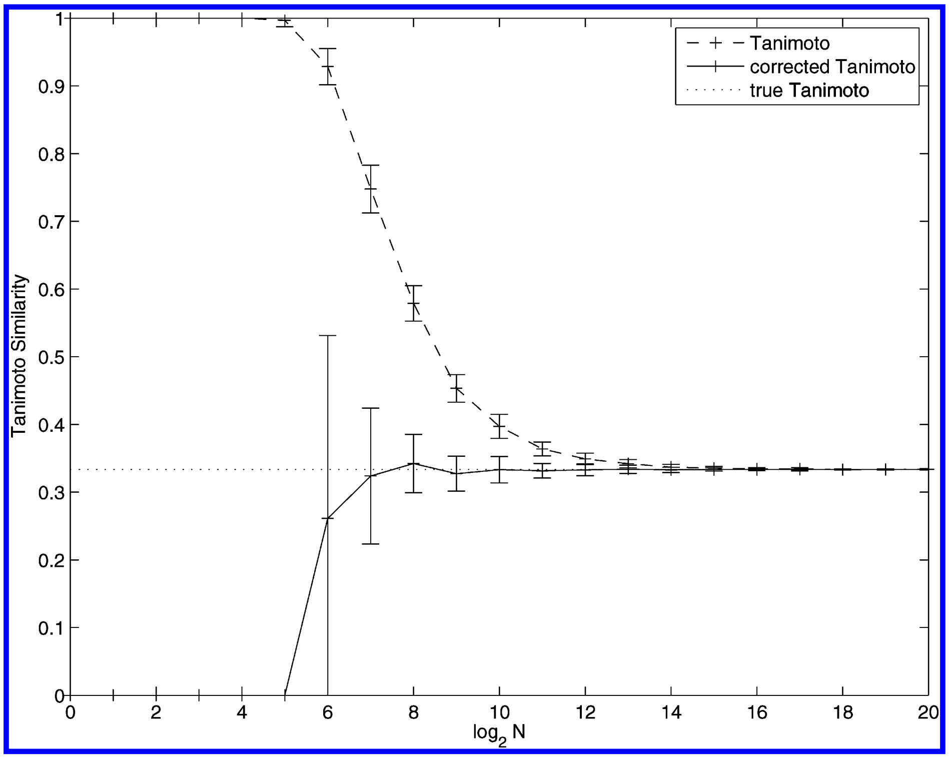

**Figure 4.** Tanimoto and corrected Tanimoto similarity for fingerprints generated uniformly at random with $A_* = 200$, $B_* = 200$, and $C_* = A_* \cap B_* = 100$ and different compression lengths $N$. The true (uncompressed) Tanimoto similarity remains constant and equal to 1/3. Between $N = 2^7$ and $N = 2^{12}$ the corrected Tanimoto value yields a much better approximation of the true value than the uncorrected Tanimoto measure.

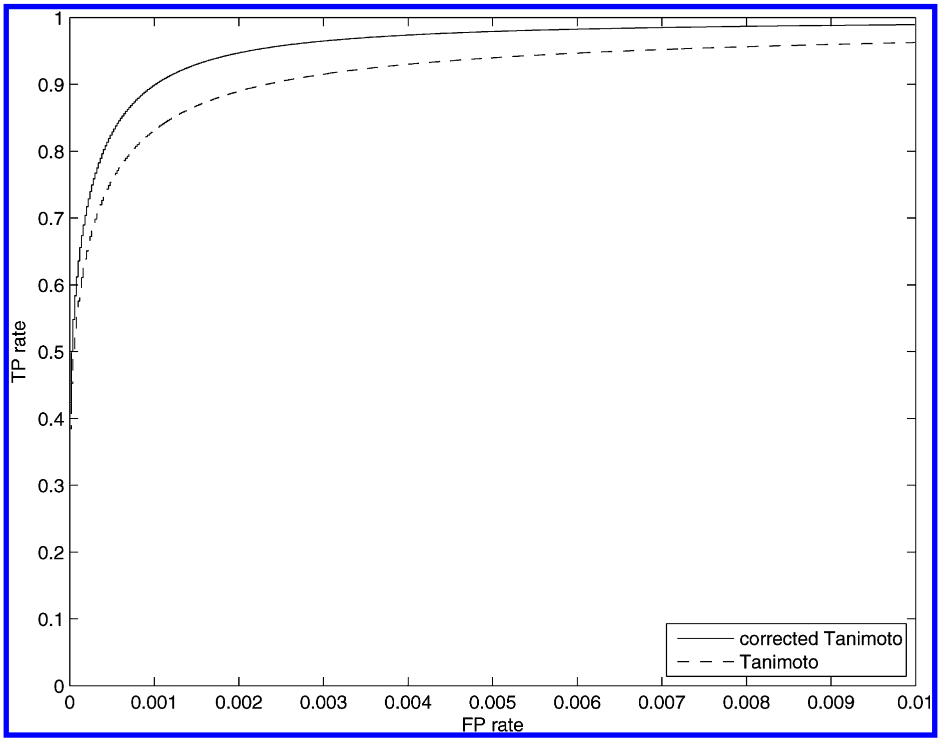

**Figure 5.** ROC curve obtained using each molecule in the ChemDB subset of 50 000 molecules as a query and fingerprints of length $N = 512$. The gold standard is the ranking provided by the uncompressed Tanimoto measure. For each query, the top 50 hits are considered to be positive.

Tanimoto provides the correct answer as soon as $N = 2^7$, whereas the same level of performance is reached by the uncorrected Tanimoto measure only around $N = 2^{14}$.

Figure 5 displays the ROC curve of the False Positive (FP) rate versus the True Positive (TP) rate for the corrected and uncorrected Tanimoto similarity measures for fingerprints of length $N = 512$ at different thresholds, computed on the random sample of 50,000 molecules from the ChemDB. The gold standard is the ranking provided by the Tanimoto measure applied to the uncompressed fingerprints. For each query, the top 50 hits are considered to be positive. At all thresholds or FP rates, the corrected Tanimoto curve is above

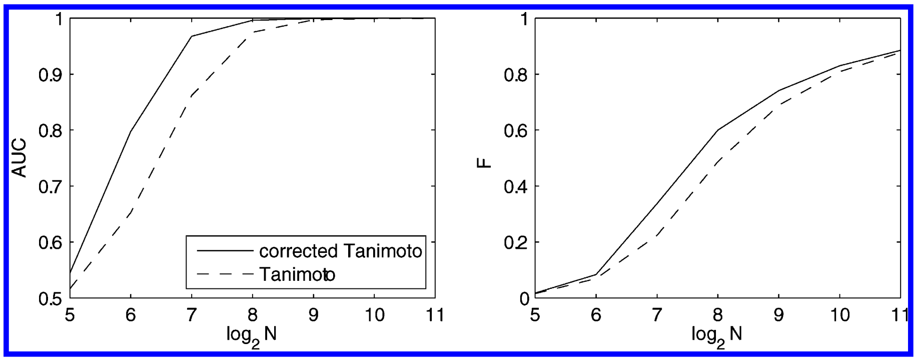

**Figure 6.** Area under the ROC curve (left) and F measure (right) as a function of fingerprint length *N*. In both cases, the corrected fingerprints lead to noticeable improvements. Both curves are derived using the top 50 hits.

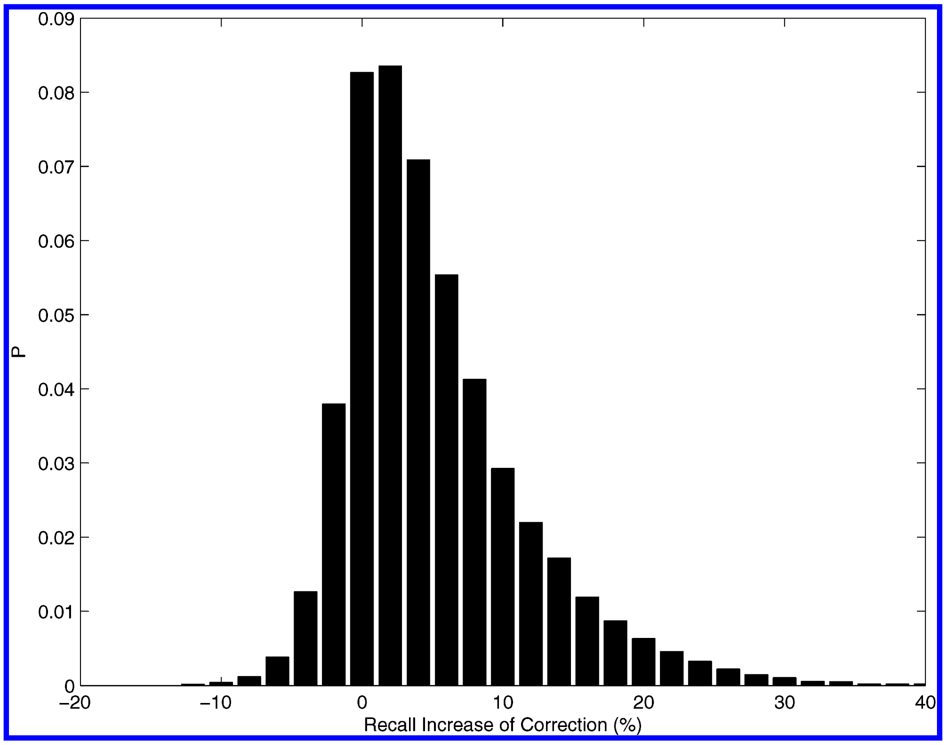

**Figure 7.** Recall difference histogram. The *x*-axis represents the difference in Recall between the corrected and uncorrected Tanimoto measures. The *y*-axis corresponds to a number of occurrences across the 50 000 queries. In rare cases, recall performance can deteriorate, while in most case it improves with the correction. Histogram created using the top 50 hits as positives.

the uncorrected Tanimoto curve. The curves obtained using the top 100 hits, instead of the top 50, look very similar to these. The corrected Tanimoto measure yields a small but clearly noticeable gain associated with an increase in the area under the ROC curve.

Figure 6 displays the area under the ROC Curve (AUC) and the F measure, as a function of fingerprint length on a logarithmic scale, for the corrected and uncorrected similarity measures using the random set of 50,000 molecules from the ChemDB. At all compression lengths, the corrected Tanimoto measure achieves better AUC and F performance measures. While as $N \rightarrow \infty$ the values associated with the corrected and uncorrected measures converge to the same values, as predicted by the theory, the gain for typical current values of *N* (512 or 1024) remains significant. Furthermore, the gain can depend on other factors that are beyond the scope of this analysis, such as the fingerprint density and the specific set of features.

The histogram of the difference in Recall between the corrected and uncorrected Tanimoto measures in the previous experiment, across 50,000 queries, is shown in Figure 7. In most cases, the correction improves the Recall by a significant percentage. Similar histograms are obtained for other measures such as Accuracy, Precision, and F measure. In all cases, the corrected Tanimoto performs significantly better on average. The figure is derived using the top 50 hits as positives. Using the top 100 hits, the Recall improvement is similar but even more pronounced.

The average Recall improvement is plotted in Figure 8, as a function of *A*, with $N = 512$ and using all 50 000 queries. Using the top 50 hits as positives, there is on average a 7.48% increase in Recall. Using the top 100, the average improvement in Recall is 9.47%. Similar results are observed with the other classification performance measures.

Figure 9 displays the corrected and uncorrected Tanimoto similarity values versus the uncompressed Tanimoto similarity value for a random subset of 1,000 pairs of molecules. On the left, the 1,000 pairs are randomly sampled uniformly from the ChemDB database. Again the improvement of the correction is obvious. However, with randomly sampled

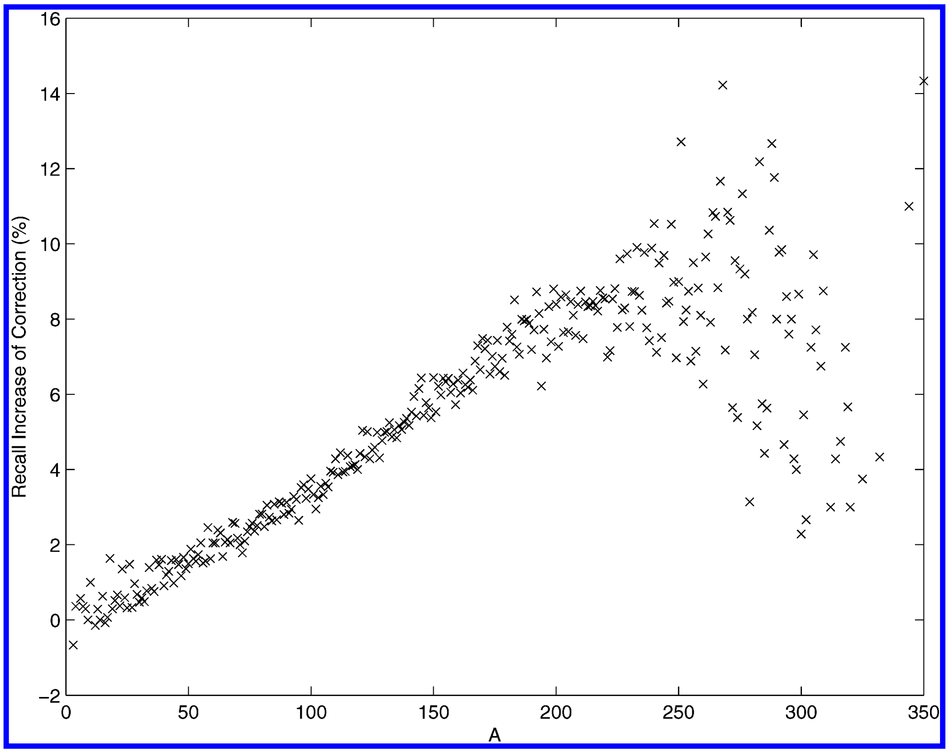

**Figure 8.** Average improvement in Recall as a function of the number $A$ of bits ($N = 512$). Jitter increases for larger values of $A$ because there are fewer data points. A similar, but smaller, effect is also observed for very small values of $A$. Top 50 hits are used as positives.

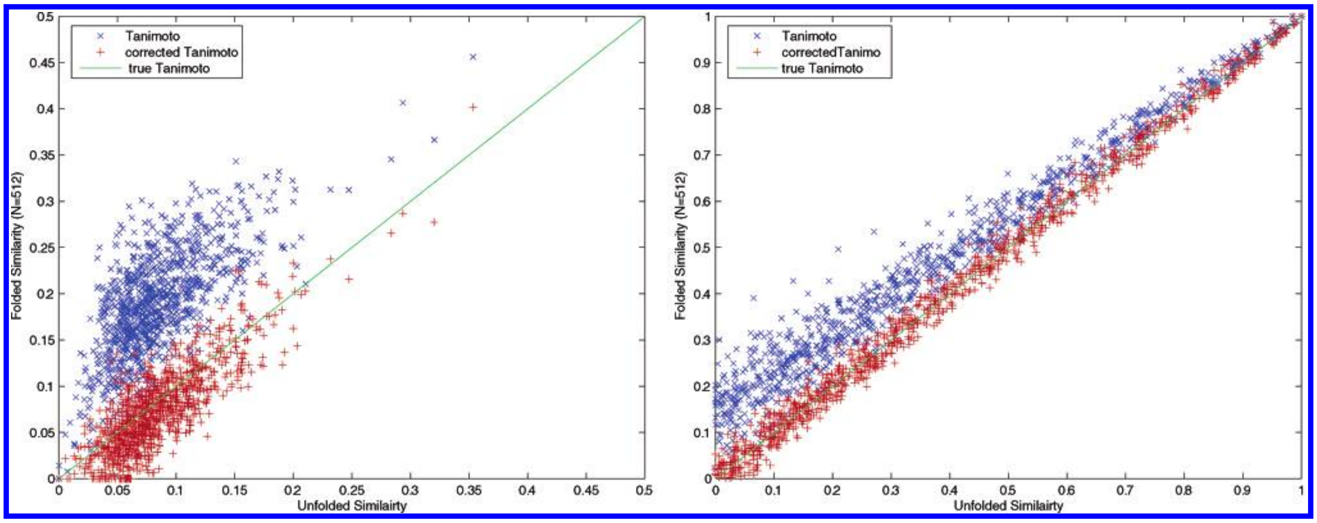

**Figure 9.** Left: Plot of uncompressed Tanimoto similarity versus corrected (red) and uncorrected (blue) Tanimoto similarity computed on a random sample of 1,000 pairs of molecules extracted from the ChemDB. Right: Similar plot obtained on 1,000 pairs of randomly generated bit vectors. To generate these vectors, a value $A_*$ is first sampled according to its distribution in ChemDB, then a value $C_*$ is sampled uniformly between 0 and $A_*$. Two bit vectors $\vec{A}_*$ and $\vec{B}_*$ are then generated uniformly satisfying the constraints $A_* = B_*$ and $A_* \cap B_* = C_*$.

molecules, the similarity in general is bound to be low which explains why the points are clustered toward the lower left corner. The right plot overcomes this limitation to some extent by sampling also high similarity values, by generating random bit vectors as described in the figure legend.

The previous results clearly demonstrate that the corrected similarity measure computed on compressed fingerprints provides a better approximation to the similarity measure computed on full length fingerprints than the uncorrected similarity measure computed on compressed fingerprints. However, in a somewhat contrived way, one could argue that the goal is not to approximate the value of the similarity measure applied to uncompressed fingerprints but rather to retrieve chemically relevant molecules. In other words, there is a small chance that the uncorrected Tanimoto measure could provide a weaker approximation to the true Tanimoto measure but still provide more chemically meaningful results. To rule out this possibility, we test the corrected and uncorrected similarity measures on six biochemically relevant data sets of molecules. These sets are combined with the random subset of 50,000 molecules from the ChemDB. Retrieval of each data set is tested against this random background. Figure 10 demonstrates how correcting fingerprints ($N = 512$) improves also the retrieval for these six data sets. Each plot corresponds to a different set of molecules which are known to interact with the same protein.[13] Each ROC curve is constructed by aggregating the ROC curves calculated by using each molecule in the group

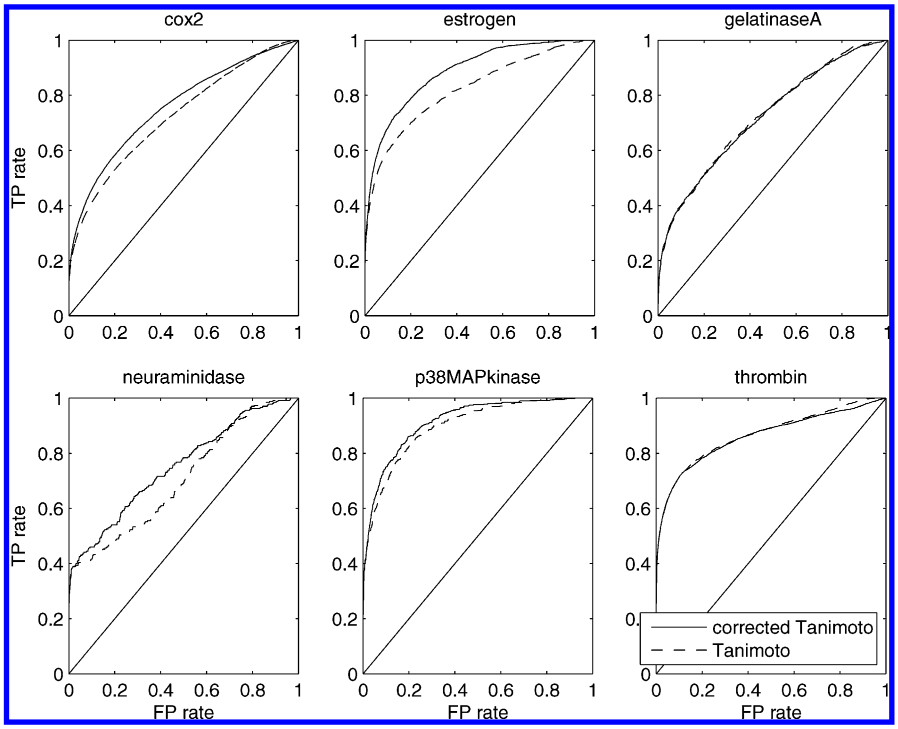

**Figure 10.** ROC curves for corrected and uncorrected Tanimoto measures applied to compressed fingerprints ($N = 512$) for six biologically relevant data sets. Each plot corresponds to a different set of molecules which are known to interact with the same protein. Each ROC curve is constructed by aggregating the ROC curves calculated by using each molecule in the group to search for the rest of the group against the background provided by the random subset of 50 000 molecules form the ChemDB.

as a query to search for the rest of the group. The data show that in all six cases correcting the similarity measure increases the area under the ROC curve, often by a significant amount. Similar results are obtained using other performance measures, such as the F measure.

Finally, while the correction clearly improves the retrieval, one needs also to consider its computational cost. From eqs 17−20, one can see that the correction introduces a small, essentially constant, overhead in the computation of the similarity. To further assess the value of this overhead, we compute the corrected and uncorrected similarity measures 1,000,000 times for a random pair of fingerprints of length $N = 512$. Using an optimized C implementation wrapped in Python using SWIG on a 2.4 MHz AMD Opteron processor, the computation takes 37.2 s without the correction and 37.6 s with the correction. Thus, in short, the computational overhead is negligible. Furthermore, the correction methods can easily be combined with fast search algorithms.[17]

## 6. CONCLUSION

As databases of compounds continue to grow in size, fingerprint representations and their similarity measures become increasingly important. Compressed fingerprints provide efficient compact representations for searching these databases. The most widely used compression algorithm by modulo operation, however, discards some information and has traditionally been associated with a level of distortion in the similarity measures that increases as the length of the fingerprints decreases. Here we have derived a mathematical correction to alleviate this problem as much as possible. The key to the correction is a better estimate, given the compressed fingerprints, of the overlap between the uncompressed fingerprints. While the correction has been illustrated on the most widely used similarity measure—the Tanimoto measure—with fixed-size fingerprints, in the Appendixes we show how the same methods can be applied to all commonly used similarity measures as well as to fingerprints of variable size. With minimal and essentially constant computational overhead, the correction leads to a noticeable improvement in retrieval performance.

## ACKNOWLEDGMENT

Work supported by an NIH Biomedical Informatics Training grant (LM-07443-01) and an NSF MRI grant (EIA-0321390) to P.B., by the UCI Medical Scientist Training Program, and by a Harvey Fellowship to S.J.S. We would like also to acknowledge the OpenBabel project and OpenEye Scientific Software for their free software academic license and Drs. Chamberlin, Nowick, and Weiss for their useful feedback.

## 7. APPENDIX 1: EXACT MODEL

**7.1. Exact Treatment.** $P(A|A_*)$ can be computed exactly, without resorting to the binomial model approximation, under the probabilistic framework induced by assuming that all random permutations of the bits in the uncompressed fingerprint vector are equally likely to occur or, equivalently, that the distribution on $\vec{A}_*$ given the count $A_*$ is uniform. Under this assumption, the probability $P(A|A_*)$ is given by

$$P(A|A_*) = \frac{\binom{N}{A} G(A,A_*)}{\binom{N_*}{A_*}} \tag{21}$$

where $\binom{N_*}{A_*}$ counts all the possible uniform realizations of the vector $\vec{A}_*$, and $\binom{N}{A}$ counts all the corresponding realizations of the vector $\vec{A}$. We let $G(A,A_*)$ represent the number of ways a particular set of $A_*$ bits could result in a particular set of $A$ bits after compression. In other words, $G(A,A_*)$ counts all the possible ways of placing $A_*$ balls (or bits) in a $A \times k$ array of slots, so that there is at least one ball in each column. We can derive several recurrence relations on the numbers $G(A,A_*)$. First note that $G(A,A_*) = 0$ if $A_* < A$ or if $A_* > A \times k$, $G(A,0) = 0$ when $A > 0$, $G(0,A_*) = 0$ when $A_* > 0$, and $G(0,0) = 1$.

By counting all possible ways of putting $A_*$ balls in the $A \times k$ array and subtracting those that have one empty column, two empty columns, and so forth, we get

$$G(A,A_*) = \binom{Ak}{A_*} - \sum_{j=1}^{A}\binom{A}{j}G(A-j,A_*) \tag{22}$$

when $A_* \leq Ak$, and 0 otherwise. In the sum, we must have $(A - j)k \geq A_*$ so that in general the summation goes only up to the $l$th term, with $l = A - \lfloor A - \frac{A_*N}{N_*} \rfloor$, beyond which $G(A - j,A_*) = 0$. From this, it is easy to see that each term $G(A,A_*)$ can be expanded in the form

$$G(A,A_*) = \sum_{n=0}^{A} C_{A,n}\binom{(A-n)k}{A_*} \tag{23}$$

when $A_* \leq Ak$, and 0 otherwise. Finally, by solving this recurrence relation or directly applying the inclusion-exclusion principle, we get

$$G(A,A_*) = \sum_{n=0}^{A}(-1)^n\binom{A}{n}\binom{(A-n)k}{A_*} \tag{24}$$

so that $C_{A,n} = (-1)^n(A/n)$. Thus, in principle, we can compute $P(A|A_*)$ and the expectation $E(A|A_*)$ under the fixed size mode, although the calculations appear to be considerably more involved in comparison to the simple fixed density binomial approximation.

## 8. APPENDIX 2: EXTENSIONS TO OTHER SIMILARITY MEASURES

Fingerprint similarity can be calculated using many different measures. Holliday et al. (2002)[10] compare a comprehensive list of fingerprint similarities and distances. Using their nomenclature, we show here how to apply our correction by substituting $A$, $B$, $A \cup B$, and $A \cap B$ with the estimates of the corresponding uncompressed values $A_*$, $B_*$, $A_* \cup B_*$, and $A_* \cap B_*$, given by eqs 17−20 for large $N_*$. Based on how the value of $N$ (or $N_*$) appears in the measures, the measures can be divided into four classes. The first class of measures and their correction can be written without using $N$ (or $N_*$). The second class of measures are defined using $N$, corrected using $N_*$, but can be written in a rank-equivalent (i.e., which ranks molecules in the exact same order of similarity) form which does not involve $N_*$. The third class of measures are defined using $N$, corrected using $N_*$, but as $N_* \to \infty$ they can be expressed in a rank-equivalent form which does not involve $N_*$. The fourth class of measures can be corrected using $N_*$ but converge to a constant as $N_* \to \infty$, without yielding an asymptotic rank-equivalent form.

**8.1. Class One: Similarity Measures Independent of *N*.** The measures in this class can be written in a form that does not contain $N$. They can be trivially corrected by using the corresponding estimates. For example, as we have seen, the *Jaccard/Tanimoto* measure defined by $S(\vec{A},\vec{B}) = (A \cap B)/(A \cup B)$ can be corrected by

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{A_* \cup B_*} \tag{25}$$

The *Tversky* measure can be written as $S_{\alpha\beta}(\vec{A},\vec{B}) = (A \cap B)/(\alpha A + \beta B + (1 - \alpha - \beta)(A \cap B))$, and, therefore, it can be corrected by

$$S_{\alpha\beta}(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{\alpha A_* + \beta B_* + (1 - \alpha - \beta)(A_* \cap B_*)} \tag{26}$$

The *Ochiai/Cosine* measure defined by $S(\vec{A},\vec{B}) = (A \cap B)/(\sqrt{AB})$ can be corrected by

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{\sqrt{A_*B_*}} \tag{27}$$

The *McConnaughey* measure is defined by $S(\vec{A},\vec{B}) = ((A \cap B)(2 - A - B) + AB)/(AB)$ and can be corrected by

$$S(\vec{A},\vec{B}) = \frac{(A_* \cap B_*)(2 - A_* - B_*) + A_*B_*}{A_*B_*} \tag{28}$$

The *Dice* and *Sokal*/*Sneath(1)* measures are equivalent to Tversky similarity with $\alpha = \beta = 0.5$ and $\alpha = \beta = 2$, respectively. Noting this relationship, we can use the previous derived corrections for the Tversky similarity. For example, the *Dice* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{2(A_* \cap B_*)}{A_* + B_*} \tag{29}$$

The *Kulczynski(1)* measure can be written as $S(\vec{A},\vec{B}) = (A \cap B)/((A \cup B) - (A \cap B))$ and corrected by

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{(A_* \cup B_*) - (A_* \cap B_*)} \tag{30}$$

The *Kulczynski(2)* measure can be written as $S(\vec{A},\vec{B}) = ((A \cap B)(A + B))/(2AB)$ and corrected by

$$S(\vec{A},\vec{B}) = \frac{(A_* \cap B_*)(A_* + B_*)}{2A_*B_*} \tag{31}$$

And finally, the *Simpson* measure can be written as $S(\vec{A},\vec{B}) = (A \cap B)/(\min (A,B))$ and corrected by

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{\min (A_*,B_*)} \tag{32}$$

**8.2. Class Two: Similarity Measures Rank-Equivalent to Similarity Measures Independent of *N*.** This class of measures can be rewritten in a rank-equivalent, sometimes strictly proportional, form which does not use $N$ (or $N_*$).

The *Russel/Rao* measure is defined by $S(\vec{A},\vec{B}) = (A \cap B)/N$. Noting that $1/N$ and $1/N^*$ are positive constants, it can be corrected by

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{N_*} \propto A_* \cap B_* \tag{33}$$

Here the symbol ∝ is used to denote rank equivalence.

Likewise, noting that $f(x) = (C + x)/C$ is a monotonically increasing transformation of $x$, the *Simple Matching* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_* + (A_* \cap B_*) - (A_* \cup B_*)}{N_*} \propto (A_* \cap B_*) - (A_* \cup B_*) \tag{34}$$

Similarly, the *Hamann* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_* + 2(A_* \cap B_*) - 2(A_* \cup B_*)}{N_*} \propto (A_* \cap B_*) - (A_* \cup B_*) \tag{35}$$

The *Mean Manhattan* distance is the L1 norm, also known as the city-block distance. It is the complement of the Simple Matching measure, and, in the binary case, it is rank-equivalent to Euclidian distance. Its correction can be written as

$$D(\vec{A},\vec{B}) = \frac{(A_* \cup B_*) - (A_* \cap B_*)}{N_*} \propto (A_* \cup B_*) - (A_* \cap B_*) \tag{36}$$

Here we use $D$ to denote distance and distinguish it from similarity $S$.

Similarly, the *Normalized Euclidian Distance* is the familiar L2 norm or Euclidian Distance. Its corrections can be written as

$$D(\vec{A},\vec{B}) = \lim_{N \to \infty} \sqrt{\frac{(A_* \cup B_*) - (A_* \cap B_*)}{N_*}} \propto \sqrt{(A_* \cup B_*) - (A_* \cap B_*)} \tag{37}$$

The *Forbes* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_*(A_* \cap B_*)}{A_* B_*} \propto \frac{A_* \cap B_*}{A_* B_*} \tag{38}$$

The *Fossum* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_*\left((A_* \cap B_*) - \frac{1}{2}\right)^2}{A_* B_*} \propto \frac{\left((A_* \cap B_*) - \frac{1}{2}\right)^2}{A_* B_*} \tag{39}$$

Noting that $f(x) = (C - x)/(C + x)$ is a monotonic decreasing function of $x$, the *Rogers/Tanimoto* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_* - (A_* \cup B_*) + (A_* \cap B_*)}{N_* + (A_* \cup B_*) - (A_* \cap B_*)} \propto (A_* \cap B_*) - (A_* \cup B_*) \tag{40}$$

Likewise, the *Sokal/Sneath(3)* measure can be corrected by

$$S(\vec{A},\vec{B}) = \frac{N_* - (A_* \cup B_*) + (A_* \cap B_*)}{(A_* \cup B_*) - (A_* \cap B_*)} \propto (A_* \cap B_*) - (A_* \cup B_*) \tag{41}$$

**8.3. Class Three: Similarity Measures Asymptotically Rank-Equivalent to Similarity Measures Independent of *N*.** The corrected version of these measures reveals rank-equivalent asymptotic forms as $N_* \to \infty$ while maintaining $A_*$ and $B_*$ fixed. For instance, the corrected *Pearson* measure becomes

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \frac{N_*(A_* \cap B_*) - A_* B_*}{\sqrt{A_* B_*(N_* - A_*)(N_* - B_*)}} = \frac{A_* \cap B_*}{\sqrt{A_* B_*}} \tag{42}$$

Likewise, the *Dennis* measure yields a corrected rank-equivalent form obtained by dividing by $\sqrt{N_*}$ in the form

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \frac{N(A_* \cap B_*) - A_* B_*}{\sqrt{N_* A_* B_*}} \propto \frac{A_* \cap B_*}{\sqrt{A_* B_*}} \tag{43}$$

which is identical to the corrected asymptotic Pearson measure.

In the same manner, the rather complex *Stiles* measure diverges, but a rank-equivalent form can be obtained by subtracting $\log_{10} N_*$ in the form

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \log_{10} \frac{N_*\left(\left|N_*(A_* \cap B_*) - A_* B_*\right| - \frac{N_*}{2}\right)^2}{A_* B_*(N_* - A_*)(N_* - B_*)} \propto \log_{10} \frac{\left((A_* \cap B_*) - \frac{1}{2}\right)^2}{A_* B_*} \tag{44}$$

This form is rank equivalent with the *Fossum* measure.

**8.4. Class Four: Similarity Measures Asymptotically Converging to Constants.** These similarity measures can be corrected, but the corrected expression converges to a constant when $N_* \to \infty$, without yielding an asymptotic rank-equivalent form. For instance, for the *Baroni-Urbani/Buser* measure, the corrected version converges to 1 as seen in

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \frac{\sqrt{(A_* \cap B_*)(N_* - A_* \cup B_*)} + A_* \cap B_*}{\sqrt{(A_* \cap B_*)(N_* - A_* \cup B_*)} + A_* \cup B_*} = 1 \tag{45}$$

Note that this measure is also rank-equivalent to Tanimoto for any $N_*$ and therefore can also be included in Class Two.

The corrected *Sokal/Sneath(2)* yields

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \frac{2(N_* - (A_* \cup B_*) + (A_* \cap B_*))}{2N_* - (A_* \cup B_*) + (A_* \cap B_*)} = 1 \tag{46}$$

Similarly, the corrected *Yule* measure yields

$$S(\vec{A},\vec{B}) = \lim_{N_* \to \infty} \frac{N_*(A_* \cap B_*) - A_*B_*}{(A_* \cap B_*)(N_* - 2A_* - 2B_* + 2) + A_*B_*} = 1 \quad (47)$$

**8.5. Equivalences.** From these corrections, we can see clear groups of measures which produce identical rankings after correction, either immediately or asymptotically. These groups include the following:

(1) Special cases of the Tversky measure, which includes the Jaccard/Tanimoto, Dice, and Sokal/Sneath(1) measures:

$$S_{\alpha\beta}(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{\alpha A_* + \beta B_* + (1 - \alpha - \beta)(A_* \cap B_*)} \quad (48)$$

(2) The Ochiai/Cosine, Forbes, Pearson, and Dennis measures are all rank equivalent with

$$S(\vec{A},\vec{B}) = \frac{A_* \cap B_*}{\sqrt{A_*B_*}} \quad (49)$$

(3) The Fossum and Stiles measures are part of the next grouping. This group is nearly identical to the Ochiai/Cosine measure, producing essentially the same top rankings, with the similarity

$$S(\vec{A},\vec{B}) = \frac{\left((A_* \cap B_*) - \frac{1}{2}\right)^2}{A_*B_*} \quad (50)$$

(4) The measures which converge to Simple Matching, which includes the Rogers/Tanimoto, Hamann, and Sokal/Sneath(3), with the similarity

$$S(\vec{A},\vec{B}) = (A_* \cap B_*) - (A_* \cup B_*) \quad (51)$$

(5) The Mean Manhattan Distance and the Euclidian Distance remain rank-equivalent in the binary case with ranking generated by

$$D(\vec{A},\vec{B}) = (A_* \cup B_*) - (A_* \cap B_*) \quad (52)$$

Finally, the McConnaughey, Kulczynski(1), Kulczynski-(2), Simpson, and Russel/Rao measures remain distinct measures. After applying the correction, this leaves us with nine distinct similarity measures and one distance metric.

## 9. APPENDIX 3: VARIABLE-LENGTH FINGERPRINTS

Some chemoinformatics systems use variable length fingerprints, where the length $N$ is chosen adaptively for each molecule, in order to hold the bit density in a fixed range. Holding the density at approximately the same value minimizes the storage space of smaller molecules while preserving information for larger molecules. So given two uncompressed fingerprints $\vec{A}_*$ and $\vec{B}_*$ these are compressed to two fingerprints $\vec{A}_{N(\vec{A}_*)}$ and $\vec{B}_{N(\vec{B}_*)}$ of length $N(\vec{A}_*)$ and $N(\vec{B}_*)$ respectively, with $A_{N(\vec{A}_*)}$ and $B_{N(\vec{B}_*)}$ bits set to 1 in each compressed vector. In variable-size fingerprint systems, meaningful similarity can be defined only when there is a correspondence between the indices of fingerprints of different size. So, if the larger fingerprint is $n$ times longer than the smaller fingerprint, each bit in the smaller fingerprint corresponds with $n$ bits in the larger fingerprint. The most simple and convenient case is when the fingerprints lengths are powers of 2. All current approaches for computing similarity between two fingerprints of different lengths begin by deriving two fingerprints of equal length and then applying a standard similarity measure to these fingerprints of equal length. Therefore all the correction formula can be applied at this stage.

More precisely, to fix the idea, given two molecules $\mathscr{A}$ and $\mathscr{B}$, let us assume that after compression $N(\vec{A}_*) > N(\vec{B}_*)$. The formulas derived in section 4 allow us to immediately get estimates of $A_*$ and $B_*$ from $A_{N(\vec{A}_*)}$ and $B_{N(\vec{B}_*)}$ in the form (for large $N_*$)

$$A_* \approx -N(\vec{A}_*) \log\left(1 - \frac{A_{N(\vec{A}_*)}}{N(\vec{A}_*)}\right) \quad \text{and}$$

$$B_* \approx -N(\vec{B}_*) \log\left(1 - \frac{B_{N(\vec{B}_*)}}{N(\vec{B}_*)}\right) \quad (53)$$

The problem is how to get good estimates of the uncompressed intersection and union.

One approach to estimate the intersection would be to fold the vector $\vec{A}_{N(\vec{A}_*)}$ until it has a length equal to $N(\vec{B}_*)$ and then apply a Tanimoto or corrected Tanimoto measure to these short vectors. This may not be ideal because by further folding the vector $\vec{A}_{N(\vec{A}_*)}$ one increases its density.

A second alternative is simply to look only at the first $N(\vec{B}_*)$ components of the vector $\vec{A}_{N(\vec{A}_*)}$ and compute the intersection with $\vec{B}_{N(\vec{B}_*)}$. This approach may also not be ideal because it discards a considerable amount of information contained in the longer vector $\vec{A}_{N(\vec{A}_*)}$.

A third and perhaps preferable approach is to expand the shorter compressed vector $\vec{B}_{N(\vec{B}_*)}$ consistently with the modulo constraints until it has length $N(\vec{A}_*)$. Obviously this expansion is not unique. To minimize rejecting any molecule that may be similar to the query, we propose to do this expansion in a way that minimizes the size of the union $A_{N(\vec{A}_*)} \cup B_{N(\vec{A}_*)}$. Although there are many optimal configurations for $\vec{B}_{N(\vec{A}_*)}$, it is easy to see that at the optimum the size of the intersection is equal to $A_{N(\vec{A}_*)}$ plus the number of bits in $\vec{B}_{N(\vec{B}_*)}$ which do not have at least one projection in $\vec{A}_{N(\vec{A}_*)}$. This is given by $A_{N(\vec{A}_*)} + (A_{N(\vec{B}_*)} \cup \vec{B}_{N(\vec{B}_*)}) - \vec{A}_{N(\vec{B}_*)}$. So, for large $N_*$, the estimate of the size of the union becomes

$$A_* \cup B_* \approx N(\vec{A}_*) \log\left(1 - \frac{A_{N(\vec{A}_*)} + (A_{N(\vec{B}_*)} \cup B_{N(\vec{B}_*)}) - A_{N(\vec{B}_*)}}{N(\vec{A}_*)}\right) \quad (54)$$

The more general and more symmetric formula, which relaxes the ($N_A \geq N_B$) constraint using $N_{\min} = \min[N(\vec{A}_*), N(\vec{B}_*)]$ and $N_{\max} = \max[N(\vec{A}_*), N(\vec{B}_*)]$, is given by

$$A_* \cup B_* \approx -N_{\max} \log \left(1 - \frac{A_{N(\vec{A}_*)} + B_{N(\vec{B}_*)} + (A_{N_{\min}} \cup B_{N_{\min}}) - A_{N_{\min}} - B_{N_{\min}}}{N_{\max}}\right) \quad (55)$$

Combining this equations with eq 54 we can get the corrected estimate for the size of the intersection

$$A_* \cap B_* = A_* + B_* + N_{\max} \log \left(1 - \frac{A_{N(\vec{A}_*)} + B_{N(\vec{B}_*)} + (A_{N_{\min}} \cup B_{N_{\min}}) - A_{N_{\min}} - B_{N_{\min}}}{N_{\max}}\right) \quad (56)$$

Note, that with these formula, fingerprints of fixed length and their corrections can be viewed as a special case of fingerprints with variable lengths.

## REFERENCES AND NOTES

(1) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 177−182.

(2) Chen, J.; Swamidass, S. J.; Dou, Y.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133−4139.

(3) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard/Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.

(4) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 378−386.

(5) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; 2004. Available at http://www.daylight.com/dayhtml/doc/theory/theory-.toc.html (accessed Jan 2007).

(6) Xue, L.; Godden, J. F.; 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.

(7) Xue, L.; Stahura, F. L.; Bajorath, J. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2032−2039.

(8) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: 2005.

(9) Tversky, A. Features of similarity. *Psychological Rev.* **1977**, *84*, 327−352.

(10) Rouvray, D. Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput*. Sci. **1992**, *32*, 580−586.

(11) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155−166.

(12) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity, and anti-cancer activity. *Bioinformatics* **2005**, *21*, i359−368.

(13) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(14) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(15) Bollobas, B. *Random Graphs;* Academic Press: London, 1985.

(16) van Rijsbergen, C. J. *Information Retrieval*. *Information Retrieval;* Butterworths: London, U.K., 1978.

(17) Swamidass, S. J.; Baldi, P. Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. *J. Chem. Inf. Model.* **2007**, *47*, 302−317.

CI600526A