Comparison of Molecular Fingerprint Methods on the Basis of Biological Profile Data

Andreas Steffen,* Thierry Kogej, Christian Tyrchan, and Ola Engkvist*

DECS Global Compound Sciences, AstraZeneca R&D Mölndal, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

Received September 8, 2008

In this study we evaluated a set of molecular fingerprint methods with respect to their capability to reproduce similarities in the biological activity space. The evaluation presented in this paper is therefore different from many other fingerprint studies, in which the enrichment of active compounds binding to the same target as selected query structures was studied. Conversely, our data set was extracted from the BioPrint database, which contains uniformly derived biological activity profiles of mainly marketed drugs for a range of biological assays relevant for the pharmaceutical industry. We compared calculated molecular fingerprint similarity values between all compound pairs of the data set with the corresponding similarities in the biological activity space and additionally analyzed agreements of generated clusterings. A closer analysis of the compound pairs with a high biological activity similarity revealed that fingerprint methods such as CHEMGPS or TRUST4, which describe global features of a molecule such as physicochemical properties and pharmacophore patterns, might be better suited to describe similarity of biological activity profiles than purely structural fingerprint methods. It is therefore suggested that the usage of these fingerprint methods could increase the probability of finding molecules with a similar biological activity profile but yet a different chemical structure.

INTRODUCTION

Molecular similarity and techniques for similarity searches is a very active field of research.[1,2] In recent years a large variety of computational methods have been developed with the aim of identifying novel active molecules from compound collections.[3,4] Some of these tools compare graph abstractions of the molecular structures;[5,6] others rely on a shape representation of the molecules for calculating a shape-based similarity.[7,8] Furthermore pharmacophore-based similarity search methods are common,[9] and also physicochemical properties can be used for describing molecular similarity.[10] In this article we will focus on methods that mainly rely on bitstring representations of the molecules.[2] This group of computational tools, i.e. *in silico* derived molecular fingerprints, allows generally for fast molecule comparisons and is often applied in ligand-based virtual screenings of very large compound collections or for deriving similarity matrices of large molecular data sets as a basis for clustering. Molecular fingerprints can be used both for finding near-neighbors of known active molecules and to do scaffold-hopping to identify molecules with different structures but similar biological activity against a given target.[11] Some recent studies assessed the ability of representative fingerprint tools to retrieve known active molecules of a given drug target among a set of decoy molecules by means of retrospective enrichment studies.[8,12] These studies are very often done in such a way that one molecule from an active compound series is used as a "bait" to fish out the other active compounds from the series from a database of decoys. In many of these studies structural fingerprints have shown excellent retrieval rates, thus structural fingerprints are valuable tools for finding structurally related active compounds. Scaffold-hopping is another application field for molecular fingerprints, and several successful studies have been published.[11] Here the requirements are different compared to a near-neighbor search, since the goal is to find active molecules with a different structure.

Recently, data sets that contain biological profile data for a large number of marketed drugs have become available.[13,14] In these databases is given not only the relation 'ligand-single protein target' but also the biological activities in a panel of disease-related drug targets—the target vector. Such databases offer entirely new possibilities to validate computational methods in the bioactive space. For example, Schuffenhauer et al. assessed the ability of a number of classification methods to produce clusters of compounds with homogeneous biological activity profiles.[15] Fliri et al. linked biological activity profiles from the BioPrint database to their molecular structure.[16,17] Horvath et al. studied the neighborhood behavior of *in silico* structural spaces with respect to biological activity profiles derived from the BioPrint database.[18] Thus, databases like BioPrint offer the possibility to validate molecular fingerprints from a bioactivity perspective, *i.e.* how well can a fingerprint retrieve molecular pairs with similar bioactivity.

In this study we used the BioPrint database as the source of biological data, which essentially contains uniformly derived biological activity profiles of mainly marketed drugs for a range of proteins relevant for drug design.[13] The aim was then to investigate how the molecular similarities derived from *in silico* fingerprint methods relate to the corresponding biological profile similarities derived from the assays in the BioPrint database. Three types of analysis were done, each

* Corresponding author e-mail: An.Steffen@web.de (A.S.), Ola.Engkvist@astrazeneca.com.

10.1021/ci800326z CCC: $40.75
Published on Web 01/28/2009

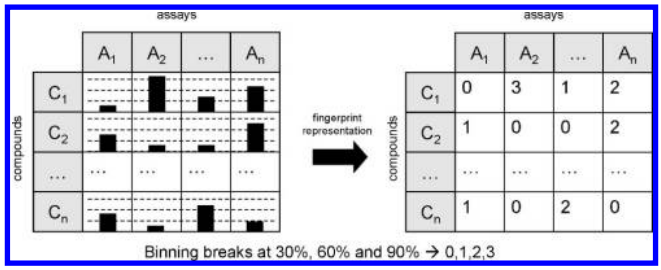

Figure 1. The experimental percent inhibition values (left) are binned to obtain a fingerprint representation of the binding profile of each compound (right). An inhibition percentage of less than 30% corresponds to a binning value of 0 (no inhibition), inhibition between 30% and 60% corresponds to a binning value of 1 (low inhibition), inhibition between 60% and 90% corresponds to 2 (medium inhibition), and inhibition of above 90% to 3 (high inhibition).

of which captures a different aspect of molecular similarity. First we analyzed the entire range of fingerprint-based similarities between the compounds of the BioPrint data set and compared them with the corresponding similarities of the biological activity profiles. This type of analysis reveals correspondences in similarities and dissimilarities with respect to the biological space and the *in silico* fingerprints. The second analysis compared the agreement of the generated clusterings from the computed and the biological similarities respectively, similarly to Schuffenhauer et al.[15] In the last analysis we focused on compound pairs that exhibit a high biological activity profile similarity to estimate the applicability of *in silico* fingerprint methods for retrieving compounds with similar biological profiles.

MATERIALS AND METHODS

In this section we describe how the biological similarities were derived from the binding profiles and explain the applied *in silico* fingerprint methods as well as the statistical quality measures used in the evaluation of them.

Calculation of Biological Screening Profile Similarities. The biological screening profile similarities were calculated from a multivalue fingerprint representation of the biological data. Each bit in this fingerprint corresponds to an assay. For increasing the sensitivity in comparison to a dichotomous representation we decided to apply the following binning scheme: an inhibition percentage of less than 30% at 10 μM corresponds to a binning value of 0 (no inhibition), inhibition between 30% and 60% corresponds to a binning value of 1 (low inhibition), inhibition between 60% and 90% corresponds to 2 (medium inhibition) and inhibition of above 90% to 3 (high inhibition) (see Figure 1).

The similarity between two compounds can then be calculated from the continuous Tanimoto coefficient (T_{cont}) equation:

$$T_{cont} = \frac{\sum_{i=1}^{i=n} x_{iA}x_{iB}}{\sum_{i=1}^{i=n}(x_{iA})^2 + \sum_{i=1}^{i=n}(x_{iB})^2 - \sum_{i=1}^{i=n} x_{iA}x_{iB}} \quad (1)$$

where i iterates over all assays, x_{iA} is the binned percent inhibition value of molecule A and x_{iB} is the binned percent inhibition value of molecule B. The continuous Tanimoto index ranges from −0.3 to 1, where 1 denotes identity. In the following text the biological profile similarity obtained from the experimental data and the continuous Tanimoto similarity as defined in eq 1 is referred to as the BP similarity.

Table 1. Used Fingerprint Methods

name	acronym	type	applied similarity equation
ALOGP	ALOGP	structural	continuous
FOYFI	FOYFI	structural	binary
Molprint2D	MOLP	structural	binary
Scitegic ext.-con. Fp	ECFP_4	structural	binary
Unity2D	UNITY	structural	binary
ALFI	ALFI	pharmacophoric	binary
CATS	CATS	pharmacophoric	continuous
TRUST2 binary	TRUST2.B	pharmacophoric	binary
TRUST2 count	TRUST2.C	pharmacophoric	continuous
TRUST3 binary	TRUST3.B	pharmacophoric	binary
TRUST3 continuous	TRUST3.C	pharmacophoric	continuous
TRUST4 binary	TRUST4.B	pharmacophoric	binary
TRUST4 continuous	TRUST4.C	pharmacophoric	continuous
Scitegic funct.-class fp	FCFP_4	functional class	binary
CHEMGPS	CHEMGPS	property	continuous
LINGO	LINGO	string based	binary

Computational Similarity Tools. The fingerprint methods applied in the present study employ different ways to represent a molecule. In principle a 2D fingerprint is a vector in which each dimension (bin) holds information extracted from the molecular topology or a molecular property. Binary fingerprints only consider the presence or the absence of a set bin, whereas continuous fingerprints contain bins that are either counts (integers) or real numbers. For the binary fingerprint methods we calculated the binary Tanimoto coefficient (T_{bin}), which is defined as follows:

$$T_{bin} = \frac{c}{a+b-c} \quad (2)$$

where "c" is the number of bits common to the two fingerprints and "a" and "b" denote the number of bits set in each of the two fingerprints. T_{bin} ranges between 0 and 1. In Table 1the applied fingerprint methods are listed together with the type of the fingerprint method and the equation used for calculating the respective fingerprint similarity. ALOGP is a functional group based fingerprint originally developed to predict LogP.[19] LINGO is a recently developed fingerprint directly based on a canonical SMILES string representation[20,21] and CATS is an in-house reimplementation of a published pharmacophore fingerprint.[22] FOYFI is an in-house developed fingerprint which is similar to the standard Daylight fingerprint.[7] CHEMGPS,[10] TRUST3 and ALFI are in-house developed fingerprints that have been described previously.[12] TRUST2 is the 2-point pharmacophore version, and TRUST4 is the 4-point pharmacophore version of TRUST3. The different TRUST fingerprints are based on the open source toolkit OpenBabel (www.openbabel.org).[24] Both TRUST2 and CATS are 2-point pharmacophore fingerprints. However, they differ in the pharmacophore definitions and in their fingerprint lengths, due to differences in how the distances between the pharmacophoric features are binned. In CATS each bond distance up to 10 bonds between the pharmacophoric features corresponds to a separate bin. In the TRUST family of fingerprints 8 distance bin are used corresponding to 2, 3, 4−5, 5−7, 7−9, 9−12, 11−15 and 14−20 bonds between the pharmacophores. The Molprint2D

source code was kindly provided by Andreas Bender.[25] The Unity2D was calculated by a command line version of the program. The Scitegic fingerprints were calculated in PipelinePilot.[26] In this article we only discuss the results for the ECFP_4 and FCFP_4 fingerprints—the corresponding ECFP_6 and FPCP_6 fingerprints showed a very similar behavior with respect to our analysis and were therefore omitted. All the fingerprints are accessible at AstraZeneca through a Web interface.[27]

Comparison of the Fingerprint Similarities to the Biological Profile Similarities. First, we calculated all pairwise fingerprint similarities as well as all biological profile (BP) similarities between the compounds of our data set. Based on the calculated similarities, the Pearson, Spearman and Kendall correlation coefficients[28] of the fingerprint similarities and their corresponding biological profile similarities were computed. However, the biological profile similarity values are not equally distributed over the whole range of possible values. A large part of the compound pairs has low similarity. Therefore, such correlation analysis largely depends on whether a fingerprint method can mainly distinguish differences in dissimilarities, whereas it is clearly more relevant for practical applications whether a method can distinguish dissimilar compounds from similar compounds. Indeed, the calculated correlations are low (see Table S1 in the Supporting Information). For this reason we additionally conducted a simple quartile correspondence analysis, which is in some respect similar to the calculation of rank correlation but is less stringent. The quartile correspondence analysis works as follows. First we generated a sorted list of nonredundant similarity values from the corresponding similarity matrices for each fingerprint method as well as for the BP similarity (see Figure 2). Each of these lists was then split into four equally sized bins (quartiles). For each similarity method (the fingerprint methods as well as the BP similarity) every compound pair was assigned to a quartile depending on their similarity. Thereafter, we compared the assigned quartile of each fingerprint similarity to the corresponding quartile of the BioPrint similarity and calculated the number of compound pairs that were in the same quartile.

Analysis of the Generated Clusterings. Distance matrices for clustering were obtained by transforming the Tanimoto similarities into Soergel distances (Soergel distance = 1-Tanimoto similarity).[23] The obtained distance matrices were submitted to a complete-linkage clustering using the statistical programming language R.[29] All fingerprint similarity clustering trees were compared to the clustering tree based on the similarities of the biological data. For the distance matrix based on the biological profiles we used a tree cutoff value of 0.7 (refer to the section entitled Evaluation of Fingerprint Methods for the Prediction of Biological Profiles). Next, we tested for all fingerprint methods which distance cutoff value leads to a maximal adjusted *Rand-Index* (tested in steps of 0.01 between 1 and 0).[30] The adjusted *Rand-Index* compares two clusterings by computing the probabilities that a pair of compounds is classified with both clustering methods in the same cluster or respectively in different ones. An adjusted *Rand-Index* value of 1 corresponds to a complete agreement between the clustering, whereas 0 corresponds to an agreement expected by random

$$R = \frac{\sum_{i,j}\binom{n_{ij}}{2} - E}{0.5\left[\sum_i \binom{n_{i\bullet}}{2} + \sum_i \binom{n_{\bullet j}}{2}\right] - E}$$

$$\text{with } E = \frac{\sum_i \binom{n_{i\bullet}}{2}\sum_i \binom{n_{\bullet j}}{2}}{\binom{N}{2}} \quad (3)$$

$$\text{and } n_{i\bullet} = \sum_j n_{ij}$$

$$\text{and } n_{\bullet j} = \sum_i n_{ij}$$

where R is the adjusted *Rand-Index* and n_{ij} are the elements to be clustered.

Evaluation of Fingerprint Methods for the Prediction of Biological Profiles. The third analysis focused on the ability of a fingerprint method to identify compound pairs with a high BP similarity. We quantified this ability by measures commonly used for assessing the quality of classifications in cheminformatics (see eqs 4−6):

$$PRECISION = \frac{TP}{TP + FP} \quad (4)$$

$$RECALL = \frac{TP}{TP + FN} \quad (5)$$

where *TP* is the number of true positives, *FP* is the number of false positives, and *FN* is the number of false negatives. None of the two measures can be used alone since a high *recall* rate can be obtained by simply predicting every pair of compounds as similar, which then, however, results in a low precision. Conversely, a high *precision* can be obtained by only predicting a small number of compound pairs as similar, whereas most of the true positive compound pairs are predicted as false negative. Therefore, a third measure was used that combines the two:

$$FSCORE = \frac{2 \times RECALL \times PRECISION}{RECALL + PRECISION} \quad (6)$$

F-Score is the harmonic average of the precision and recall and is bound between 0 and 1 (where 1 corresponds to perfect classification). In our analysis we defined compound pairs exhibiting a BP similarity above a cutoff value of 0.7 as the active data set (positive data) and all remaining pairs as the inactive data set (negative data). We then calculated the maximal F-Score for each fingerprint method and at which similarity cutoff value the maximal F-Score occurred for each fingerprint.

Calculation of the Maximum Common Substructure Tanimoto. In order to gain further insights into the behavior of the fingerprint methods, we compared the fingerprint similarities against the maximum common substructure Tanimoto (TanimotoMCSS), which was calculated with an in-house Python script based on the OEChem library (http://www.eyesopen.com). The MCSS of two molecules was calculated taking into account matching atomic elements and bond types. The Tanimoto coefficient was calculated only on the matched heavy atoms as described by eq 7:

$$T_{MCSS} = \frac{N_C}{N_A + N_B - N_C} \quad (7)$$

where T_{MCSS} is the MCSS Tanimoto similarity, N_C is the number of matched heavy atoms, N_A is number of heavy atoms in molecule A and N_B is the number of heavy atoms in molecule B.

RESULTS AND DISCUSSION

Data Preprocessing. We extracted the biological activity profile data from our in-house version of the BioPrint database. Percent inhibition data was used for our analysis, since IC_{50} values are only available if the percent inhibition was equal to or greater than 30%. For those compound-assay combinations where an IC_{50} was determined we could, however, calculate the Spearman rank correlation coefficient to the corresponding percent inhibition value and found a correlation of r^2 of 0.75, which further justifies the use of percent inhibition data.

The original file contained percent inhibition data of 2483 structures in 206 assays. In order to focus on compounds that are druglike, a set of property and substructural filters was applied.[31] 1320 molecules passed these filters. Next, all assays in which less than 90% of the compounds had experimental data were removed. Furthermore, we removed CYP, MAO, ADME and cell based toxicity assays. If an assay was present in both a human and a nonhuman version, only the human assay was used. 146 assays remained after the pruning. Thereafter, all compounds were removed which did not have measured data for all of these assays or did not have a percentage inhibition of more than 30% in any of the assays. After a final manual inspection, the data set consisted of 608 high quality compounds with measured percent inhibition data for all 146 assays.

Comparison of the Fingerprint Similarities to the Biological Profile Similarity Based on the BioPrint Data Set. The biological activity profile similarity based on the BioPrint data set (BP similarity) was calculated based on a fingerprint representation of the biological profiles and with the continuous Tanimoto equation (see eq 1) as the similarity measure. Also in recent studies other similarity measures have been used (for example Euclidean distance).[15,18] Here, we preferred the continuous Tanimoto equation, where two compounds essentially need to demonstrate a comparable degree of inhibition in the same assays to obtain a high similarity value, whereas the absence of inhibition in a particular assay does not increase their similarity. Contrarily, if Euclidean distance is used, two compounds are also considered as similar if they do not show inhibition in any of the assays. The absence of biological activity in an assay does not necessarily imply similar structures.[18]

The histogram of all pairwise BP similarities obtained from the experimental data of the compounds within the data set is shown in Figure 3. Most of the compound pairs share a low BP similarity: 73% of the pairs demonstrate a BP similarity below 0.1. Only a small percentage of the compound pairs (~1%) share a similarity greater than or equal to 0.7. The corresponding histograms of the computational fingerprint methods differ clearly between different fingerprints (see Figure 4 and Figures S1−S4 in the Supporting Information). TRUST4.C gives very low similarity values for most of the compound pairs. The similarity distribution exhibits a clear peak in the 0−0.05 bin. 98% of the pairs share a similarity of less than 0.1. Similarities calculated with the FOYFI fingerprint show a narrow distribution around a peak in the 0.15−0.2 bin. The ALOGP (peak in the 0.3−0.35 bin) and particularly the CHEMGPS (peak around 0) fingerprints spread the similarity values over the entire range of possible values. Histograms for the other fingerprint distributions are given in Figures S1−S4 in the Supporting Information.

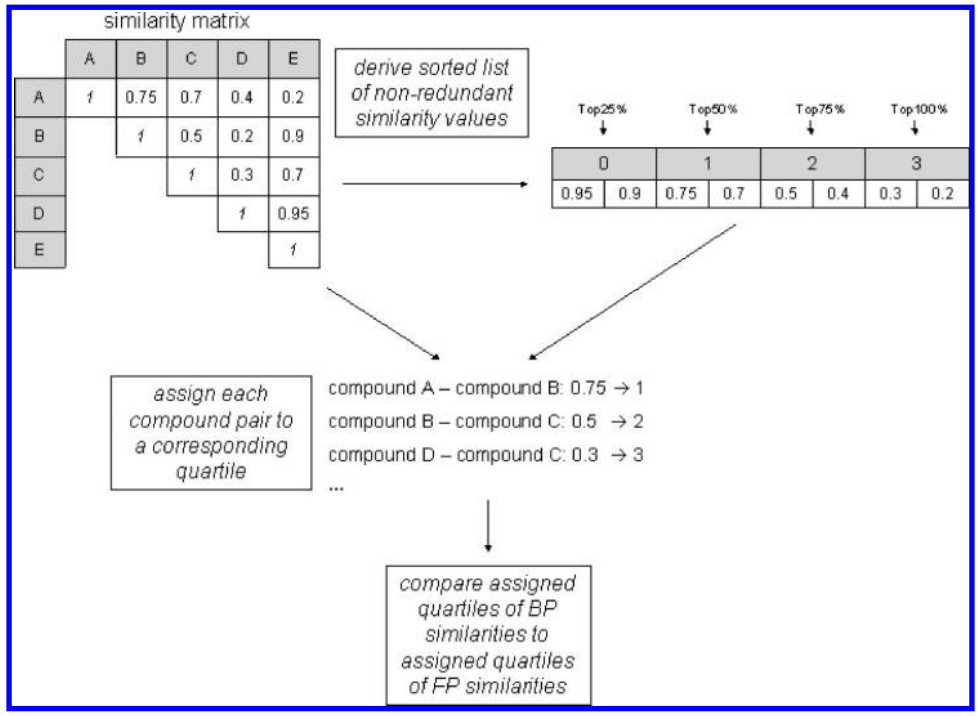

Figure 2. A sorted list of nonredundant similarity values is generated for each similarity matrix, where the similarity value of each compound pair is assigned to the corresponding quartile. The assigned quartiles of the *in silico* fingerprint similarities are then compared to the corresponding BP similarities.

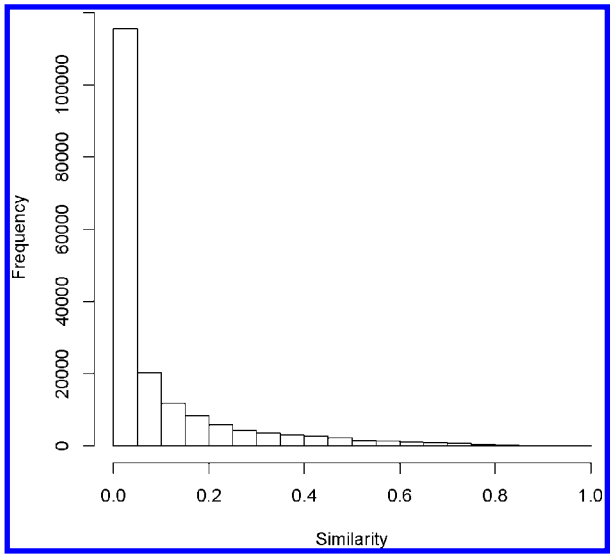

Figure 3. Histogram of all pairwise similarities obtained from the experimental assays (BP similarities) between the compounds in the BioPrint data set. 73% of the compound pairs have a similarity lower than 0.1, while about 1% of the pairs have a similarity of above 0.7.

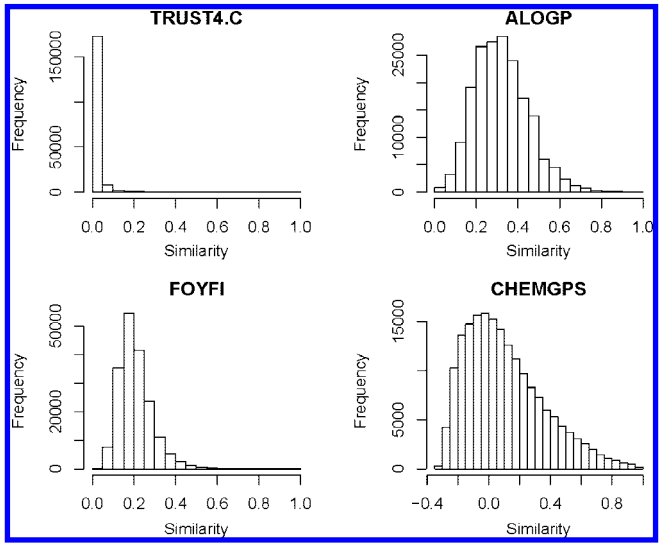

Figure 4. The similarity histograms of selected FP methods. In the case of TRUST4.C 98% of the compound pairs share a similarity of less than 0.1. FOYFI similarities display a narrow normal distribution around a peak in the 0.15−0.2 bin. The ALOGP (peak in the 0.3−0.35 bin) and particularly the CHEMGPS (peak around 0) fingerprints spread their similarity values over the entire range of possible values.

In order to analyze correspondences of the similarity values of the fingerprint methods to the BP similarities we conducted a simple quartile correspondence analysis (see the section entitled Comparison of the FP Similarities to BP Similarities). This analysis is based on the quartile assignments of the single similarity values with respect to the sorted list of nonredundant similarity values of each FP method. We determine the number of agreeing quartile assignments. As a comparison we additionally added two random sets of similarity values: the first (RANDOM.U) randomly assigned a similarity value to each of the compound pairs (range 0−1, uniformly distributed), the second set (RANDOM.B) consists of the similarity values as calculated for the experimental BP similarities, which were then randomly shuffled. In this way the RANDOM.B set had the same distribution of

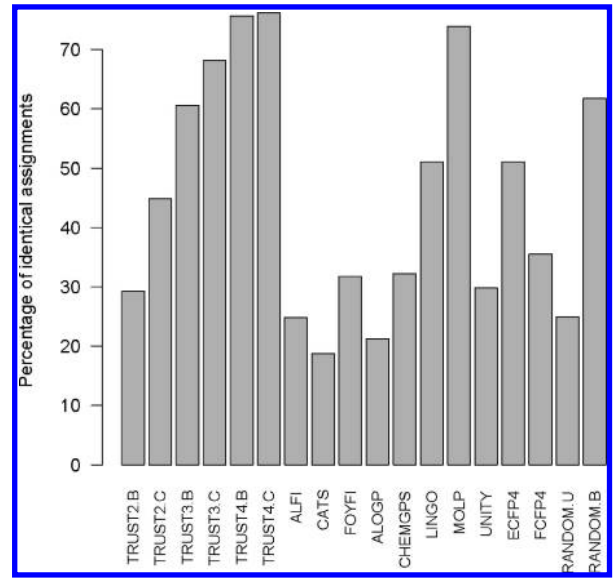

Figure 5. The diagram shows the percentage of identical quartile assignments of the similarity values for each FP method with respect to the BP similarity.

similarity values as the BP similarities but differs regarding the assignments to compound pairs. Each of the two RANDOM set types was generated 20 times.

Figure 5 shows the percentage of identical similarity quartile assignments between each FP method and the similarity assignments based on the BP similarity. According to this analysis the TRUST4 fingerprints (both in binary and count versions) performed best with 76% identical assignments, followed by MOLPRINT2D (73.89%) and TRUST3.C (68.13%). All of them demonstrated a better performance than both random functions, RANDOM.U (25.00% ± 0.08) and RANDOM.B (62.08% ± 0.07), respectively. CATS and ALOGP performed worst in this analysis and had a lower percentage of identical quartile assignments than the randomly assigned similarity values. The comparison of the best performing fingerprint methods to RANDOM.B demonstrates that the better performance is not only due to the principle behavior of these fingerprints that they have most of the similarity values at less than 0.10. TRUST4 and MOLPRINT2D have more pairs in the same quartile with the similarity pairs from BioPrint than with RANDOM.B. It should be noted that this finding does not imply that all remaining fingerprint tools perform worse than random, since the knowledge of the BP similarity distribution was used to create RANDOM.B. However, it is clear that the analysis is dependent on the bin distributions shown in Figure 4 (and in the Supporting Information). It is also noted that the results depend on the length of the fingerprint. The probability of common bits set in two compared molecular fingerprints decreases relative to the fingerprint length. A more specific analysis is described in the following sections. In conclusion fingerprints like TRUST4 and MOLPRINT2D have a pairwise similarity distribution that is more similar to the pairwise similarity distribution calculated from bioprofiles (BP similarities) obtained through uniformly *in vitro* screenings.

Comparison of Fingerprint Similarity Based Clustering with the Biological Profile Based Clustering. The second analysis compares the generated clusterings based on FP similarities to a clustering based on the BP similarities.

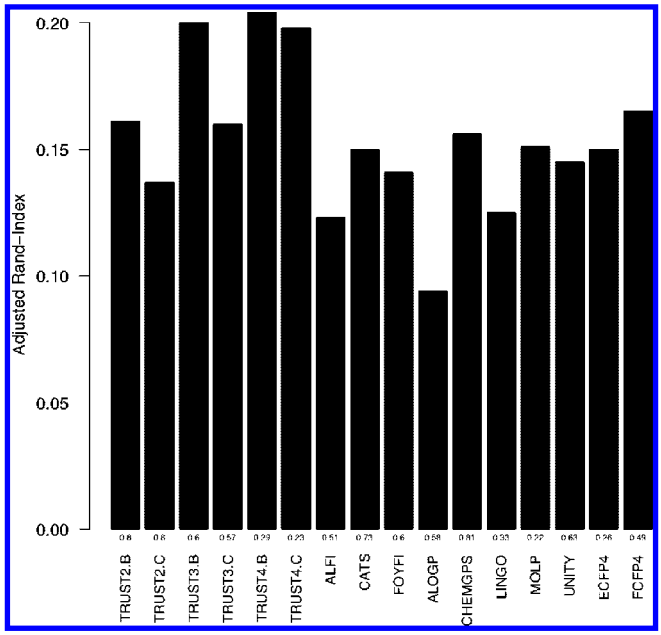

Figure 6. The bars show the value of the adjusted *Rand-Index*; the numbers below indicate the cutoff value where the maximal adjusted *Rand-Index* was found for each method.

The clusterings were compared by the adjusted *Rand-Index* as described in the Materials and Methods section. Each fingerprint clustering was performed on the distance matrix, which was obtained from the similarity matrices by transforming the Tanimoto similarity values into Soergel distances. However, for the ease of the discussion in the text we will refer to the original Tanimoto similarity values.

We applied a hierarchical clustering procedure (complete-linkage) as implemented in the statistical software package R.[29] Cutoff values for obtaining cluster classes had to be defined at which the calculated clustering trees were cut. For the BP similarity clustering (the reference) we set the cutoff value to 0.7. This cutoff value was chosen in order to guarantee a reasonable number of non-singletons (54% of the molecules have at least one nearest neighbor with a similarity equal to or greater than 0.7), while still assuring that only compounds are grouped that share similar activities in the same assays. Figure 6 shows the maximal adjusted *Rand-Index* derived for each of the fingerprint methods as well as the cutoff values at which the maximal adjusted *Rand-Indices* were found. TRUST4.B exhibits the highest adjusted *Rand-Index* (0.2) at a cutoff value of 0.29. The lowest adjusted *Rand-Index* was obtained for the ALOGP fingerprint (0.09) at a cutoff value of 0.58. All generated clusterings based on fingerprints clearly outperform clusterings produced on randomly generated similarities (RANDOM.U: Rand-Index 0.002 ± 0.003 at cutoff 0.49 ± 0.41, RANDOM.B 0.003 ± 0.002 at cutoff 0.34 ± 0.37). The results indicate that pharmacophore-based fingerprints are better in describing the similarity between structurally diverse sets of compounds with similar binding profiles than structural fingerprints. The overall rather low level of cluster class agreement, however, might have several reasons. It could be suggested that in general the rather simple way of describing molecules in fingerprint methods is not particularly suited for mirroring the complex nature of biological profile similarity in many cases. A typical example where computational similarity methods fail is, for example, when two compounds modulate a protein at different binding sites (see Prediction of Biological Profiles). It has to be stated that biological similarity does not necessarily determine structural similarity. This reverse of the similarity principle was also observed by others.[15,18]

Nevertheless, even though the overall cluster class agreements between the molecular fingerprints and the fingerprints created from the BioPrint database are low, there is a marked difference in how well the different molecular fingerprints agree with the bioactive profiles from BioPrint.

Prediction of Biological Profiles. The BioPrint data offer a new possibility for assessing how well biological profiles of molecules can be predicted with *in silico* fingerprint tools. Here, we tested how well a specific fingerprint can retrieve biologically similar compounds from the BioPrint database. All compound pairs were defined as similar in the biological profile space (positive data points in classification terminology), when they exhibited a similarity of greater than or equal to 0.7 (see Comparison of Fingerprint Similarity Based Cluster with the BP Based Clusters and Figure S5 in the Supporting Information). 1772 (out of 184,528) compound pairs in the BioPrint database fulfilled this criterion and were defined as the positive set (BP similar compound pairs). The choice of the cutoff value that separates the BP similar compound pairs from the BP dissimilar compound pairs is somewhat arbitrary. It is clear from the definition that the closer this cutoff value is to 0, the higher all resulting *F-Scores* are, since an increasing number of compound pairs are defined as the positive set. We analyzed the dependence of our results to the choice of the cutoff value, to check if the results were sensitive to the cutoff value, and found that the qualitative results were independent of the cutoff value (see Figure S5 in the Supporting Information). Figure S6 in the Supporting Information shows the number of assays in which both of the compounds in the pairs with a BP similarity above the cutoff value are active in.

The results for the retrieval of compound pairs exhibiting a similarity of equal to or above 0.7 are shown in Figure 7. The numerical values on the *x*-axis denote the cutoff values for which the maximal *F-Scores* were found. In this analysis TRUST4.B shows the best performance with an *F-Score* value of 0.25 at a similarity cutoff of 0.14, followed by TRUST4.C (0.23/similarity cutoff 0.19), TRUST3.B (0.22/similarity cutoff 0.36) and CHEMGPS (0.21/similarity cutoff 0.79). Of the structural fingerprints, ECFP4 performed best (0.19/similarity cutoff 0.26). The worst performance was, as expected, observed for ALOGP (0.1/similarity cutoff 0.69).

For comparison we also computed the *F-Scores* for the two random subsets. The F-Score for the RANDOM.U subset was 0.0195 ± 0.0004 (*recall*: 0.6444 ± 0.2889, *precision*: 0.0100±0.0001, cutoff: 0.40±0.29) and for RANDOM.B 0.0195 ± 0.0005 (*recall*: 0.4494 ± 0.2803, *precision*: 0.0101 ± 0.0005, cutoff: 0.1 ± 0.11) indicating that all *in silico* fingerprints exhibit a certain specificity that mirrors biological activity similarity.

The four best performing similarity tools all represent molecules in a manner that abstracts from the molecular structure (such as pharmacophores or physicochemical properties), whereas fingerprints based on structural fragments such as FOYFI and Unity2D perform worse. It is interesting to compare the binary and the count versions ('.B'

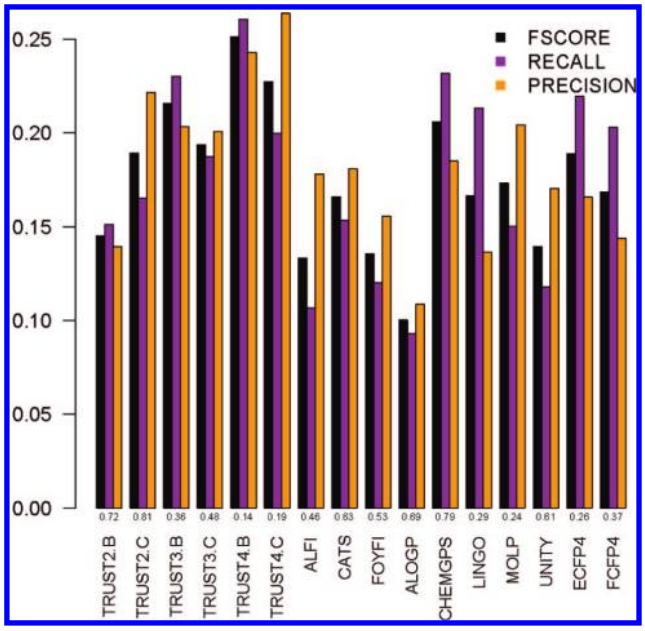

Figure 7. The obtained maximal *F-Score* (with corresponding recall and precision values) for each fingerprint method with respect to the retrieval of compound pairs exhibiting a BP similarity of above or equal to 0.7. The numerical values on the *x*-axis denote the cutoff value with the maximal *F-Score.*

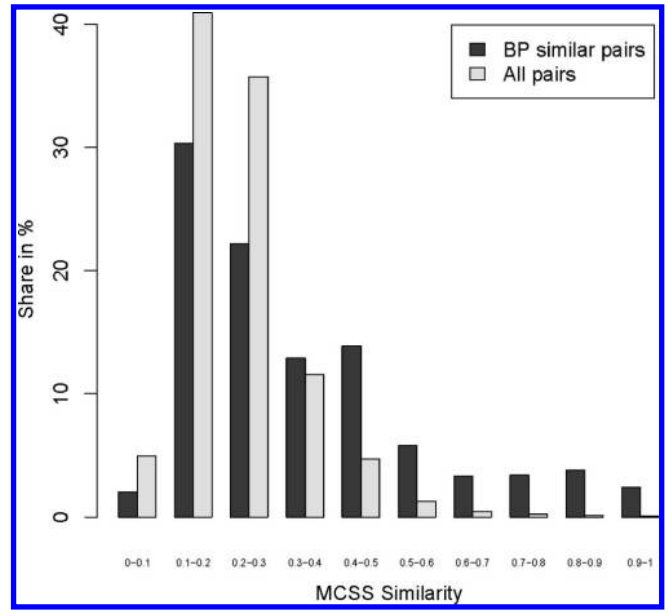

Figure 8. Histogram of the distribution of the MCSS similarity values among the compound pairs that exhibit a BP similarity of above 0.7 (dark gray) and between all compound pairs (light gray).

vs '.C') of TRUST2, TRUST3 and TRUST4, since the pharmacophore definitions and the distance bins are identical, but they differ in the number of points considered for the pharmacophores. TRUST2.C, which takes into account not only which bin are occupied but also how many times the bin are occupied, clearly outperforms TRUST2.B, which only takes into account which bins are set or not. Thus, for the 2-point pharmacophore fingerprint it is important to take into account how many times a bin is set in order to obtain specificity. Interestingly, for the 3- and 4-point pharmacophore fingerprints the opposite trend is seen, and the binary description performs better than the continuous.

To further understand the relationship between structural similarity and the BP similarity we computed the MCSS similarity as defined in eq 6. Figure 8 shows the distribution of MCSS values among those compound pairs that have a BP similarity of equal or above 0.7 in comparison to the distribution calculated for all compound pairs in the data set. Although a clearly higher share of BP similar compound pairs compared to all compound pairs have a MCSS Tanimoto similarity of above 0.5, still 80% of the BP similar compound pairs show a MCSS Tanimoto similarity of below 0.5. Thus, similar biological profiles do not necessarily correspond to a structural similarity as defined by the Tanimoto MCSS in eq 7. The majority of the compounds having a high BP similarity have a low MCSS similarity, indicating that they have a rather low structural similarity.

In general the analysis is hampered by the uncertainty about the binding modes of the molecules. Although two molecules bind to the same target and therefore exhibit a high BP similarity, they might bind to different pockets of the protein or have different binding modes in the same pocket. In all these cases, a similarity-based method cannot be expected to retrieve a compound pair with a high BP similarity. The compound pairs that were not retrieved from any of the fingerprint methods have an average MCSS similarity of 0.24 ± 0.11, while the compound pairs that were retrieved by all methods (in total 35 pairs) exhibited an average MCSS similarity of 0.86 ± 0.15. In the following examples we highlight some of compound pairs that were never found as similar by any *in silico* fingerprint method (see Figure 9). Diclofenac and bevonium (see Figure 9a) both bind only to the histamine type-1-receptor and therefore have an identical BP similarity. Their structures are clearly different and possibly their binding modes differ. Cyclopen-

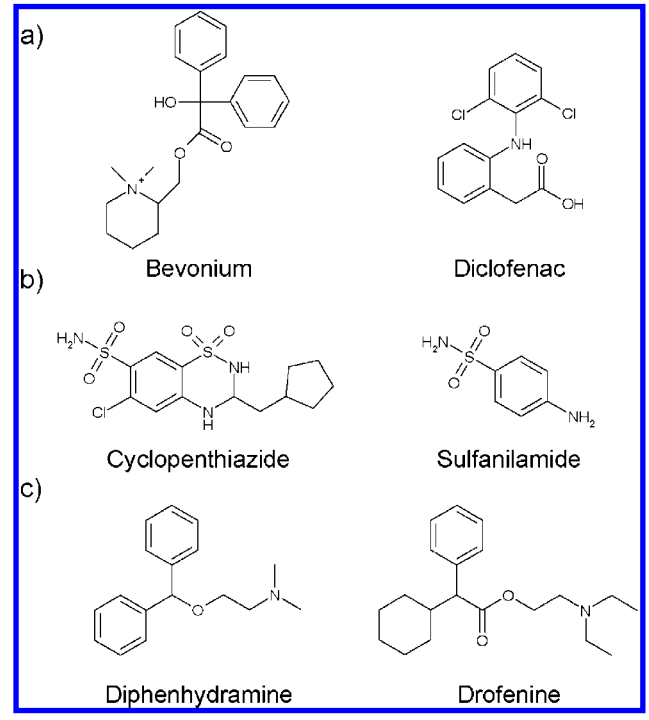

Figure 9. Examples of compound pairs, which were not retrieved by any of the fingerprint methods. a) Bevonium (IC_{50}=8.9 μm) and diclofenac (IC_{50}=11.0 μm) bind to the human H1 receptor, b) cyclopenthiazide (IC_{50}=2.0 μm) and sulfanilamide (IC_{50}=1.7 μm) inhibit the carbonic anhydrase II enzyme, and c) diphenydramine and drofenine both bind to a panel of assays.

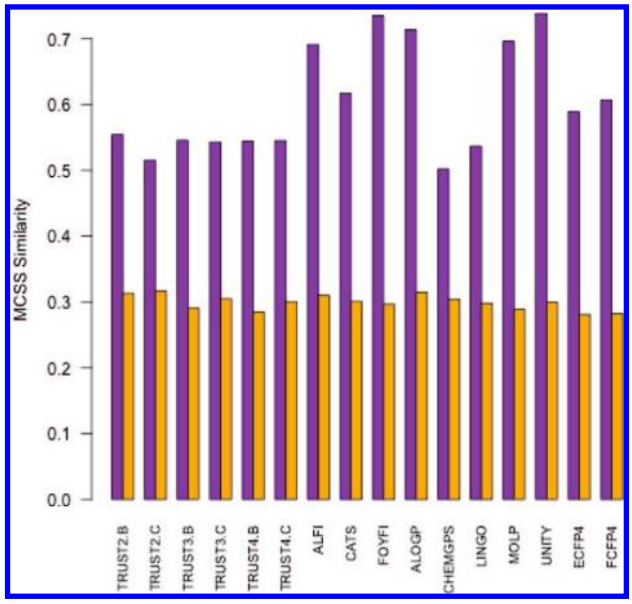

Figure 10. The average MCSS similarity values for the pairs that were retrieved (violet) and for those that were not retrieved (orange). The retrieved BP similar pairs of the pharmacophore based fingerprint methods as well as the property based fingerprint CHEMGPS show a signficantly less strong dependence on the MCSS similarity than the BP similar pairs retrieved by structural fingerprints.

Table 2. Correlation of the Fingerprint Similarities to the MCSS Similarities[a]

fingerprint	r_P^2	fingerprint	r_P^2
TRUST2.B	0.25	FOYFI	0.68
TRUST2.C	0.21	ALOGP	0.49
TRUST3.B	0.34	CHEMGPS	0.17
TRUST3.C	0.3	LINGO	0.34
TRUST4.B	0.36	MOLP	0.6
TRUST4.C	0.34	UNITY	0.65
ALFI	0.41	ECFP4	0.51
CATS	0.36	FCFP4	0.56

[a] Considering only the retrieved BP similar pairs.

tathiazide and sulfanilamide (see Figure 9b) are both inhibitors of carbonic anhydrase II and bind to the calcium ion. Sulfanilamide is a substructure of cyclopentathiazide. Molecules with similar activities but of different sizes might not be detected as similar by fingerprint methods if Tanimoto similarity is applied. Interestingly none of the methods retrieved the structurally closely related compound pair diphenhydramin−drofenine (in the case of LINGO the calculated similarity was however close to the identified cutoff). The two compounds bind to a panel of identical assays.

To understand how different fingerprint methods represent compounds we compared the MCSS similarity values and the similarity values for the BP similar pairs that were retrieved of each fingerprint method individually (see Figure 10). For all *in silico* fingerprint methods the average MCSS similarity of the compound pairs that were found is significantly higher than of those which were not found (p-value $<2.2*10^{-16}$, Wilcoxon-Mann−Whitney test). The pharmacophore-based fingerprint methods as well as the property-based fingerprint CHEMGPS are significantly less dependent on MCSS similarity and accordingly show a lower average MCSS similarity for the retrieved BP similar compound pairs

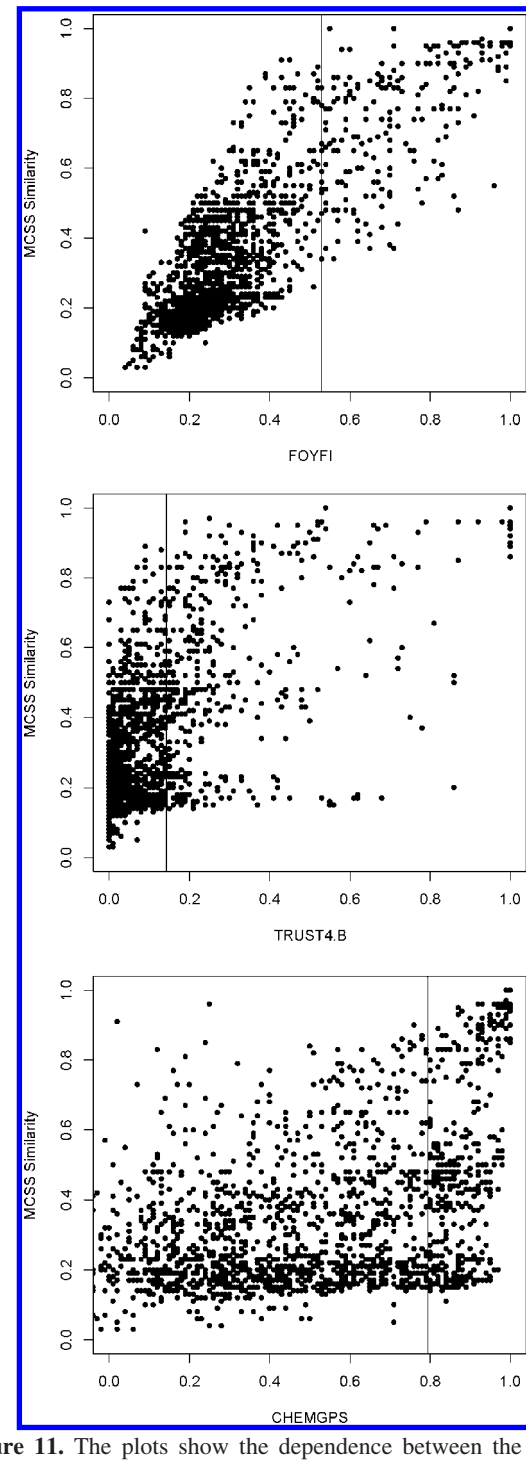

Figure 11. The plots show the dependence between the MCSS similarity and the fingerprint similarity values for the compound pairs having a BP similarity of equal or greater than 0.7. The vertical lines denote the cutoff values for which the optimal *F-Score* was found.

compared to structure based fingerprints such as FOYFI and UNITY (the pairwise p-values computed with the Wilcoxon-Mann−Whitney test are reported in Table S2 in the Supporting Information) Subsequently, we calculated the correlation coefficients between the MCSS similarity and fingerprint similarity for the retrieved BP similar compound pairs for each fingerprint (see Table 2 and Figure 11). The FOYFI fingerprint has a squared linear correlation coefficient r_p^2 of 0.68 with the MCSS similarity. In Figure 11 it is shown that among the retrieved pairs with FOYFI none had a MCSS Tanimoto of less than 0.3, while both TRUST4.B and

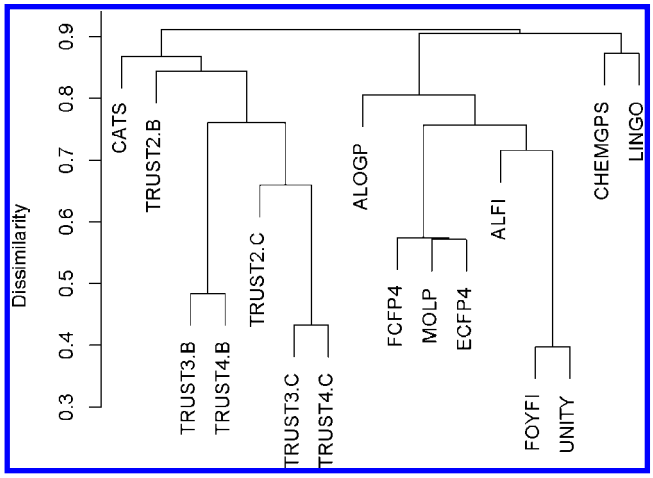

Figure 12. Hierarchical tree of the fingerprint methods derived from similarities with respect to the retrieval of common compound pairs at the maximal F-Score.

CHEMGPS retrieved a larger amount of BP similar compound pairs having a low MCSS similarity.

Our analysis has shown that the investigated fingerprint methods in general capture different aspects of the compounds and therefore retrieve different compound pairs. Figure 12 shows how the fingerprint methods relate to each other by means of a hierarchical tree (*complete-linkage* clustering). This tree was obtained by calculating a Tanimoto similarity on the basis of identical and non-identical compound pairs retrieved by two methods at the maximal *F-Score*. Not surprisingly, the tree reflects the relations as expected from the underlying methodologies. The TRUST3.C and TRUST4.C fingerprints cluster together at a similarity level of 0.55; FOYFI and UNITY show a preference for similar pairs and share a Tanimoto similarity of 0.6. The extended connectivity fingerprints and MOLPRINT2D cluster together at a similarity level of approximately 0.4. LINGO and CHEMGPS show the most orthogonal behavior to all other fingerprints indicating that they represent molecules in a different way compared to other fingerprints.

CONCLUSIONS

A comparison of similarity as experimentally measured in biochemical assays and as computed with *in silico* fingerprints was performed. We based our analysis on the BioPrint database, which contains molecules that are of relevancy for the pharmaceutical industry. We therefore consider our findings as relevant for drug molecules of current interest for drug design. The investigation indicated that fingerprints that describe molecules in a more abstract way based on their physicochemical properties (CHEMGPS) or their pharmacophores (TRUST) are generally able to retrieve more and different BP similar compound pairs than fingerprints based on structural features. It is interesting to contrast the findings in this paper with other publications, which have focused on choosing a set of active molecules for a specific protein and trying to retrospectively retrieve the active compounds from a large set of assumed inactive compounds. In these studies structural fingerprints usually perform very well. This is not surprising in the case when these series contain molecules which are structurally very similar. It should be pointed out that this behavior is very desirable in cases where near-neighbors of an active molecule are sought. For instance, if near-neighbors are sought for a screening hit in either the corporate compound collection or among commercially available compounds, structural fingerprints are a fast way to retrieve these compounds. However, as it has been described in this article more "fuzzy" fingerprints are better than structural fingerprint in retrieving compounds with similar bioactive profile but with different chemical structures.

ACKNOWLEDGMENT

We thank Dr. Mike Rolf for providing us with the BioPrint data set. Dr. Dave Cosgrove (FOYFI, ALFI), Dr. Andrew Grant (LINGO), and Dr. Andreas Bender (MOLPRINT2D) are gratefully acknowledged for providing us with programs to calculate the molecular fingerprints. Dr. Claire Gavaghan, Dr. Niklas Blomberg, Dr. Sorel Muresan, Dr. Stefan Schmitt, and Dr. Dave Cosgrove are acknowledged for valuable suggestions.

Supporting Information Available: Similarity histogram plots of all fingerprint methods and some further statistical assessment. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45* (19), 4350–4358.
(2) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.
(3) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7* (17), 903–911.
(4) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1177–1185.
(5) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 338–345.
(6) Rarey, M.; Dixon, J. S. Feature trees: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12* (5), 471–490.
(7) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99* (11), 3503–3510.
(8) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519.
(9) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42* (17), 3251–3264.
(10) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157–166.
(11) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25* (12), 1162.
(12) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46* (3), 1201–1213.
(13) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6* (4), 470–480.
(14) Strausberg, R. L.; Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **2003**, *300* (5617), 294–295.
(15) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.* **2007**, *47* (2), 325–336.

(16) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **2005**, *1* (7), 389–397.

(17) Fliri, A. F.; Loging, W. T.; Volkmann, R. A. Analysis of system structure-function relationships. *ChemMedChem* **2007**, *2* (12), 1774–1782.

(18) Horvath, D.; Jeandenans, C. Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces--a benchmark for neighborhood behavior assessment of different in silico similarity metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 691–698.

(19) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem.* **1998**, *102* (21), 3762–3772.

(20) Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45* (2), 386–393.

(21) Grant, J. A.; Haigh, J. A.; Pickup, B. T.; Nicholls, A.; Sayle, R. A. Lingos, finite state machines, and fast similarity searching. *J. Chem. Inf. Model.* **2006**, *46* (5), 1912–1918.

(22) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38* (19), 2894–2896.

(23) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(24) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.* **2006**, *46* (3), 991–998.

(25) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.

(26) Pipeline Pilot. In *Basic Chemistry User Guide;* Scitegic Inc.: San Diego, 2007.

(27) Kogej, T. Manuscript to be published.

(28) Miller, I.; Miller, M. *John E. Freund's Mathematical Statistics,* 6th ed.; Prentice-Hall, Inc.: Upper Saddle River, 1999.

(29) R Development Core Team. *A Language and Environment for Statistical Computing, R package version 2.5.0*; R Foundation for Statistical Computing: Vienna, Austria, 2007.

(30) Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2* (1), 193–218.

(31) Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of successful lead generation. *Curr. Top. Med. Chem.* **2005**, *5* (4), 421–439.

CI800326Z