

Multifingerprint Based Similarity Searches for Targeted Class Compound Selection

Thierry Kogej, Ola Engkvist, Niklas Blomberg, and Sorel Muresan*

AstraZeneca R&D Mölndal, GDECS Computational Chemistry, Pepparedsleden 1, 431 83 Mölndal, Sweden

Received October 27, 2005

Molecular fingerprints are widely used for similarity-based virtual screening in drug discovery projects. In this paper we discuss the performance and the complementarity of nine two-dimensional fingerprints (Daylight, Unity, AIFi, Hologram, CATS, TRUST, Molprint 2D, ChemGPS, and ALOGP) in retrieving active molecules by similarity searching against a set of query compounds. For this purpose, we used biological data from HTS screening campaigns of four protein families (GPCRs, kinases, ion channels, and proteases). We have established threshold values for the similarity index (Tanimoto index) to be used as starting points for similarity searches. Based on the complementarities between the selections made by using different fingerprints we propose a multifingerprint approach as an efficient tool to balance the strengths and weaknesses of various fingerprints.

INTRODUCTION

Similarity searching techniques are increasingly important for the ligand based virtual screening of large compound collections.¹ Similarity searches are based on the observation that, intuitively, similar chemical structures often display activity toward the same biological targets.^{2,3} Similarity search methods are easy to handle and usually fast, and they do not require other information besides the chemical structure of the molecules. Thus these methods are ideally suited to define targeted screening sets, to follow up on actives from high-throughput screening, and for library design from virtual combinatorial libraries. Fingerprints, i.e., vectors where each element encodes some aspects of molecular structures, are commonly used as the input for similarity calculations since, once defined and generated, they can be quickly compared by using a similarity index. One of the commonly used similarity indexes is the Tanimoto index, but there are many other similarity measures (see, for example, the work of Willett et al.^{4,5}). Notably the efficiency of similarity searches depends not only on the fingerprints employed but also on the chosen similarity index.

Sheridan et al. have investigated different methods for similarity selection.⁶ They have shown that different methods select different active molecules, and, in most cases, different fingerprints will select complementary actives. This is an important observation since one possible drawback of fingerprint based virtual screening is the risk of missing structurally dissimilar actives. Another known issue with fingerprint based similarity searching is the fact that the Tanimoto coefficient does not correlate between different fingerprint types. Martin et al.² have shown that using Daylight fingerprints a Tanimoto similarity of 0.85 corresponds to a 30% chance that two compounds would have the same biological activity, but this is not valid for other fingerprints.^{2,7} Some studies have already established similarity threshold values for some fingerprint types.^{8–10} Here, we set out to systematically address the biological significance

Table 1. Name, Type, and the Number of Bins for the Different Fingerprints Used in the Present Study

name	type	number of bins
AIFi	dichotomous	2048
Daylight	dichotomous	1024
Unity	dichotomous	992
Hologram	continuous (integer)	258
ALOGP	continuous (integer)	133
ChemGPS	continuous (real)	9
CATS	continuous (integer)	155
TRUST	continuous (integer)	12005
Molprint 2D	continuous (integer)	varies according to molecular complexity

of similarity thresholds for a given set of 2D-fingerprints by establishing a framework where such thresholds can be easily determined. In addition we wanted to investigate the complementarity of different fingerprints in retrieving active molecules to create an “optimal” combination of fingerprints for our virtual screening projects.

Here, we present the results of this retrospective analysis of the Tanimoto index similarity for a set of different 2D (two-dimensional) fingerprints. For our study we used four sets of compounds with binding affinity data against four superfamilies of target proteins (GPCRs, ion channels, kinases, and proteases). We analyzed the efficiency of each fingerprint in retrieving active compounds, the degree of overlap between the selected sets, and the number of actives found by using different fingerprints.

METHODS

Fingerprints Used in the Study. A generalized 2D fingerprint is a vector where each element holds information about the 2D structure. In Table 1 the name, the type, and the number of bins (i.e. bits for binary fingerprints) for the different fingerprints used in the present study are listed. The term “dichotomous” refers to binary fingerprints, i.e., taking values 0 or 1; “continuous” fingerprints contain bins that are either counts (integers) or real numbers. Daylight,¹¹ Unity

* Corresponding author e-mail: sorel.muresan@astrazeneca.com.

2D,¹² and AlFi¹³ fingerprints are binary hashed fingerprints. In short, all possible paths in a molecule are enumerated up to a predefined length (number of bonds). Atom information (e.g. type, charge, hybridization etc.) is recorded and recursively updated via integer hashes. The resulting, large, integer is iteratively divided by fingerprint length, and the remainder is used to set the corresponding bit. Usually several bits are set for a given path, and consequently there is no direct correspondence between a specific bit and an atom or substructure. Whereas the Daylight and Unity fingerprints encode the atomic information in terms of atomic number etc., the AlFi fingerprint allows for the atoms to be described according to functional properties, as defined by the user, using SMARTS¹⁴ patterns. In the current work, SMARTS definitions were used to describe the atoms in terms of hydrogen-bond donor and acceptor abilities, acid or basic tendencies, and whether they are aliphatic or aromatic. Under this scheme, for example, phenol and aniline have very similar fingerprints as both consist of an aromatic ring and a donor/acceptor. So that they do not produce identical fingerprints, 9 bits in the fingerprint are set aside to indicate the presence of the elements carbon, nitrogen, oxygen, sulfur, fluorine, chlorine, bromine, iodine, and 'other'. The Hologram fingerprint¹⁵ counts the occurrence of substructural fragments/atom paths in the molecule. Some parameters can be fine-tuned to generate different holograms for the same molecule. Here a single set of parameters¹⁵ is used which gives a good compromise between the level of discriminative power of the Hologram and the computational effort required (data not shown). The ALOGP fingerprint counts the occurrence of 133 atom types defined by Crippen et al. for estimating logP values.¹⁶ The ChemGPS method was developed by Oprea et al.¹⁷ as a rapid alternative to principal component analysis (PCA) for large data sets. The fingerprint bins corresponds to 9 'chemical' coordinates obtained by projecting molecular properties into an orthogonal physicochemical space defined by a set of 'core' and 'satellite' compounds. The CATS fingerprint¹⁸ comprises the occurrence of topological distances (namely, the number of bonds) between pairs of 5 pharmacophoric features up to 10 bonds apart. Only the shortest number of bonds between two given features is considered. In our implementation the counts for the five features (donor, acceptor, anion, cation, and hydrophobic) are also included in the fingerprint. The TRUST fingerprint¹⁹ records 3-point topological pharmacophore distances. TRUST is similar to the Similog keys published by Novartis;²⁰ however, slightly different pharmacophore features have been used in our work (hydrogen-bond donor and acceptor, lipophilic, positive, and negative charge). As for the CATS fingerprint, the distance between two pharmacophore features is calculated as the number of bonds between them. The distances are binned into seven distance bins, corresponding to distances of 2,3,4,5,6,7–8 and 8–10 bonds, respectively. A vector is calculated for each molecule, where each bit in the vector corresponds to a specific combination of three pharmacophore features and the corresponding binned bond distances between them. The algorithm is implemented in C++ using OpenBabel.²¹ Molprint 2D fingerprints have been developed by Bender et al.^{22,23} and are 2D connectivity fingerprints that encode the atom environment up to a specified distance (two in the present study).

Similarity Index. We used the Tanimoto similarity index to assess molecular similarity. The Tanimoto index between two molecules described by a fingerprint with “*n*” bins is calculated using the following expression

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

Thus, for continuous vectors Tanimoto index ranges from –0.3 to 1 where 1 denotes identity. In the case of dichotomous fingerprints, the expression can be simplified to

$$T_{AB} = \frac{c}{a + b - c}$$

where “*c*” is the number of bits common to the two molecules, and “*a*” and “*b*” denote the number of bits set in each of the two fingerprints. For binary fingerprints the Tanimoto index range is 0 to 1.

Examples of Tanimoto indices between four pairs of compounds are depicted at Table 2 for the different fingerprints. A number of observations can be made from this table; for instance, the Tanimoto values increase from the first pair (first row of Table 2) to the second pair of compounds using the Daylight or the ALOGP fingerprints, whereas they decrease using the Hologram or the ChemGPS fingerprints. To the CATS fingerprint, the molecules in the third row are identical (i.e. Tanimoto index equals 1) since the thiophene and the phenyl ring are associated with the same feature (i.e. hydrophobic). As this modification also does not change the physicochemical properties significantly, the ChemGPS Tanimoto index is also equal to 1. The ALOGP based fingerprint is insensible to the isomerization (see fourth row of the Table 2).

Clustering. We have chosen to use Daylight fingerprints to perform the cluster analysis and near-neighbors mapping of the target sets. This was done with an in-house program, FLUSH,²⁴ which uses the Taylor sphere exclusion algorithm.²⁵ A Tanimoto cutoff value of 0.80 was chosen to define the cluster size.

Workflow. The data for this study was taken from AstraZeneca's screening database. We have chosen to pool the compounds and the associated data into target class data sets. The analysis and results in this paper reflect the common scenario of using known actives from homologous proteins to virtually screen for hits toward a novel target. Table 3 shows the number of different screening assays for each class of proteins. Confirmed IC₅₀s were used to classify a compound as active or inactive. A compound is considered active for a class of proteins if it satisfied the project defined threshold for at least one of the screening assays within a class. Figure 1 illustrates the workflow followed to obtain the similarity data presented in this work. The workflow involves four main steps. In the first step data sets were collected for each of the four classes of compounds. For ease of comparison we defined a uniform active/inactive distribution for all target classes. Hence, all *target* sets contain 10 000 compounds (9500 inactives and 500 actives) selected ran-

Table 2. Tanimoto Indices between Four Pairs of Structures (the First Two Pairs Are 5-HT Reuptake Inhibitors) with Respect to the Different Fingerprints

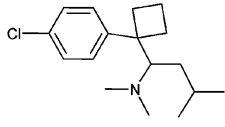
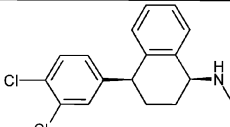
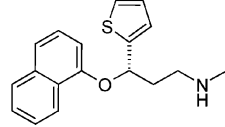
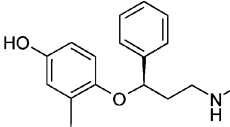
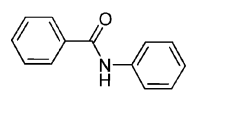
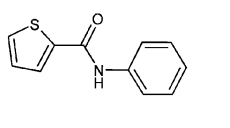
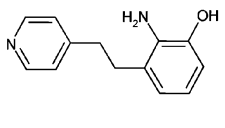
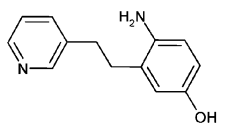
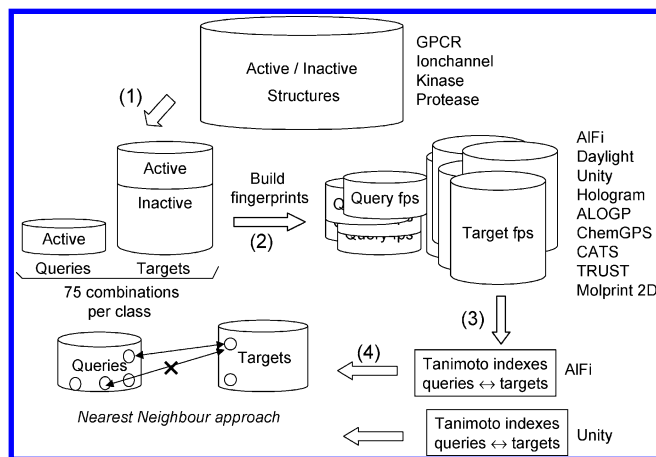
Compound A	Compound B	Fingerprint	Tanimoto Index
 Sibutramine	 Sertraline	AlFi	0.11
		Daylight	0.47
		Unity	0.54
		Hologram	0.41
		ALOGP	0.62
		ChemGPS	0.79
		CATS	0.59
		TRUST	0.72
 Duloxetine	 4-hydroxyatomoxetine	Molprint 2D	0.11
		AlFi	0.30
		Daylight	0.61
		Unity	0.57
		Hologram	0.21
		ALOGP	0.94
		ChemGPS	0.52
		CATS	0.60
 N-phenylbenzamide	 N-(thiophen-2-yl)benzamide	TRUST	0.59
		Molprint 2D	0.18
		AlFi	0.57
		Daylight	0.59
		Unity	0.54
		Hologram	0.62
		ALOGP	0.86
		ChemGPS	1.00
 2-amino-3-(pyridin-2-ylmethyl)phenol	 3-amino-4-(pyridin-2-ylmethyl)phenol	CATS	1.00
		TRUST	0.87
		Molprint 2D	0.19
		AlFi	0.28
		Daylight	0.82
		Unity	0.82
		Hologram	0.87
		ALOGP	1.00
		ChemGPS	1.00
		CATS	0.58
		TRUST	0.63
		Molprint 2D	0.40

Table 3. Number of Assays for Each Protein Class

class	number of target protein in the class	class	number of target protein in the class
GPCR	102	kinase	60
ion channel	20	protease	25

**Figure 1.** Workflow followed for the retrospective analysis.

domly from the available data, while the *query* sets contain 50 active compounds (not present in the corresponding target set). Thus, all target sets have the same size and a constant ratio of actives and inactives, 5% and 95%, respectively. The query sets are 10% the size of the number of actives in the target set. To avoid spurious similarities by particular compound sets 75 such query/target sets were defined at random for each protein family, and all data showed in this paper are averages over those 75 pairs. Note that while all compounds (actives and inactives) have been tested against individual targets in a given class, they might not have been tested against all the proteins considered in that class. This means that the similarity-based selections made in this study are contaminated by a certain (but constant) number of false

negatives. Furthermore one can reasonably assume that compounds were originally selected and tested for their ability to mimic active ligands. Thus a proportion of the inactive compounds might present some pharmacophoric features or substructures that are required for binding the target class proteins. Such compounds would show up as false positives and lower the overall hit-rate for that class. Given the fixed a priori distributions these issues should not influence the definition of robust similarity thresholds for the different fingerprints for virtual screening. Our assessment of fingerprint complementarity should not be greatly influenced by these issues.

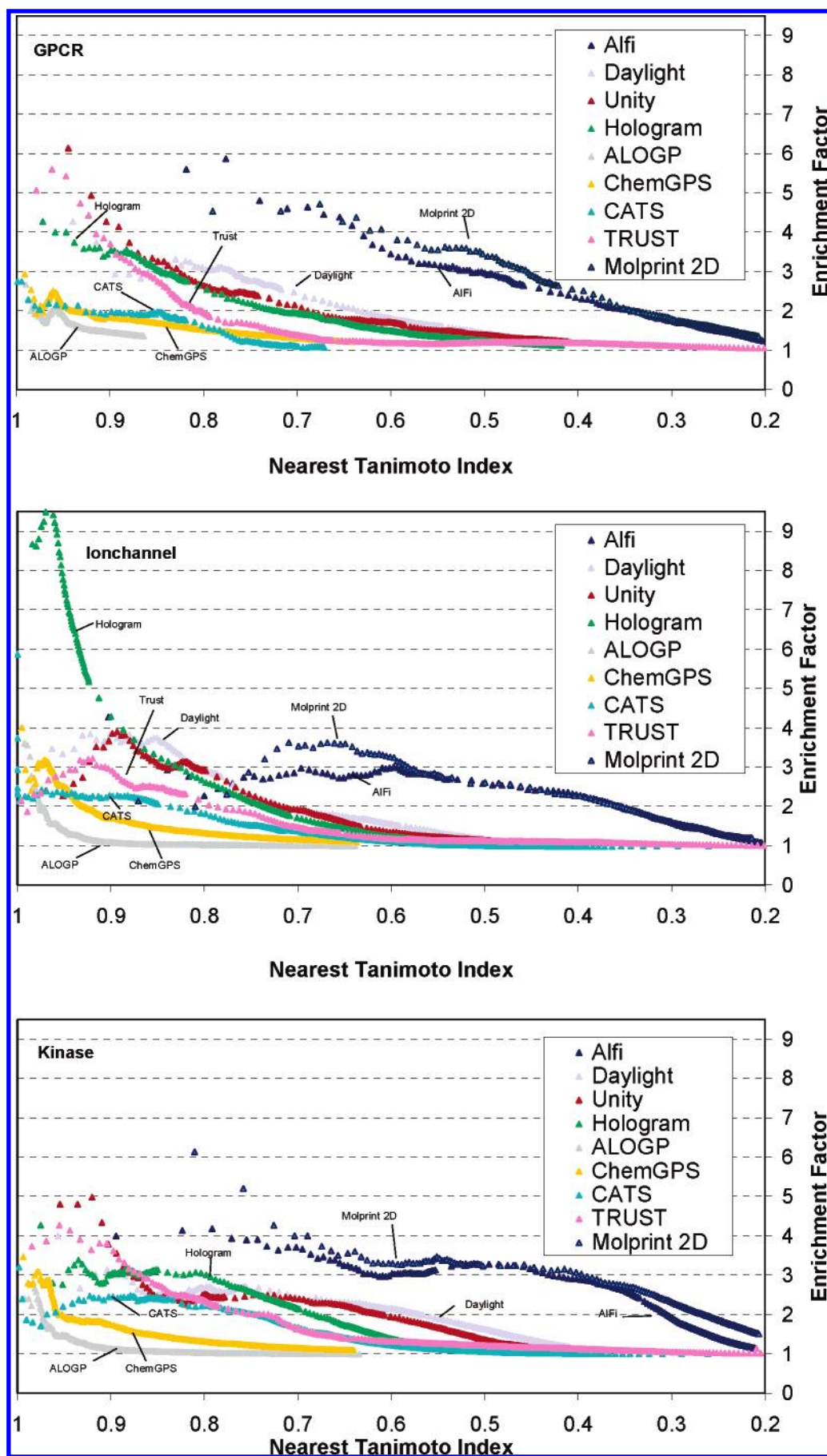
In the second and third steps of our workflow we generated all fingerprint types for all data sets and computed the Tanimoto index between all query and target molecules within a query-target set. Finally, in the fourth step, only the most similar (highest Tanimoto index) query for a target was stored per fingerprint type, i.e., “nearest neighbor approach”.^{2,3} The Tanimoto index between a query molecule and its nearest target set neighbor is referred to as the “nearest Tanimoto index”.

RESULTS AND DISCUSSION

Enrichment Factor. To analyze the relative efficiency of the different fingerprints to select active compounds from the target set, we computed enrichment factors using the following expression

$$E = \frac{\text{number of actives selected}}{\text{number of targets selected} \cdot R}$$

where R is the ratio of active to inactive compounds in our target set, i.e., 0.05. Thus, 20 is the maximum achievable enrichment, and a random selection would give an enrichment factor of 1. The efficiency of a selection technique is often assessed by plotting the number of active compounds



found with respect to the percentage of the target database screened. This is useful when focusing on coverage of actives

as it shows what percentage of the target set needs to be screened to retrieve a certain proportion of the actives.

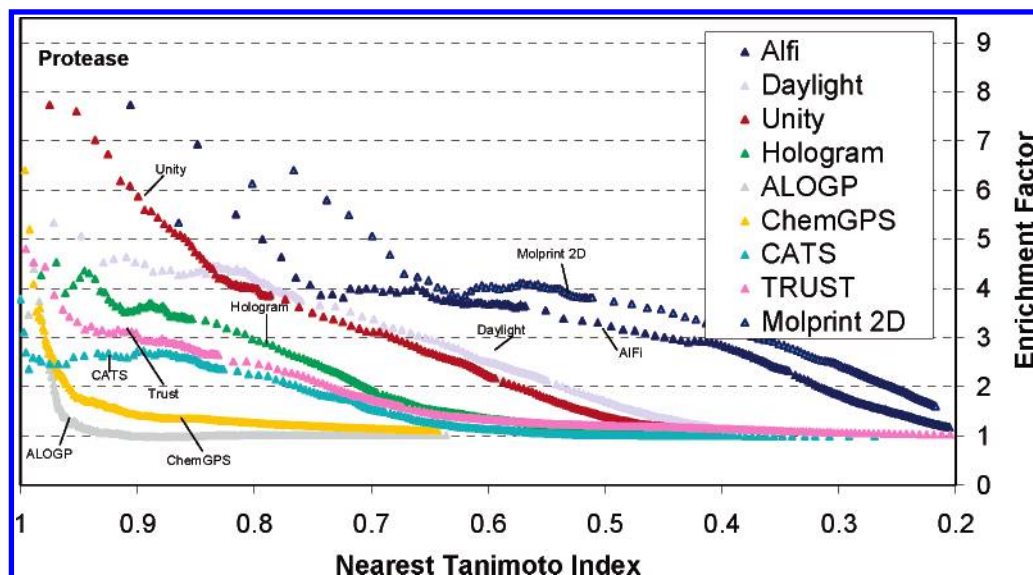


Figure 2. Enrichment factors versus the nearest neighbor Tanimoto index for the different fingerprints and target classes. The enrichment factors are the average values based on 75 different query-target set combinations.

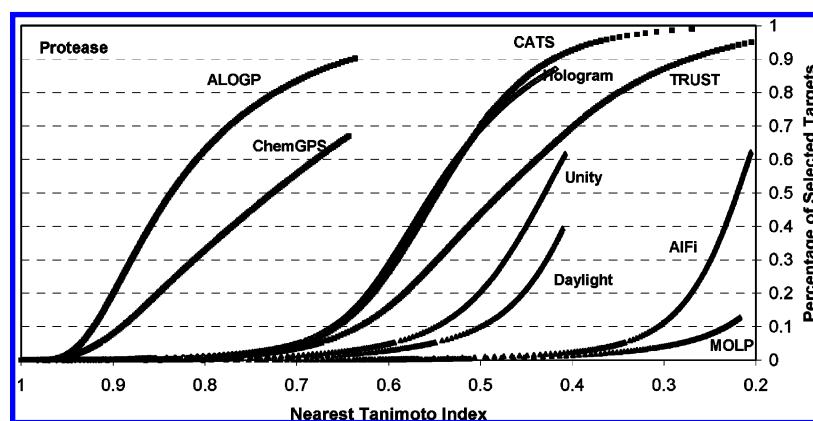


Figure 3. Percentage of selected targets versus the ranked highest Tanimoto index between the target compounds and the query set for the different fingerprints in the case of the protease target class.

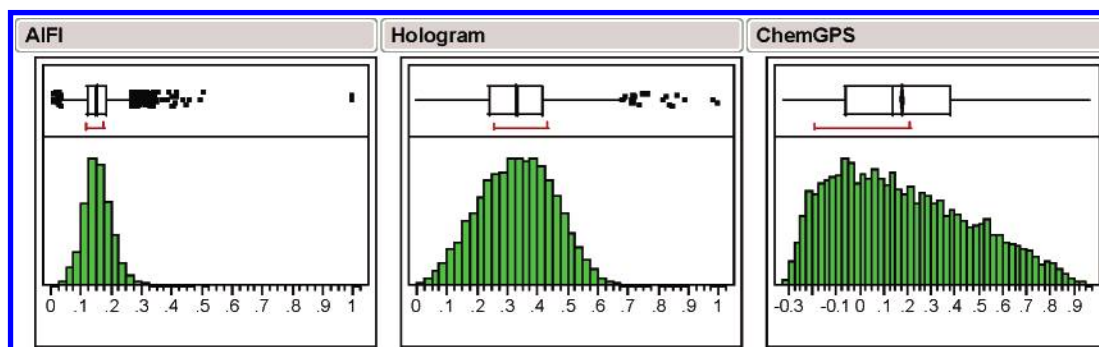
However, the efficiency with which a given fingerprint recovers actives gives little information about the reliability of a given selection when applied in a prospective setting. Indeed, selecting the top 1%, 5%, etc. from a database will not necessarily find any hits since the selected compounds can be very dissimilar. This risk is enhanced in the case where the query and/or the target sets present a poor molecular diversity. Thus one of the goals of this study is to establish a threshold value for the Tanimoto index defining (biologically) relevant similarity between the target and the query (for a given fingerprint). These thresholds can be used as guidelines when searching for new molecules in a project. For this reason we investigated how enrichment factors vary as a function of molecular similarity (here the Tanimoto index between a given target and the nearest neighbor query compound, averaged over the 75 query/target sets). This plot emphasizes the Tanimoto threshold values for the different fingerprints. In Figure 2 this is shown for compounds from all target classes.

Reassuringly, all fingerprint curves decrease when reducing the nearest neighbor Tanimoto index, i.e., all fingerprints can discriminate between active and inactive compounds. Artifacts occur at very high similarity since the number of selected target molecules is very low, thus the enrichment factors are highly influenced by the ratio between actives

and inactives. These artifacts can be explained by the fact that some groups of atoms may increase the overall similarity to the active query compound, while they are detrimental for the activity of the molecule (for example, through steric hindrance or are simply not needed for binding to the protein). Again reassuringly, the different fingerprints show a similar pattern of enrichment for the different classes. Except for the high similarity range for the ion-channel class, the AIfi and the Molprint 2D fingerprints produce higher enrichments over the whole range of similarity compared to the other fingerprints. We observe that only a few target molecules present a Tanimoto index with the query compounds higher than 0.7–0.75 using these two fingerprints. Although this observation does not imply anything about the efficiency of this fingerprint for discriminating between actives and inactives, it emphasizes the possibility of tuning the AIfi and the Molprint 2D parameters in such a way that the target compounds distribute over all the Tanimoto range. We observe the opposite trends in the case of the ALOGP and the ChemGPS fingerprints where a large number of target molecules present a nearest Tanimoto index close to 1. This is shown in Figure 3, where we plot the percentage of targets selected with respect to the nearest Tanimoto index for the protease class (the same pattern is seen for the other target classes). At a Tanimoto value of 0.80, around 30%

Table 4. Cutoff Values for the Tanimoto Index for the Different Fingerprints and Different Classes as Well as Their “Consensus”

	AlFi	Daylight	Unity	Hologram	ALOGP	ChemGPS	CATS	TRUST	Molprint 2D
GPCR	0.34	0.64	0.69	0.72	0.96	0.94	0.93	0.81	0.35
ion channel	0.36	0.72	0.72	0.74	0.98	0.93	0.84	0.80	0.35
kinase	0.31	0.57	0.61	0.68	0.97	0.95	0.68	0.73	0.27
protease	0.32	0.54	0.57	0.71	0.97	0.96	0.71	0.74	0.26
“consensus cutoff”	0.30	0.65	0.65	0.70	0.97	0.95	0.85	0.75	0.30

**Figure 4.** Distribution of 10 000 randomly selected Tanimoto indices between query and target compounds for three fingerprint types in the case of the protease class.

and 60% of the target set compounds are selected in the case of the ChemGPS and ALOGP, respectively. Inversely, at a relatively low Tanimoto value of 0.25, around 30% and 10% of the targets are selected using the AlFi and the Molprint 2D fingerprints, respectively. The other fingerprints show intermediate trends in Figure 3.

The Hologram fingerprint gave very high enrichment in the ion-channel case when considering a Tanimoto index greater than or equal to 0.9. This is probably due to the relatively poor diversity of molecular fragments important for the activity found in the active compounds.

By decreasing the nearest Tanimoto index we increase the selection of inactive compounds. Enrichment equals 1 (i.e. random selection) at different Tanimoto index values with respect to fingerprint type and the target class. We define an arbitrary threshold value of 2 for the enrichment factor; an enrichment value of 2 means here a selection of 40 targets per active compound found. This establishes practical cutoff Tanimoto values for each fingerprint. Table 4 summarizes the cutoff values defined this way for each target class set. We notice different cutoffs for a given fingerprint type through the different target classes. The biggest variations of cutoff value through the different classes of compounds are observed in the case of Daylight, Unity, and CATS. For example, the CATS cutoff is significantly higher for the GPCR than for the other classes showing that very high Tanimoto values should be employed when selecting GPCR analogues by the CATS fingerprint. The other fingerprints show similar cutoff values for the different classes. While this obviously presents difficulties for the establishment of a general similarity cutoff that can be used as a guidance for compound selections, we still suggest the use of “consensus” values for each fingerprint type (Table 4) while bearing in mind that for an individual target set results may vary.

As an example four 5HT-reuptake inhibitors are presented at the first two rows of Table 2 to illustrate similarities with different fingerprints. Using AlFi one finds e.g. duloxetine by a similarity search using the 4-hydroxyatomoxetine compound and our “consensus cutoff”, whereas the other

Table 5. Distribution of 10 000 Randomly Selected Tanimoto Indices between Query and Target Compounds for the Different Fingerprint Types in the Case of the Protease Class

fingerprint	mean	SD	Tanimoto index statistical cutoff ^a	median	99.5% quantiles
AlFi	0.1562	0.0490	0.3038	0.1527	0.3113
Daylight	0.2492	0.0650	0.4445	0.2480	0.4568
Unity	0.2646	0.0790	0.5019	0.2603	0.5032
Hologram	0.3332	0.1240	0.7052	0.3357	0.6363
ALOGP	0.4988	0.2144	1.1420	0.5124	0.8966
ChemGPS	0.1811	0.2857	1.0390	0.1407	0.8734
CATS	0.2822	0.1318	0.6776	0.2762	0.6250
TRUST	0.1811	0.1442	0.6137	0.1509	0.6272
Molprint2D	0.0338	0.0365	0.1433	0.0256	0.1852

^a Tanimoto index statistical cutoff = mean + 3 × SD.

substructure based fingerprints (Daylight and Unity) this pair scores just outside our range (Table 4). In this case, pharmacophore-based fingerprints such as CATS and TRUST fail since the associated Tanimoto indices between the two compounds (0.60 and 0.59, respectively) are far from the “consensus” cutoff values (0.85 and 0.75, respectively). However, the inverse situation is observed between the sibutramine and the sertraline compounds; the TRUST based Tanimoto index, 0.72, is closed to the “consensus” value 0.75.

Statistical Validation of the Cutoff Values. The Tanimoto cutoffs for the different fingerprints show that they differ in their “scales” of biological relevance. We reason that this scale is related to the distribution of similarities of “random” (unrelated druglike) compounds, and the cutoff is related to the probability of a given similarity between two active compounds being above the “noise” of random similarities. In Figure 4 the distribution of 10 000 Tanimoto indices chosen randomly for some fingerprints of the protease class is presented. Statistical data for all fingerprints are summarized in Table 5. Thus, we can compare the empirical cutoffs defined above with the properties of these a priori distributions. In other words, assuming normal distributions one can estimate the probability that a given Tanimoto index is associated with the “noise”, i.e., inactive compounds. Taking three standard deviations above the mean as a cutoff

Table 6. Percentage of Overlap in the Top 5% of the Selection Made by the Different Fingerprints, in the Different Classes of Compounds

GPCR	AlFi									
Daylight	29	Daylight								
Unity	28	60	Unity							
Hologram	31	35	35	Hologram						
ALOGP	10	13	14	16	ALOGP					
ChemGPS	14	12	14	15	18	ChemGPS				
CATS	14	13	12	14	7	8	CATS			
TRUST	15	12	12	12	7	9	16	TRUST		
Molprint 2D	34	45	45	36	17	12	14	13	Molprint 2D	
ion channel	AlFi									
Daylight	44	Daylight								
Unity	41	64	Unity							
Hologram	45	43	40	Hologram						
ALOGP	14	14	13	18	ALOGP					
ChemGPS	18	14	16	18	16	ChemGPS				
CATS	26	24	22	22	10	11	CATS			
TRUST	23	22	21	20	10	12	24	TRUST		
Molprint 2D	51	52	51	45	19	15	26	24	Molprint 2D	
kinase	AlFi									
Daylight	48	Daylight								
Unity	47	67	Unity							
Hologram	50	48	48	Hologram						
ALOGP	15	16	17	19	ALOGP					
ChemGPS	20	16	17	20	16	ChemGPS				
CATS	29	26	25	28	11	14	CATS			
TRUST	24	20	20	20	10	12	22	TRUST		
Molprint 2D	56	57	57	51	19	17	28	23	Molprint 2D	
protease	AlFi									
Daylight	46	Daylight								
Unity	43	69	Unity							
Hologram	41	43	41	Hologram						
ALOGP	11	13	13	18	ALOGP					
ChemGPS	16	14	16	18	15	ChemGPS				
CATS	22	21	20	22	9	11	CATS			
TRUST	26	23	23	22	11	13	21	TRUST		
Molprint 2D	52	57	55	47	17	16	23	27	Molprint 2D	

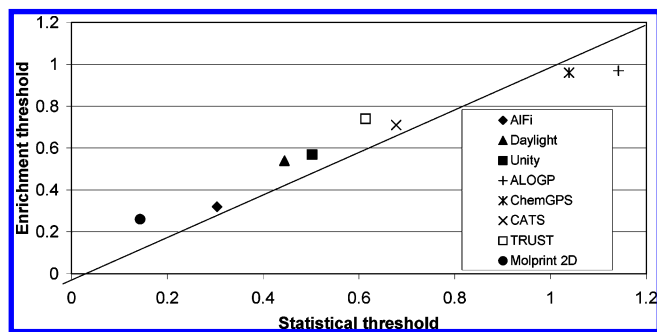


Figure 5. Correlation between the threshold values of the Tanimoto index for the different fingerprints derived from the enrichment study with a cutoff based on mean + three standard deviations for the protease class compounds.

for relevant similarity (in case of normal distributions this corresponds to 0.1% probability of random similarity) we can relate the empirical cutoffs to the a priori distribution of similarities for a given fingerprint (Table 5). A good correlation between the cutoff values derived from the two methods can be observed and gives a statistical explanation of the similarity cutoffs established in the enrichment study (Figure 5). Note that some fingerprint distributions are distinctly non-normal (e.g. ChemGPS, and ALOGP fingerprint), and this leads to the clearly unreasonable result of statistical cutoffs larger than unity. However a more meaningful parameter for non-normal distributions is the 99.5% quantile which correlates well with our empirical cutoffs (Tables 4 and 5). The results from this study clearly explain the different “scales” seen for the different fingerprints in

the first section; while there is a low probability that two unrelated compounds will have a Tanimoto similarity of 0.5 with AIFi fingerprints, this is a very likely event with ChemGPS or ALOGP. Note again that this analysis says nothing about how efficient the different fingerprints are for compound selection, rather it provides a guideline for relevant Tanimoto similarities.

Overlap Study: “One to One” Fingerprint Comparison. When selecting compounds using different methods, a certain amount of redundancy may appear in the selections. To investigate this we analyzed the overlap in the top 5% (500 compounds) of the selections from each fingerprint (Table 6, values in bold refer to overlap higher than 40%). In all compound classes Daylight and Unity selections are highly redundant. This is not surprising given that the fingerprints are highly similar in nature and in length. In fact, relatively high overlap ($\sim 30\%$ or more) is seen for all substructure-related fingerprints (AlFi, Daylight, Unity, and Hologram). The Molprint 2D is also overlapping with the substructural fingerprints. Less overlap is seen for selections from the other fingerprints (CATS, TRUST, ChemGPS, and ALOGP); we can expect a high complementarity of actives between these selections. This is seen in Table 7, showing the overlap for active compounds found in the top 5% selected sets. We notice that the overlaps seen for *actives* are similar to those observed for the *selections* in Table 6. For CATS and TRUST particularly, proportionally higher overlaps for active compounds are seen versus the substructure based fingerprints (the overlap ranges from 24% to 29% for GPCR, from 32% to 51% for the other classes). Smaller

Table 7. Percentage of Overlap between the Active Found in the Top 5% of the Selection Made by the Different Fingerprints, in the Different Classes of Compounds

GPCR	AlFi	Daylight	Unity	Hologram	ALOGP	ChemGPS	CATS	TRUST	Molprint 2D
AlFi		35	35	39	11	21	19	19	41
Daylight	39		69	46	18	20	19	19	54
Unity	38	67		46	19	22	18	19	53
Hologram	38	41	42		18	21	19	17	44
ALOGP	13	17	19	21		26	7	9	20
ChemGPS	21	18	20	22	23		10	11	18
CATS	29	27	26	31	11	16		25	29
TRUST	27	24	24	26	11	16	22		26
Molprint 2D	45	54	55	50	20	20	20	20	

ion channel	AlFi	Daylight	Unity	Hologram	ALOGP	ChemGPS	CATS	TRUST	Molprint 2D
AlFi		52	51	55	19	22	32	29	59
Daylight	62		74	55	22	24	35	33	67
Unity	62	74		55	22	26	34	32	66
Hologram	57	46	47		24	24	29	24	53
ALOGP	27	26	25	32		25	17	19	31
ChemGPS	25	21	24	26	19		16	16	23
CATS	42	37	36	35	16	19		32	39
TRUST	40	37	36	32	18	21	34		40
Molprint 2D	67	62	62	58	25	24	34	33	

kinase	AlFi	Daylight	Unity	Hologram	ALOGP	ChemGPS	CATS	TRUST	Molprint 2D
AlFi		53	52	60	20	26	37	29	64
Daylight	65		75	64	21	24	38	29	71
Unity	63	75		62	23	26	37	29	68
Hologram	68	59	57		23	26	41	28	65
ALOGP	37	31	34	37		26	24	20	31
ChemGPS	37	27	30	32	20		23	17	33
CATS	51	43	43	51	18	23		31	46
TRUST	50	41	41	42	19	21	38		51
Molprint 2D	66	67	66	55	15	22	30	36	

protease	AlFi	Daylight	Unity	Hologram	ALOGP	ChemGPS	CATS	TRUST	Molprint 2D
AlFi		57	53	49	13	21	28	33	64
Daylight	61		80	54	13	20	30	34	73
Unity	56	79		50	12	20	28	33	71
Hologram	59	61	58		18	26	31	34	66
ALOGP	26	25	24	31		27	19	22	38
ChemGPS	32	28	29	32	19		19	22	31
CATS	44	44	42	40	14	19		37	48
TRUST	47	46	45	41	15	21	35		48
Molprint 2D	70	65	64	64	22	24	38	30	

overlaps are seen for selections with ALOGP or ChemGPS fingerprints. Reassuringly there is in general more redundancy in the actives compared to the full selections. However this stresses that while adding a new fingerprint for similarity searches does recover novel actives it can also have a detrimental effect on the overall performance of the similarity-based selection (since more “noise” will be selected).

Cluster Coverage. In a similarity search procedure, it is interesting to know how a set of selected target compounds covers the whole ensemble of target clusters. Hence, we investigate to what extent the different fingerprint-based selections “cover” the database clusters (see methods for description of clustering procedure). Table 8 gives the average (based on the 75 query-target combinations) number of clusters present in the target sets as well as the number of clusters containing at least one active compound for the different protein classes.

Overlap: One to the Ensemble Fingerprint Selection. The comparison between the top 5% selected compound sets shows the degree of redundancy between different fingerprint selections. However, even if a selection made by one fingerprint presents low overlaps with any of the other

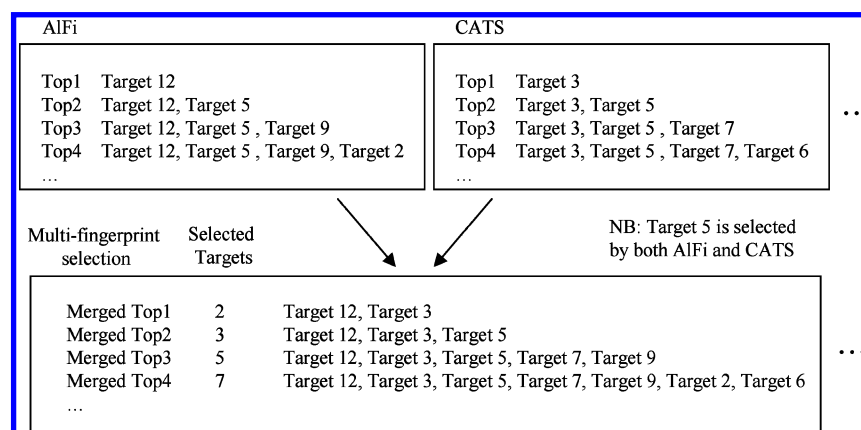
Table 8. Average Number of Clusters in the Target Sets and Clusters Containing at Least One Active Compound for the Different Classes of Protein

	number of clusters	number of active clusters
GPCR	8725	485
ion channel	6702	473
kinase	8177	480
protease	7954	465

fingerprint selections, it can happen that its overall contribution to the selection is poor. This occurs if the pairwise overlaps are small, but all the compounds found with one fingerprint are already present in the selection by the ensemble of fingerprints. Inversely, the high overlap between the selection made by Daylight and Unity, for instance, does not necessarily mean a poor contribution of both methods to the overall similarity based selection. In this section, we investigate the complementarity between one fingerprint and the others taken as an ensemble. Thus, compounds belonging to the selection made by one fingerprint can be also chosen by one or several other methods, while others molecules are exclusively selected by a given fingerprint. In addition, in a

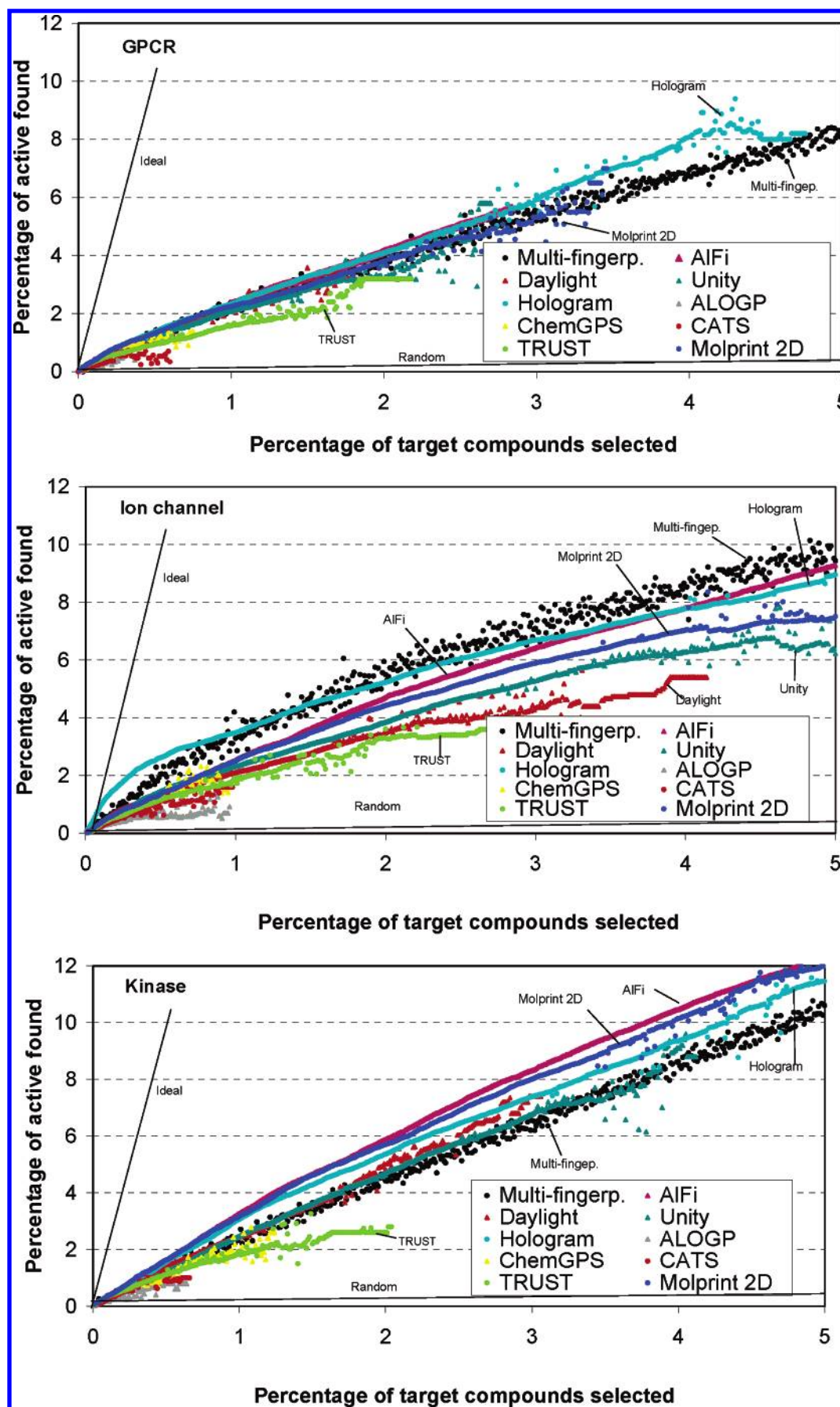
Table 9. Comparison of the Numbers and the Percentage of Exclusivity of Selected Compounds, Active Selected, Cluster Selected by the Selected Compounds, and the Cluster Explored by the Active Selected for the Different Fingerprints, for the Different Classes of Molecules

fingerprint	number					percentage of exclusivity				
	Tanimoto threshold	enrich.	selected target	active target	cluster selected	cluster active	selected target	active target	cluster selected	cluster active
GPCR										
AlFi	0.30	1.73	681	59	572	57	64	60	65	61
Daylight	0.65	2.11	123	13	89	12	11	6	11	6
Unity	0.65	1.85	195	18	149	16	23	18	24	19
Hologram	0.70	1.96	327	32	279	31	43	37	44	38
ALOGP	0.97	1.82	22	2	21	2	50	54	50	55
ChemGPS	0.95	2.00	50	5	48	5	60	52	59	52
CATS	0.85	1.38	29	2	26	2	49	19	49	19
TRUST	0.75	1.57	127	10	117	10	61	39	62	39
Molprint 2D	0.30	1.82	264	23	221	22	29	17	30	18
Ion Channel										
AlFi	0.30	1.66	959	75	570	68	50	39	53	41
Daylight	0.65	1.78	259	23	121	20	7	4	7	4
Unity	0.65	1.61	422	34	201	29	22	9	19	10
Hologram	0.70	1.69	591	50	353	45	34	27	39	27
ALOGP	0.97	1.86	43	4	33	3	36	19	33	19
ChemGPS	0.95	2.46	65	8	55	8	45	34	45	37
CATS	0.85	2.04	49	5	35	5	25	18	28	16
TRUST	0.75	1.66	145	12	106	11	39	19	45	20
Molprint 2D	0.30	1.61	561	45	332	40	25	14	28	15
Kinase										
AlFi	0.30	2.01	904	91	694	89	50	37	52	38
Daylight	0.65	2.36	195	23	125	20	6	2	7	2
Unity	0.65	2.22	306	34	209	30	14	8	15	9
Hologram	0.70	2.17	517	56	389	50	30	15	31	15
ALOGP	0.97	1.67	24	2	22	2	33	20	32	20
ChemGPS	0.95	1.94	72	7	64	7	52	29	51	29
CATS	0.85	2.16	37	4	32	4	25	5	27	5
TRUST	0.75	2.00	100	10	86	10	37	16	40	16
Molprint 2D	0.30	2.33	472	55	357	50	19	11	21	12
Protease										
AlFi	0.30	1.83	1127	103	811	91	60	48	60	48
Daylight	0.65	2.86	168	24	97	20	4	1	5	1
Unity	0.65	2.72	265	36	165	30	13	10	15	11
Hologram	0.70	1.93	486	47	362	41	34	18	37	19
ALOGP	0.97	2.22	18	2	17	2	33	10	31	10
ChemGPS	0.95	1.82	110	10	96	9	51	28	50	18
CATS	0.85	2.70	37	5	28	5	26	13	29	12
TRUST	0.75	2.04	167	17	129	15	33	15	38	16
Molprint 2D	0.30	2.46	426	51	305	45	18	9	21	10

**Figure 6.** Illustration of the multifingerprint selection procedure. At “Top 1” the multifingerprint will include all unique Top1 selected compounds coming from all fingerprints.

prospective project application there is neither a requirement nor an advantage to pick up the same number of compounds from the different fingerprints (for example, the top 5%) as these sets may present quite different enrichments. Instead, we decided to use the similarity threshold values established

above. For this reason, we show here the data from a selection made using the “consensus” value of Tanimoto thresholds. In Table 9, we present the total number of the selected actives and the number of clusters covered by the selected compounds and by the selected actives for each



fingerprint as well as the “percentage of exclusivity” (i.e. the percentage of compounds that are only selected by the given fingerprint). We observe that some enrichment factors

are below the arbitrary cutoff of 2 chosen for establishing the Tanimoto threshold values. This is explained by the fact that the consensus values derive from an average over the

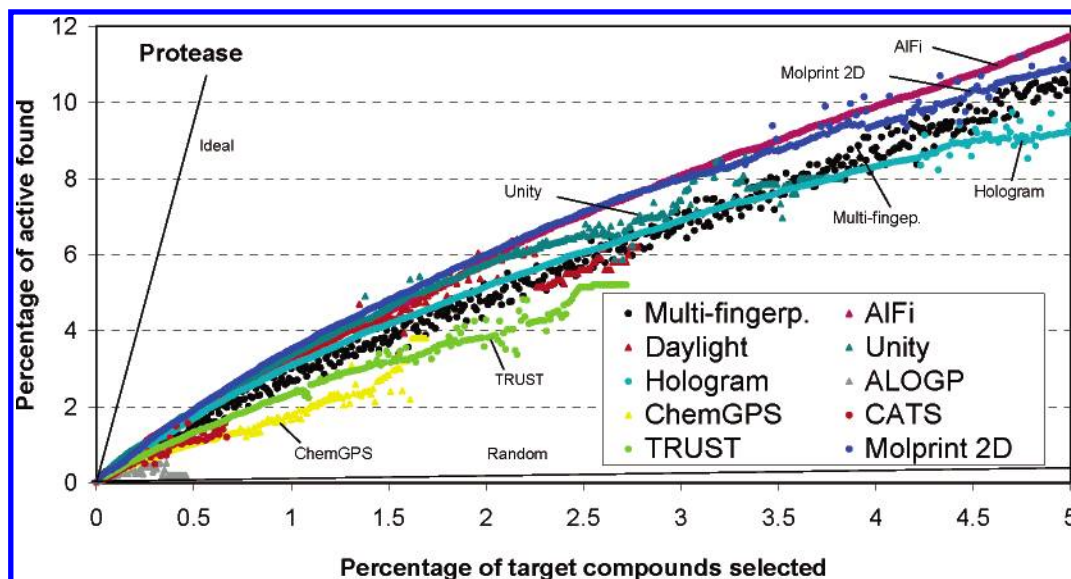


Figure 7. Number of active compounds as a function of the number of target compounds selected by various fingerprints and the multifingerprint method for different classes of proteins.

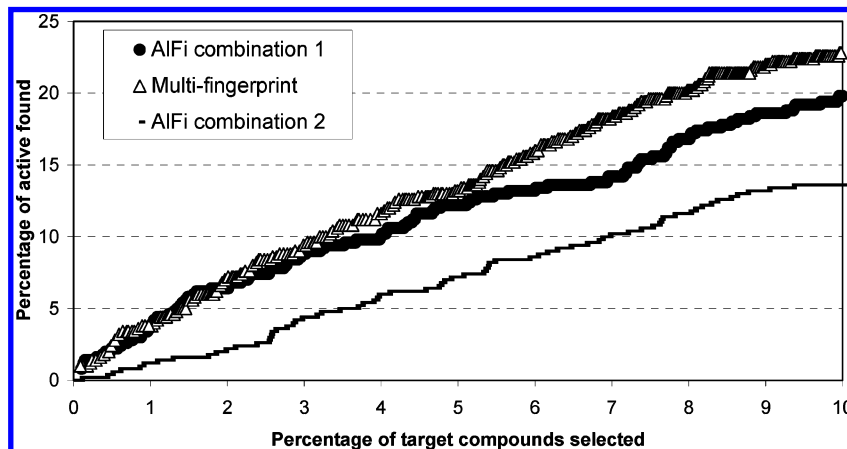


Figure 8. Number of active compounds with respect to the number of targets selected for the AIFi fingerprint type and the multifingerprint methods in the case of one query-target set combination within the protease class.

four target class sets. One could argue that this simulates a real case where the Tanimoto limit to have an enrichment factor of a certain value is not known.

The use of consensus values for the Tanimoto index prevents the selection of too many inactive molecules. Indeed, as an example, the enrichment factor for CATS selection within the GPCR class is already weak, 1.38; that would have been worse if we had “blindly” selected 500 molecules (i.e. the top 5%). Even though the number of selected compounds can vary for the different fingerprints, all sets present some novel molecules as none of the fingerprint has null values in the “exclusivity” part of Table 9. A consistently poor contribution is found for Daylight fingerprint (ranges from 1 to 6% unique active compounds in the selections). This is due to the high correlation of this fingerprint to both AIFi and Unity fingerprints. The AIFi fingerprint leads for any protein class to the highest ratio of exclusive actives (37–60% unique actives), and this enables exploring a larger biologically relevant chemical space. For other fingerprints exploring ~1000 target compounds would correspond to lower enrichment values and, consequently, to higher ratios of false positives. For example, selecting

~1100 compounds from the protease set (corresponding to a 0.30 cutoff for the AIFi fingerprint, see Table 9) by using Daylight, ALOGP, and CATS fingerprints would require Tanimoto cutoff's of 0.50, 0.92, and 0.65, respectively (see Figure 3). Figure 2 shows that at these Tanimoto cutoff's CATS and ALOGP fingerprints present very poor enrichment values (1.27 and 1.05, respectively), while for the Daylight fingerprint the enrichment value is still reasonable, 1.66.

Thus, to a first approximation there is a clear advantage to use all fingerprints when searching for active compounds in chemical collections as all fingerprints bring some novel actives. On the other hand, no significant differences were observed between the coverage and diversity of different selections as the *number of selected targets/cluster selected* and the *number of active targets/cluster active* ratios, see Table 9, are rather constant.

Multiple versus Single Fingerprint Method. In this section we compare the multifingerprint selection (all fingerprints using their corresponding consensus Tanimoto threshold values) and the selection made by using one fingerprint at a time. A multifingerprint selection is generated simply by merging the compounds coming from the ranked

list of selected targets corresponding to all fingerprints as exemplified in Figure 6. One can produce “ n ” multifingerprint selections containing “ m ” compounds where $m = n \times (\text{number of fingerprint}) - r$ (redundancy). Once again, this procedure is applied to the 75 combinations of the query-target set for each class, and in the following we present the average values of the results.

Figure 7 presents a comparison between the single fingerprint and the multifingerprint methods for selecting active compounds (all selected target compounds fulfill the consensus values of similarity threshold). We can readily see that the enrichment factor patterns are different for the four target classes. This finding is in agreement with the studies of similarity-based virtual screening using multiple bioactive reference structures for different activity classes by Hert et al.^{26,27} (although they used different fingerprints and different sets of compounds). For proteases and kinases, AIFi fingerprint retrieves a larger number of active compounds followed closely by Molprint 2D. For ion channels the multifingerprint approach has the best enrichment factors. For GPCR targets Hologram and multifingerprint present the same efficiency in retrieving active compounds.

It is important to stress that, in the present study, all query sets contain 50 randomly selected active compounds from the total number of actives in each target class and that the results presented in Figure 7 are averages over 75 trials. This explains the rather small differences between different fingerprints. In any two particular runs the fingerprints ranking on the basis of their enrichment factors might be different, see the examples in Figure 8, and this is mainly due to the differences in the diversity of randomly selected queries that will directly affect the optimum performance of each fingerprint. This behavior is smoothed by the averaging procedure. When translating this into practice we may find that for one particular application one fingerprint gives better results than another in selecting an active compound. However this depends on the project and the query compounds rather than the fingerprint, and there is no clear way to decide a priori which of the fingerprints will perform better. However using the multifingerprint selection one can explore better the chemical space described by the query set and have a better chance to get higher enrichment factors. It is also more robust since for a particular screening situation the weakness of one fingerprint can be balanced by the strengths of other fingerprints.

CONCLUSIONS

A detailed analysis was performed to assess the ability of different fingerprints in capturing biological similarity. We have established threshold values of similarity for a set of nine 2D fingerprints by analyzing enrichment factor data, and it was shown that the threshold values of similarity and statistical analysis of the Tanimoto index distribution are in relatively good agreement. These cutoff values can be used as guidelines when searching for new active compounds by similarity to a set of known active molecules. A large overlap between the selections made by using different fingerprints can be observed. However, the fingerprints rank the molecules rather differently, and therefore their combined output gives a better alternative in terms of enrichment factors and diversity of the final selection.

NOTE

After finishing this paper we were notified of a similar work by Godden et al. (ref 28, JCIM articles ASAP). The authors used five biological activity data from ACD and MDDR and two fingerprints (MACCS and MPMFP) to benchmark fingerprint search calculations and estimate their probability of success.

ACKNOWLEDGMENT

The authors are grateful to Andreas Bender from the Unilever Centre for Molecular Science Informatics (Department of Chemistry, University of Cambridge) for the codes to perform Molprint 2D fingerprint calculations. We also thank our colleagues at AstraZeneca: Hongming Chen (Mölnådal, Sweden) for providing Tanimoto similarity index code, David Cosgrove (Alderley Park, U.K.) who created and implemented the AIFi fingerprint as well as the FLUSH program, and Jens Sadowski (Mölnådal, Sweden) for the ALOGP code and general support.

REFERENCES AND NOTES

- (1) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
- (2) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (3) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of “Molecular Diversity” Descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
- (4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (5) 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.
- (6) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911.
- (7) Bajorah, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.
- (8) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
- (9) Xue, L.; Godden, J. W.; Bajorah, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227–1234.
- (10) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorah, 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.
- (11) Daylight Chemical Information Systems Inc. <http://www.daylight.com>.
- (12) UNITY; Tripos Inc. <http://www.tripos.com>.
- (13) Rabow, A.; Cosgrove, D. Manuscript in preparation.
- (14) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc., Jun. 2003. <http://www.daylight.com/release/manuals.html>.
- (15) The Hologram fingerprint is generated with Sybyl (HQSAR module required) available from Tripos Inc. <http://www.tripos.com>; the parameters used in this study are as follows: length = 257, minimum number of atoms in a fragment = 4, maximum of atoms in a fragment = 7, atom flag = 1, bond flag = 1, connections flag = 1, hydrogen's flag = 0, chirality's flag = 0.
- (16) 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. A* **1998**, *102*, 3762–3772.
- (17) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157–166.
- (18) Schneider, G.; Neidhard, W.; Giller, T.; Schmid, G. “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.

- (19) Engkvist, O.; Kogej, T.; Blomberg, N.; Muresan, S. Manuscript in preparation.
- (20) Jacoby, E.; Schuffenhauer, A.; Popov, M.; Azzaoui, K.; Havill, B.; Schopfer, U.; Engeloeh, C.; Stanek, J.; Acklin, P.; Rigollier, P.; Stoll, F.; Koch, G.; Meier, P.; Orain, D.; Giger, R.; Hinrichs, J.; Malagu, K.; Zimmermann, J.; Roth, H.-J. Key Aspects of the Novartis Compound Collection Enhancement Project for the Compilation of a Comprehensive Chemogenomics Drug Discovery Screening Collection. *Curr. Top. Med. Chem.* **2005**, *5*, 397–411.
- (21) <http://openbabel.sourceforge.net>.
- (22) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reilling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
- (23) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reilling, S. Similarity searching of chemical databases using atom environment descriptors (MOL-PRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
- (24) FLUSH is an in-house program for cluster analysis and neighborhood mapping. It was used in combination to the Daylight fingerprint (Dave Cosgrove, AstraZeneca, Alderley Park).
- (25) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.
- (26) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org., Biomol. Chem.* **2004**, *2*, 3256–3266.
- (27) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- (28) Godden, J. W.; Stahura, F. L.; Bajorath, J. Anatomy of Fingerprint Search Calculations on Structurally Diverse Sets of Active Compounds. *J. Chem. Inf. Comput. Sci.* **2005**, in print.

CI0504723