

Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations

Ling Xue,[†] Florence L. Stahura,[†] Jeffrey W. Godden,[†] and Jürgen Bajorath^{*,†,‡}

New Chemical Entities, Inc., 18804 North Creek Parkway, Suite 100, Bothell, Washington 98011, and
Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received November 27, 2000

Results of systematic virtual screening calculations using a structural key-type fingerprint are reported for compounds belonging to 14 activity classes added to randomly selected synthetic molecules. For each class, a fingerprint profile was calculated to monitor the relative occupancy of fingerprint bit positions. Consensus bit patterns were determined consisting of all bits that were always set on in compounds belonging to a specific activity class. In virtual screening calculations, scale factors were applied to each consensus bit position in fingerprints of query molecules. This technique, called “fingerprint scaling”, effectively increases the weight of consensus bit positions in fingerprint comparisons. Although overall prediction accuracy was satisfactory using unscaled calculations, scaling significantly increased the number of correct predictions but only slightly increased the rate of false positives. These observations suggest that fingerprint scaling is an attractive approach to increase the probability of identifying molecules with similar activity by virtual screening. It requires the availability of a series of related compounds and can be easily applied to any keyed fingerprint representation that associates bit positions with specific molecular features.

INTRODUCTION

Computational screening of databases for molecules with a specific activity has become a popular approach in drug design and discovery.¹ A variety of methods for virtual screening of compound databases has been introduced including approaches based on macromolecular target structures,^{2,3} often called docking, and techniques that start from small molecular hits or leads.^{1,4} Small molecular methods of increasing complexity and sophistication include relatively simple 2D substructure search techniques,⁵ more abstract 2D molecular fingerprint representations,⁶ static or flexible 3D pharmacophore search methods,^{7,8} 3D- or 4D-QSAR models,^{9,10} and 3D pharmacophore fingerprints.^{11,12} In addition, different statistical methods have been developed to derive predictive models of biological activity that can then be used for virtual screening of compound databases.^{13,14}

Binary molecular fingerprints that encode various property descriptors, structural fragments, molecular patterns, or pharmacophores are popular computational tools for virtual screening based on small molecular hits or leads. In such calculations, fingerprints are usually generated for one or more query molecules and searched against corresponding fingerprints of database compounds. As a measure of molecular similarity, fingerprint overlap is then calculated for pairwise comparison of molecules using different metrics.¹⁵ Calculation of fingerprints for thousands of database compounds is usually computationally expensive. However, the translation of compound databases into “fingerprint space” is a one-time calculation per molecule and thus provides the basis for many virtual screens.

Different types of fingerprint representations have been introduced.⁴ Such fingerprints typically encode molecular information in the form of binary bit strings. Each bit may account for the presence or absence of a specific molecular feature, for example, a structural fragment¹⁶ or a potential pharmacophore,¹² or may represent a value or range of a property descriptor. This type of design is often called “keyed” because each bit position is associated with a specific feature, regardless of the molecule for which the fingerprint is calculated. Alternatively, fingerprint designs may utilize “hashing” or folding where molecular patterns are mapped to overlapping bit positions to produce characteristic molecular bit patterns.⁶ Consequently, in these fingerprints, a single bit position can no longer be associated with only one specific property.

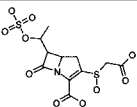
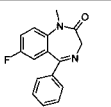
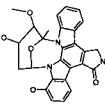
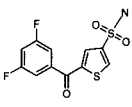
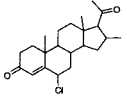
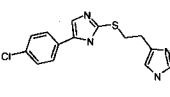
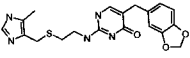
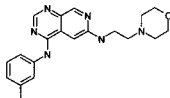
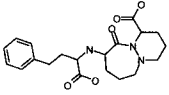
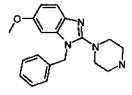
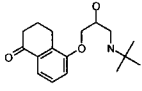
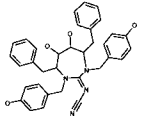
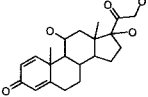
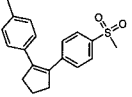
The current study had several objectives. We intended to carry out a systematic fingerprint-based virtual screening experiment, involving more than 200 compounds belonging to 14 different classes of drug or drug-like molecules, to determine overall performance when searching for molecules having very diverse activities. Such insights are still limited in the study of virtual screening approaches. Furthermore, we have developed and applied a new technique called “fingerprint scaling” that emphasizes fingerprint bit patterns conserved in specific compound classes and, in test calculations, determined whether the predictive performance of virtual screening calculations could be improved. Thus, we have assembled a benchmark compound database for systematic virtual screening. For our calculations, a structural fragment or key-type fingerprint, consisting of MACCS keys,^{16,17} was used, for several reasons. This fingerprint is widely available, its keyed design provides an excellent basis for the application of our scaling technique, and its limited complexity makes it simple to visualize consensus bit

* Corresponding author phone: (425)424-7297; fax: (425)424-7299;
e-mail: jbajorath@nce-mail.com.

[†] New Chemical Entities, Inc.

[‡] University of Washington.

Table 1. Classes of Active Compounds in the Test Database^a

Class	Number of compounds	Activity	Representative structure	Class	Number of compounds	Activity	Representative structure
NP_BLC	14	β -Lactamase inhibitors		BA_BEN	22	Benzodiazepine receptor ligands	
NP_PKC	15	Protein kinase C inhibitors		BA_CAE	22	Carbonic anhydrase II inhibitors	
CMC_ESTR	11	Estrogen antagonists		BA_H3E	21	H3-antagonists	
CMC_H2	12	Histamine H2-antagonists		BA_TKE	20	Tyrosine kinase inhibitors	
CMC_ACE	17	ACE inhibitors		BA_5HT	21	Serotonin receptor (5-HT) ligands	
CMC_ADR	16	Anti-adrenergic (β -receptor) ligands		BA_HIV	18	HIV protease inhibitors	
CMC_GLU	14	Glucocorticoid analogues		BA_COX	17	Cyclooxygenase-2 (Cox-2) inhibitors	

^a Abbreviations for each of the 14 classes of biologically active or drug-like compounds are given under "class" and are used throughout the text. For each class, an example structure is shown.

patterns. Furthermore, structural keys have been shown, in independent studies, to be powerful 2D descriptors to, for example, cluster or partition molecules according to biological activity.^{18,19} In our virtual screening calculations, MACCS keys gave satisfactory results. However, fingerprint scaling significantly improved the performance for the majority of compound classes and increased overall prediction accuracy.

METHODS

Our test database consisted of 2240 compounds including 240 compounds belonging to 14 different activity classes. Active molecules were obtained by combining two compound collections that were independently assembled from the literature and specific databases as described previously.^{20,21} The activity classes and number of compounds per class are reported in Table 1. The database contains a diverse array of biologically active compounds. Seven of the classes were taken from the literature,²⁰ five from the Comprehensive Medicinal Chemistry Database,²² and two from the Chapman and Hall Dictionary of Natural Products.²³ Thus, the latter two classes exclusively consisted of naturally occurring molecules. The 240 molecules were then added to 2000 randomly selected synthetic compounds from ACD²⁴ that were considered "inactive" in our studies. These ACD compounds were added to increase the "noise" in virtual screening calculations, i.e., the probability of identifying false

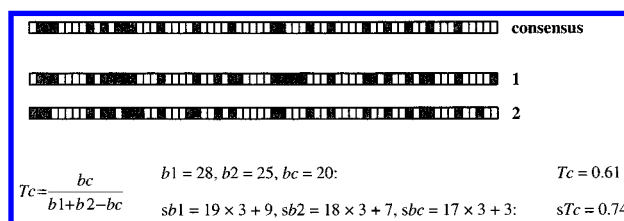


Figure 1. Fingerprint scaling. At the top, the consensus pattern of a hypothetical drug class and a model fingerprint is shown. This fingerprint is generated for a molecule A belonging to this class and a similar molecule B. At the bottom, the conventional Tc value is calculated for comparison of the two fingerprints and a scaled Tc value (sTc) using a scale factor of 3.

positives. The compound database was imported into the Molecular Operating Environment (MOE)²⁵ that served as the computational platform for all calculations.

As a fingerprint, we used the set of 166 publicly available MACCS keys (MACCS-FP), as mentioned before. Each of the 166 bit positions in this keyed fingerprint accounts for the presence or absence of a specific structural motif in a molecule. As a similarity metric, fingerprint overlap was calculated using the conventional Tanimoto coefficient,¹⁵ defined as $Tc = bc / (b1 + b2 - bc)$, with $b1$ being the number of bits set on in the first molecule, $b2$ the number of bits set on in the second molecule, and bc the number of bits that both molecules have in common.

Table 2. Search Results for a Compound Belonging to Class “CMC_ADR”^a

Tc	scale factor = 1 (unscaled)					scale factor = 2				
	score	A	Na	B	Nb	score	A	Na	B	Nb
0	0.00	16	16	2224	2224	0.00	16	16	2224	2224
0.05	0.00	16	16	2222	2224	0.00	16	16	2222	2224
0.1	0.00	16	16	2210	2224	0.00	16	16	2215	2224
0.15	0.00	16	16	2191	2224	0.00	16	16	2200	2224
0.2	0.00	16	16	2098	2224	0.00	16	16	2183	2224
0.25	0.00	16	16	1895	2224	0.00	16	16	2110	2224
0.3	0.00	16	16	1524	2224	0.00	16	16	1945	2224
0.35	0.00	16	16	1158	2224	0.00	16	16	1694	2224
0.4	0.00	16	16	830	2224	0.00	16	16	1349	2224
0.45	0.00	16	16	532	2224	0.00	16	16	981	2224
0.5	0.00	16	16	283	2224	0.00	16	16	662	2224
0.55	0.01	16	16	116	2224	0.00	16	16	367	2224
0.6	0.03	16	16	34	2224	0.01	16	16	160	2224
0.65	0.16	15	16	6	2224	0.02	16	16	59	2224
0.7	0.63	10	16	0	2224	0.06	16	16	17	2224
0.75	0.50	8	16	0	2224	1	16	16	1	2224
0.8	0.44	7	16	0	2224	0.63	10	16	0	2224
0.85	0.38	6	16	0	2224	0.44	7	16	0	2224
0.9	0.06	1	16	0	2224	0.38	6	16	0	2224
0.95	0.06	1	16	0	2224	0.06	1	16	0	2224
1	0.06	1	16	0	2224	0.06	1	16	0	2224

Tc	scale factor = 3					scale factor = 4				
	score	A	Na	B	Nb	score	A	Na	B	Nb
0	0.00	16	16	2224	2224	0.00	16	16	2224	2224
0.05	0.00	16	16	2222	2224	0.00	16	16	2222	2224
0.1	0.00	16	16	2218	2224	0.00	16	16	2218	2224
0.15	0.00	16	16	2201	2224	0.00	16	16	2201	2224
0.2	0.00	16	16	2189	2224	0.00	16	16	2189	2224
0.25	0.00	16	16	2144	2224	0.00	16	16	2153	2224
0.3	0.00	16	16	2048	2224	0.00	16	16	2092	2224
0.35	0.00	16	16	1886	2224	0.00	16	16	1958	2224
0.4	0.00	16	16	1630	2224	0.00	16	16	1766	2224
0.45	0.00	16	16	1271	2224	0.00	16	16	1453	2224
0.5	0.00	16	16	931	2224	0.00	16	16	1123	2224
0.55	0.00	16	16	582	2224	0.00	16	16	745	2224
0.6	0.00	16	16	327	2224	0.00	16	16	472	2224
0.65	0.01	16	16	142	2224	0.00	16	16	234	2224
0.7	0.02	16	16	51	2224	0.01	16	16	95	2224
0.75	0.06	16	16	16	2224	0.03	16	16	33	2224
0.8	1	16	16	0	2224	0.14	16	16	7	2224
0.85	0.63	10	16	0	2224	0.94	15	16	0	2224
0.9	0.44	7	16	0	2224	0.50	8	16	0	2224
0.95	0.06	1	16	0	2224	0.19	3	16	0	2224
1	0.06	1	16	0	2224	0.06	1	16	0	2224

^a Search calculations are reported for one of the compounds belonging to activity class CMC_ADR. “Tc” gives the similarity threshold value for each calculation. “A” is the number of correctly identified compounds, “Na” the total number of compounds in this class, “B” the number of false positives, and “Nb” the total number of “inactive” compounds in the database. Nb varies dependent on the number of compounds per activity class. Results of unscaled calculations and scaled calculations using three different scale factors are shown.

In our calculations, each active compound was removed from the database and searched against the remaining molecules under systematic variation of the Tc cutoff value for the detection of similar compounds. Since the number of compounds in each activity class ranged from 11 to 22, approximately 0.5–1% of database compounds were correct “hits” in each calculation, but the vast majority of compounds, if recognized, would be incorrect. The following scoring function was implemented to evaluate the search results: $S = A/(Na*B)$, where A is the number of correctly identified compounds belonging to the same activity class as the query, Na is the total number of compounds in this

Table 3. Search for a Compound Belonging to Class “BA_H3E”^a

Tc	scale factor = 1 (unscaled)					scale factor = 2				
	score	A	Na	B	Nb	score	A	Na	B	Nb
0	0.00	21	21	2219	2219	0.00	21	21	2219	2219
0.05	0.00	21	21	2209	2219	0.00	21	21	2206	2219
0.1	0.00	21	21	2190	2219	0.00	21	21	2193	2219
0.15	0.00	21	21	2153	2219	0.00	21	21	2163	2219
0.2	0.00	21	21	2041	2219	0.00	21	21	2105	2219
0.25	0.00	21	21	1772	2219	0.00	21	21	1915	2219
0.3	0.00	21	21	1409	2219	0.00	21	21	1687	2219
0.35	0.00	21	21	1029	2219	0.00	21	21	1369	2219
0.4	0.00	21	21	667	2219	0.00	21	21	1063	2219
0.45	0.00	21	21	419	2219	0.00	21	21	744	2219
0.5	0.00	21	21	254	2219	0.00	21	21	512	2219
0.55	0.01	16	21	120	2219	0.00	21	21	309	2219
0.6	0.02	16	21	42	2219	0.00	21	21	172	2219
0.65	0.04	13	21	14	2219	0.01	19	21	85	2219
0.7	0.10	6	21	3	2219	0.03	16	21	29	2219
0.75	0.07	3	21	2	2219	0.09	13	21	7	2219
0.8	0.14	3	21	1	2219	0.29	6	21	1	2219
0.85	0.14	3	21	0	2219	0.14	3	21	1	2219
0.9	0.14	3	21	0	2219	0.14	3	21	0	2219
0.95	0.14	3	21	0	2219	0.14	3	21	0	2219
1	0.05	1	21	0	2219	0.05	1	21	0	2219

Tc	scale factor = 3					scale factor = 4				
	score	A	Na	B	Nb	score	A	Na	B	Nb
0	0.00	21	21	2219	2219	0.00	21	21	2219	2219
0.05	0.00	21	21	2206	2219	0.00	21	21	2206	2219
0.1	0.00	21	21	2192	2219	0.00	21	21	2195	2219
0.15	0.00	21	21	2169	2219	0.00	21	21	2173	2219
0.2	0.00	21	21	2122	2219	0.00	21	21	2122	2219
0.25	0.00	21	21	1970	2219	0.00	21	21	2007	2219
0.3	0.00	21	21	1786	2219	0.00	21	21	1828	2219
0.35	0.00	21	21	1516	2219	0.00	21	21	1602	2219
0.4	0.00	21	21	1225	2219	0.00	21	21	1315	2219
0.45	0.00	21	21	942	2219	0.00	21	21	1036	2219
0.5	0.00	21	21	693	2219	0.00	21	21	800	2219
0.55	0.00	21	21	453	2219	0.00	21	21	560	2219
0.6	0.00	21	21	292	2219	0.00	21	21	373	2219
0.65	0.01	21	21	159	2219	0.00	21	21	230	2219
0.7	0.01	21	21	75	2219	0.01	21	21	125	2219
0.75	0.03	18	21	32	2219	0.02	21	21	60	2219
0.8	0.15	16	21	5	2219	0.04	18	21	22	2219
0.85	0.29	6	21	1	2219	0.62	13	21	1	2219
0.9	0.14	3	21	0	2219	0.14	3	21	0	2219
0.95	0.14	3	21	0	2219	0.14	3	21	0	2219
1	0.05	1	21	0	2219	0.04	1	21	0	2219

^a Search results are shown for one of the H3-antagonists in the test database. The representation is according to Table 2.

class, and B is the number of false positive recognitions. If no compound was incorrectly identified, B was set to one so that S was still defined. This scoring function was applied because it emphasizes the number of correctly identified compounds, penalizes incorrectly identified molecules (false positives), and normalizes the score relative to the number of compounds per activity class. Thus, in case of a perfect, or nearly perfect, experiment (i.e., all compounds belonging to the same class as the query were correctly identified and zero or one false positives were observed), S would be one. Following our nomenclature, Nb is the total number of “inactive” compounds for each search calculation, which is not part of the scoring function but reported in tables summarizing the results. “Prediction accuracy” was defined as $PA = A/Na$. All programs required for this analysis were generated using SVL²⁶ code and implemented in MOE.

For each of the 14 activity classes, fingerprint profiles were calculated as described previously²⁷ by summation of bit

Table 4. Distribution of Scores for Unscaled Search Calculations^a

class	Tc	sf	PA			score			A			Na	B			Nb
			av	min	max	av	min	max	av (%)	min	max		av	min	max	
BA_5HT	0.70–0.85	1	0.27	0.14	0.43	0.25	0.11	0.43	5.7 (27)	3	9	21	0.48	0	4	2219
BA_BEN	0.70–0.80	1	0.36	0.23	0.45	0.36	0.23	0.45	8 (36)	5	10	22	0.21	0	1	2218
BA_CAE	0.60–1.00	1	0.30	0.09	0.82	0.29	0.09	0.82	6.5 (30)	2	18	22	0.27	0	2	2218
BA_COX	0.65–0.90	1	0.32	0.18	0.59	0.30	0.18	0.59	5.4 (32)	3	10	17	0.48	0	3	2223
BA_H3E	0.65–1.00	1	0.30	0.10	0.57	0.29	0.10	0.57	6.2 (30)	2	12	21	0.42	0	4	2219
BA_HIV	0.60–1.00	1	0.35	0.06	0.78	0.34	0.06	0.78	6.3 (35)	1	14	18	0.40	0	2	2222
BA_TKE	0.65–0.85	1	0.27	0.10	0.50	0.27	0.10	0.50	5.4 (27)	2	10	20	0.23	0	1	2220
CMC_ACE	0.70–0.90	1	0.32	0.12	0.53	0.30	0.12	0.53	5.4 (32)	2	9	17	0.54	0	2	2223
CMC_ADR	0.60–0.80	1	0.74	0.13	0.94	0.74	0.13	0.94	11.8 (74)	2	15	16	0.50	0	1	2224
CMC ESTR	0.70–1.00	1	0.27	0.09	0.73	0.19	0.09	0.45	2.4 (22)	1	8	11	0.31	0	2	2229
CMC_GLU	0.60–1.00	1	0.25	0.07	0.64	0.25	0.07	0.64	3.5 (25)	1	9	14	0.41	0	1	2226
CMC_hH2	0.60–1.00	1	0.12	0.08	0.33	0.11	0.08	0.33	1.4 (12)	1	4	12	0.20	0	2	2228
NP_BLC	0.55–0.70	1	0.77	0.21	0.93	0.77	0.21	0.93	10.7 (76)	3	13	14	0.20	0	1	2226
NP_PKC	0.60–1.00	1	0.25	0.07	0.60	0.25	0.07	0.60	3.7 (25)	1	9	15	0.13	0	1	2225

^a For compounds in each activity class, the observed maximum (max), minimum (min), and average (av) scores, prediction accuracies (PA), and number of correctly (A) and incorrectly (B) identified compounds are reported. Scores and prediction accuracy were calculated as defined in the Methods section. Percentage values are provided in parentheses. “Tc” reports the interval of Tc cutoff values in which minimum and maximum scores were observed, and “sf” means scale factor.

settings over all compounds in a class followed by division by the total number of compounds in this class. Thus, fingerprint profiles, represented as histograms, monitor the relative occupancy (between 0 and 1) of each of the bit positions in the fingerprint (166 in this case). Consensus profiles for each class were determined by selecting bit positions that were consistently set on in all compounds belonging to a class. When using fingerprint scaling, similarity search calculations were repeated and scale factors between two and five were applied to all bit positions forming the consensus profile for calculation of Tc values.

What is the basic idea behind fingerprint scaling? The concept is summarized in Figure 1. A short model fingerprint is shown for two compounds 1 and 2 and a consensus pattern for a parent class of molecules with similar activity. For direct comparison of the 1 and 2 fingerprints a Tc value of 0.61 is calculated. If the calculation is repeated by applying a scale factor of 3 to all bits shared with the consensus pattern, the “scaled Tc” value increases to 0.74. Thus, compounds 1 and 2 are considered more similar on the basis of this modified calculation. In virtual screening calculations, a Tc similarity threshold value would need to be applied to, for example, identify compound 1 if 2 was used as a query. If the Tc threshold was set to 0.7, the similarity relationship would only have been detected by comparison of scaled but not conventional Tc values. All scripts required for fingerprint profiling and scaling were written in Perl.

RESULTS AND DISCUSSION

Systematic Virtual Screening Calculations. Each of the 240 active compounds was searched against the remainder of the database, consisting of more than 2200 compounds, under systematic variation of the Tc cutoff value for recognition of “similar” molecules. In each case, the number of correct recognitions and false positives were determined and scores were calculated. Representative search results for two compounds are shown in Tables 2 and 3. In the first analysis, discussed in this section, MACCS-FP was used without any modification for the calculation of fingerprint overlap. In subsequent analyses, as discussed below, fingerprint scaling was applied and different scale factors were

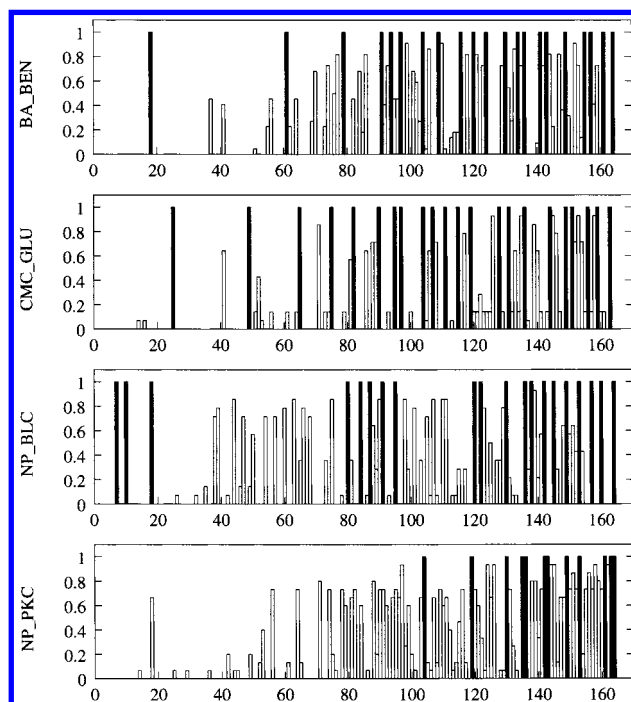


Figure 2. Fingerprint profiles. For four activity classes, profiles are shown that monitor the relative occupancy (vertical axis) of each of the 166 bit positions in MACCS-FP. Bars representing relative bit occupancy of one (i.e., bits set on in all compounds belonging to this class) are shown in black. The profiles display some significant differences.

used. First we wished to analyze the overall performance of (unscaled) virtual screening experiments for the diverse array of activity classes studied here. The results of our exhaustive search calculations using unscaled MACCS-FP are summarized in Table 4. As one may expect, the results show significant variations for different activity classes. However, all classes have in common that only very few false positives were recognized at optimum performance levels with MACCS-FP, i.e., between one and four of more than 2200 possible “inactive” compounds. In these calculations, poor predictive performance was characterized by identification of similarly low numbers of correct and incorrect compounds. For example, only 12% of H2-antagonists were correctly identi-

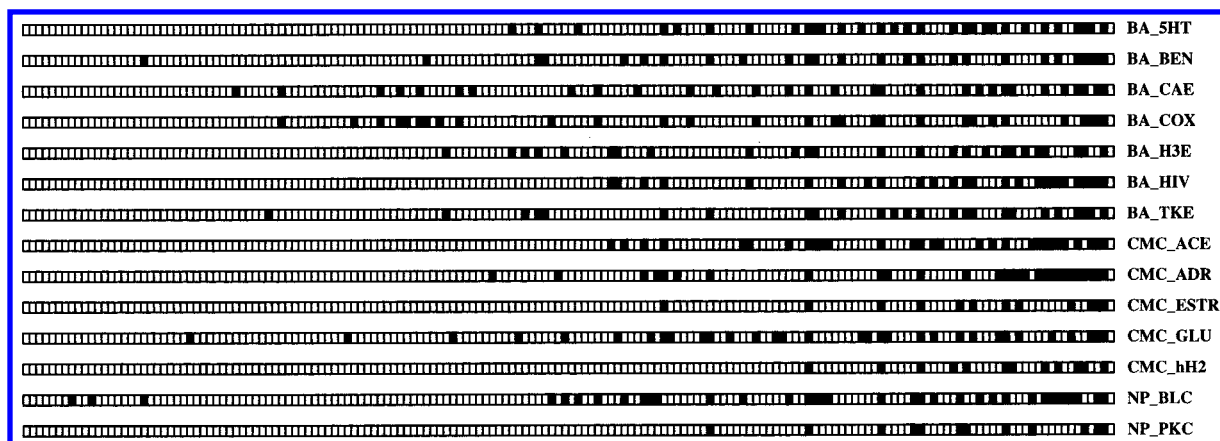


Figure 3. Consensus bit patterns. For all 14 activity classes, consensus bit patterns were generated consisting of all bits with relative bit occupancy of one. The comparison shows that the majority of activity classes shares some consensus bit positions, while many other consensus bits are highly variable among different classes or unique to a specific class.

fied. Two other classes, estrogen antagonists and glucocorticoid analogues, also showed only 22% and 25% correct recognition, respectively. Thus, MACCS-FP had obvious difficulties to identify these steroids. By contrast, compounds belonging to other classes were identified with higher accuracy. For example, more than 70% of β -lactamase inhibitors, one of two natural product classes, and anti-adrenergic ligands were correctly identified. Correct identifications in the range of 30% of active compounds, i.e., five or six compounds per class, were frequently observed. For the best Tc cutoff values, we observed on average six correctly identified compounds (of on average 17 per class) and only less than one identified compound with activity different from the one of the query molecule. Thus, taking into account that a rather diverse ensemble of activity classes was analyzed, the virtual screening results obtained with MACCS-FP provided satisfactory results.

Fingerprint Profiles and Consensus Patterns. For each of the 14 activity classes, a fingerprint profile was calculated that monitors the relative bit occupancy at each of the 166 positions of MACCS-FP. Four representative fingerprint profiles are shown in Figure 2. As can be seen, each activity class displayed a number of bit positions that were always set on in all compounds belonging to the class. These subsets of fully occupied bits were defined as “consensus patterns”. In each case, the consensus pattern contained less than 20% of the bits in MACCS-FP. Do consensus patterns of more or less randomly selected sets of active or drug-like molecules show “class-specific” features? Figure 3 compares the consensus bit patterns of all 14 activity classes studied here. The comparison shows that some bits are shared among activity classes, whereas other consensus positions are unevenly distributed or unique to specific activity classes. Thus, even for a relatively simple fingerprint such as MACCS-FP, in part overlapping yet distinct patterns of bit occupancy were observed in each case. The question of whether such differences can be exploited in order to further increase the performance level of virtual screening calculations is investigated below.

Scaling in Virtual Screening Calculations. The possibility to apply fingerprint scaling in virtual screening raises several questions. Does fingerprint scaling lead to a general increase in predictive performance or are effects limited to specific cases, if they occur at all? Do calculated consensus

bit patterns such as the ones shown in Figure 3 encode enough “specificity” to ensure meaningful applications of the method? Does fingerprint scaling lead to a significant increase in the number of false positives? To answer these questions and gain some insights into the potential of the approach, we repeated the systematic virtual screening analysis reported in Table 4 using fingerprint scaling. In these calculations, scale factors between two and five were subsequently applied to consensus bit positions of each activity class, as shown in Figure 3. As before, Tc cutoff values for detection of similarity were systematically varied and optimum performance ranges were determined. For a few compounds belonging to different activity classes, representative score profiles are shown in Figure 4. The profiles generally show distinct maxima at higher Tc values. As further discussed below, Tc threshold values that yield optimum performance usually differ in unscaled and scaled calculations and also depend on the applied scale factors. The results of all scaled calculations are summarized in Table 5. We found that for 12 of 14 classes average prediction accuracy and scores improved by fingerprint scaling. The only exceptions were estrogen and H₂-antagonists. Compared to our original calculations, the overall average of correctly recognized compounds per calculation increased from approximately six to nine. By contrast, the number of false positive recognitions remained essentially constant, thus resulting in an overall increase in predictive performance of more than 30%. Only in one case, glucocorticoid analogues, false positives increased in a somewhat significant way with, on average, one compound per calculation. The histogram in Figure 5 compares the best scores for all compounds in each class obtained in unscaled and scaled search calculations. The comparison reveals that scaling led to an overall improvement of scores and that the magnitude of improvement was dependent on the particular class of molecules. In part, these improvements were quite significant. For three well-performing classes (H₃-antagonists, benzodiazepine receptor ligands, and carbonic anhydrase inhibitors), the average number of correctly identified compounds doubled or nearly doubled. Best overall performance was observed for a class consisting of 16 anti-adrenergic ligands with an average of 15 correctly identified compounds and less than one false positive per calculation. As a control, we also repeated the search calculations without scaling and only

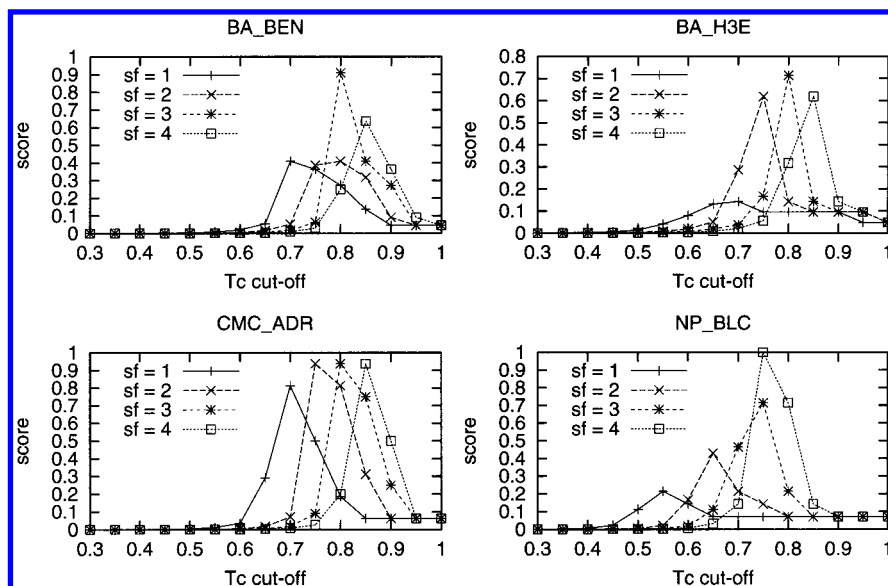


Figure 4. Score profiles. Profiles are shown for four compounds belonging to different activity classes (labeled as in Table 1), which display calculated scores (vertical axis) as a function of Tc cutoff values; “sf” means scale factor. For each compound, the profiles compare unscaled (sf = 1) and scaled calculations. These examples were selected because the profiles show some significant differences.

Table 5. Distribution of Scores under Conditions of Fingerprint Scaling^a

class	Tc	sf	PA			score			A			Na	B			Nb
			av	min	max	av	min	max	av (%)	min	max		av	min	max	
BA_5HT	0.80–0.95	2–5	0.30	0.24	0.52	0.28	0.16	0.43	6.2 (30)	5	11	21	0.40	0	3	2219
BA_BEN	0.75–0.90	2–5	0.60	0.32	0.95	0.58	0.32	0.95	13.3 (60)	7	21	22	0.41	0	3	2218
BA_CAE	0.70–0.90	2–5	0.50	0.18	1	0.50	0.18	1	11.1 (50)	4	22	22	0.48	0	2	2218
BA_COX	0.75–0.90	2–5	0.51	0.29	0.71	0.51	0.29	0.71	8.6 (51)	5	12	17	0.65	0	1	2223
BA_H3E	0.75–0.90	2–5	0.66	0.52	0.95	0.66	0.52	0.95	13.8 (66)	11	20	21	0.63	0	1	2219
BA_HIV	0.70–0.95	2–5	0.49	0.11	0.78	0.49	0.11	0.78	8.8 (49)	2	14	18	0.40	0	1	2222
BA_TKE	0.75–0.90	2–5	0.44	0.20	0.95	0.42	0.20	0.95	8.8 (44)	4	19	20	0.74	0	3	2220
CMC_ACE	0.80–0.95	2–5	0.44	0.18	0.71	0.40	0.18	0.53	7.5 (44)	3	12	17	0.65	0	2	2223
CMC_ADR	0.75–0.85	2–5	0.97	0.31	1	0.97	0.31	1	15.4 (96)	5	16	16	0.52	0	1	2224
CMC_ESTR	0.70–1.00	2–5	0.18	0.09	0.55	0.18	0.09	0.55	1.9 (17)	1	6	11	0.23	0	1	2229
CMC_GLU	0.80–0.95	2–5	0.47	0.14	0.86	0.40	0.10	0.71	6.5 (46)	2	12	14	1.18	0	9	2226
CMC_hH2	0.65–1.00	2–5	0.11	0.08	0.33	0.11	0.08	0.33	1.3 (11)	1	4	12	0.17	0	2	2228
NP_BLC	0.65–0.80	2–5	1	1	1	1	1	1	14 (100)	14	14	14	0.08	0	1	2226
NP_PKC	0.65–1.00	2–5	0.28	0.07	0.6	0.28	0.07	0.6	4.2 (28)	1	9	15	0.16	0	1	2225

^a Results of scaled search calculations are reported for scaling factors between two and five. The data representation is according to Table 4.

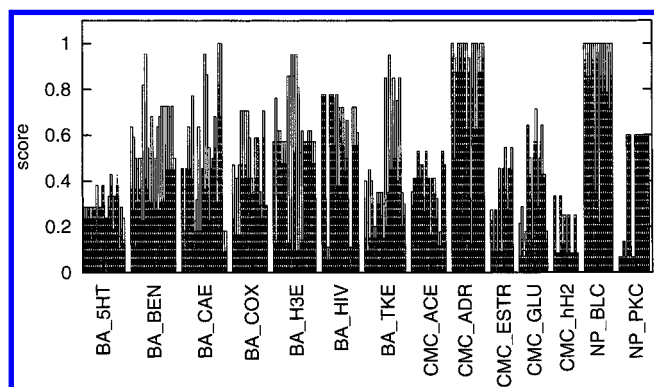


Figure 5. Best scores for each compound. The histogram records optimum scores (vertical axis) obtained for each compound in unscaled (filled) and scaled (unfilled) search calculations. Activity classes are separated and labeled as in Table 1. For each compound, the overall best score is reported, regardless of the scale factor and the frequency of occurrence.

using the consensus bit positions of every class. Since the number of bits is very small in these cases and since the consensus patterns are overlapping (see Figure 3), these bit

subsets alone have, as one may expect, very little discriminatory power. Thus, in these calculations, many false positives were recognized and only scores close to zero obtained.

Preferred Tc Threshold Values and Scale Factors. Since we systematically modified Tc values and scale factors during our calculations, we also analyzed the distribution of best scores obtained for each compound relative to those parameters. The results are summarized in Table 6, and a graphical representation is shown in Figure 6. In the absence of scaling, best results were obtained at Tc cutoff values between 0.7 and 0.8 with peak performance at 0.75. For scaling, the preferred range shifted to a higher Tc range between 0.8 and 0.9, dependent on the scale factor. When we applied these preferred parameters for exhaustive virtual screening of our test database, the effect of fingerprint scaling was further emphasized. For unscaled calculations using a Tc cutoff value of 0.75, on average 35.5% of active compounds were correctly recognized, whereas 66.9% were correctly recognized when scaling was applied using a Tc value of 0.85 and scale factor of 5. Since fixed parameters were applied in these calculations, search conditions were

Table 6. Tc-Dependent Distribution of Top Scores^a

Tc cut-off	0.6				0.65					0.7					0.75					0.8					0.85					0.9					0.95					1							
Scale factor	1	1	2	3	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5			
BA_5HT								4							9					13	3				3	9	7	2				4	14	8				1									
BA_BEN								10							13	1				6		1					1	6	16				1	1													
BA_CAE		2	3					11	1						7		2			6	5	1	6		2	4	7	4	1	2		3	5	7	2						2						
BA_COX			1					3							11	1				4	4				3	1	7	1		3			6	11													
BA_H3E			3					5							13	1	1			12	4	1	1	2	5	3	6	3	3	5		2	7	4	3					2							
BA_HIV		3	4					9	2						1	3	4			3	1	5	3		4	1	1	2	7	2	2	2	3	2	2			2	2	2							
BA_TKE			4					6							7		1			6	3		4		3	1	4	4	3			2	7	6													
CMC_ACE								5							3					6	4				8	1	3	2		2	6	3	1	4				1	6								
CMC_ADR	1	4						10							2	3				1		7	1					6	12																		
CMC ESTR								7	1						4	3	2			3	2	3	4	1	3	2	3	2	1	3	3	3	2	1	3	3	3	3	3	3	3	3	3	3	3	3	3
CMC_GLU	1	1						2											2	1					4			1	4	5		1			4		3	2	4	4							
CMC_hH2	1	2	1					8	2	2					7	8	4	2	2	7	7	5	6	4	7	7	7	7	5	6	6	7	7	7	6	6	6	6	6	6	6	6	6	6	6	6	
NP_BLC	8	3	7	8				2			7	13	8					7	12																												
NP_PKC	1	2	1					2	1	1					12	6	1	1	1	9	13	11	4	1	7	10	9	13	12	5	5	6	8	8	5	5	5	5	5	5	5	5	5	5	5	5	

^a For scaled and unscaled calculations, the Tc value range is shown where best compound scores were observed. The number of top compound scores is reported for each Tc cutoff value and scaling factor. Top scores may occur several times at more than one Tc cutoff value. Note that no scaled calculation produced a top score at a Tc value of 0.6.

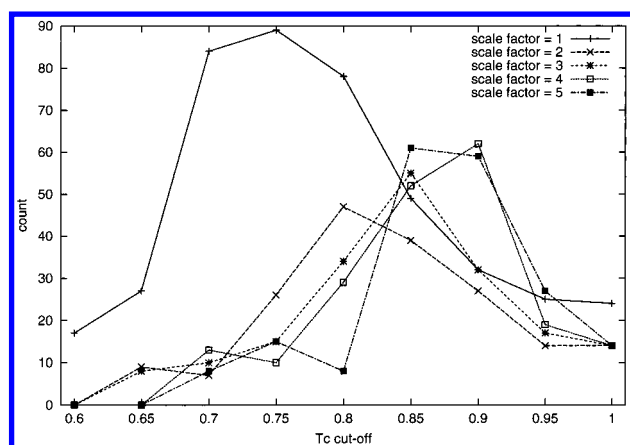


Figure 6. Distribution of top scores. For each Tc cutoff value, "count" represents the total number of best scores observed at this value, as reported in Table 6. If the top score for a compound occurs at several Tc values, it is counted several times. Score distributions are compared for different scale factors.

no longer optimal for each compound and activity class, which resulted in a slight increase in false positives. In unscaled calculations at a Tc of 0.75, on average approximately two false positives per calculation were found, whereas six false positives were detected under scaling conditions, which is in part a consequence of applying a large scale factor. However, most of this increase was due to much larger numbers of false positives in only one or two activity classes, in particular, glucocorticoid analogues, as mentioned earlier. However, if the scaled calculations were repeated at the Tc threshold value of 0.85 but using a scale factor of 4 instead of 5, 53.5% of all active compounds were still correctly identified but now the average number of false positives was reduced to approximately three compounds per calculation. Thus, scaling produced overall better results in our calculations using MACCS-FP.

Conclusions. We have carried out systematic virtual screening calculations on a variety of biologically active and

drug-like molecules using original and modified versions of a structural fragment-based molecular fingerprint. The scaling technique introduced here emphasizes consensus bit settings in keyed fingerprint representations that result from comparison of sets of compounds with similar activity. Such consensus patterns can be variably defined. Fingerprint profiles make it possible to define consensus patterns as a function of relative bit occupancy. For example, all bit positions having relative bit occupancy of at least 0.8 could be included in a consensus pattern. Here we have deliberately focused on fully occupied bit positions to minimize the number of consensus bits per class. On the basis of systematic test calculations, the overall performance of our virtual screening analysis was increased by approximately 30% when scale factors were applied to consensus bits for pairwise comparison of molecules. However, scaling did not significantly increase the number of false positives. Taken together, these findings suggest that even relatively simple and in part overlapping fingerprint consensus patterns encode "class-specific" molecular features that can be exploited when searching for similar molecules. On the basis of our calculations, preferred Tc values and scale factors for virtual screening using MACCS-FP could be suggested. Equivalent analyses can be carried out for any keyed fingerprint representation.

REFERENCES AND NOTES

- (1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening — an overview. *Drug Discovery Today* **1998**, *3*, 160–178.
- (2) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078–1082.
- (3) Gane, P. J.; Dean, P. M. Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401–404.
- (4) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen* **2000**, *3*, 363–372.
- (5) Barnard, J. M. Substructure searching methods. Old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
- (6) James, C. A.; Weininger, D. Daylight theory manual; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

- (7) Good, A. C.; Mason, J. S. Three-dimensional structure databases searching. *Rev. Comput. Chem.* **1996**, 7, 67–117.
- (8) Clark, D. E.; Jones, G.; Willett, P. Kenny, P. W.; Glen, R. C. Pharmacophoric pattern matching in files of three-dimensional structures: Comparison of conformational-searching algorithms for flexible searching. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 197–206.
- (9) Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959–5967.
- (10) Hopfinger, A. J.; Wang, S.; Tobarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, 119, 10509–10524.
- (11) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 569–574.
- (12) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, 5, 576–587.
- (13) Labute, P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, 4, 444–455.
- (14) Chen, X.; Rusinko, A. III; Young, S. S. Recursive partitioning analysis of a large structure–activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 1054–1062.
- (15) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 983–996.
- (16) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL “Keys” as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 443–448.
- (17) MACCS keys; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
- (18) Brown, R. D.; Martin, Y. C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 572–584.
- (19) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 801–809.
- (20) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 699–704.
- (21) Xue, L.; Godden, J.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1227–1234.
- (22) CMC-3D (Comprehensive Medicinal Chemistry Database), version 99.1; MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
- (23) Chapman & Hall, Dictionary of Natural Products, CD-ROM version 1999; CRC Press LLC: 2000 NW Corporate Blvd., Boca Raton, FL 33431.
- (24) Available Chemicals Directory (ACD); MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.
- (25) MOE (Molecular Operating Environment); Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
- (26) Santavy, M.; Labute, P. SVL: The Scientific Vector Language. *Journal of the Chemical Computing Group*, <http://www.chemcomp.com/feature/svl.htm>.
- (27) Godden, J. W.; Xue, L.; Stahura, F. L.; Bajorath, J. Searching for molecules with similar biological activity. *Pac. Symp. Biocomput.* **2000**, 5, 566–575.

CI000311T