Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys

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

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received March 31, 2003

The concept of compound class-specific profiling and scaling of molecular fingerprints for similarity searching is discussed and applied to newly designed fingerprint representations. The approach is based on the analysis of characteristic patterns of bits in keyed fingerprints that are set on in compounds having equivalent biological activity. Once a fingerprint profile is generated for a particular activity class, scaling factors that are weighted according to observed bit frequencies are applied to signature bit positions when searching for similar compounds. In systematic similarity search calculations over 23 diverse activity classes, profile scaling consistently increased the performance of fingerprints containing property descriptors and/or structural keys. A significant improvement of ∼15% was observed for a new fingerprint consisting of binary encoded molecular property descriptors and structural keys. Under scaling conditions, this fingerprint, termed MP-MFP, correctly recognized on average close to 60% of all active test compounds, with only a few false positives. MP-MFP outperformed MACCS keys and other reference fingerprints. In general, optimum performance in scaling calculations was achieved at higher threshold values of the Tanimoto coefficient than in nonscaled calculations, thereby increasing the search selectivity. In general, putting relatively high weight on signature bit positions that were always, or almost always, set on was found to be the most effective scaling procedure. Analysis of class-specific search performance revealed that profile scaling of MP-MFP improved the similarity search results for each of the 23 activity classes.

INTRODUCTION

Similarity searching using molecular fingerprints is among the most widely used approaches for virtual screening of compound databases.[1,2] Fingerprints capture structural or topological features and/or properties of molecules in a binary bit string format but often differ greatly in size and complexity.[3] Regardless of their design, the use of fingerprints for similarity searching requires the calculation of bit strings for both query and database compounds and their quantitative comparison, for which a variety of similarity metrics are available.[4] In essence, fingerprint overlap is expressed as a value of a chosen similarity coefficient, most frequently the Tanimoto coefficient (Tc),[4] and if this value exceeds a predefined threshold, the compared molecules are regarded as being similar.

In principle, fingerprints generate linear and rather abstract representations of molecular structure and properties that are straightforward to calculate and compare (provided, however, the molecular descriptors used for their design are easy to calculate as well). In these calculations, the assessment of molecular similarity is carried out in fingerprint-specific vector spaces and naturally depends, to a significant extent, on the design and specific features of these reference spaces. Thus, as is essentially the case with any descriptor-based molecular similarity approach, there is no generally valid answer to the question which fingerprint designs are overall the best because their suitability typically much depends on the specifics of the search problem under investigation.[3] However, what can be concluded with some degree of certainty, based on currently available data, is that 3D fingerprints are not generally superior to 2D representations and that complex designs do not necessarily perform better than simpler ones.[5−7]

One of the principal differences between various fingerprint designs is whether or not their bit positions can be associated with specific chemical features or descriptor values. This is the case in keyed designs, for example, structural key-type fingerprints,[8,9] where each bit position monitors the presence or absence of a specific structural fragment. By contrast, this is not the case in hashed or folded representations[10] where diverse topological or chemical features are mapped to corresponding or overlapping bit segments and, hence, single bit positions or bit segments lose physical meaning. Keyed fingerprint designs can vary dramatically in length, ranging from MACCS keys[9] or so-called mini-fingerprints (MFPs)[7,11] that consist of ∼100−200 bits to pharmacophore fingerprints[12,13] that can consist of millions of bits. In the latter fingerprints, each bit position detects the presence or absence of a specific geometric pharmacophore arrangement in a test molecule.

We are particularly interested in the design of keyed fingerprints, as exemplified by MFPs[11] that combine selected

* Corresponding author phone: (425)424-7297, fax: (425)424-7299, e-mail: jurgen.bajorath@albmolecular.com. Corresponding author address: Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, WA 98011.

[†] Bothell Research Center (AMRI-BRC).

[‡] University of Washington.

10.1021/ci030287u CCC: $25.00
Published on Web 06/20/2003

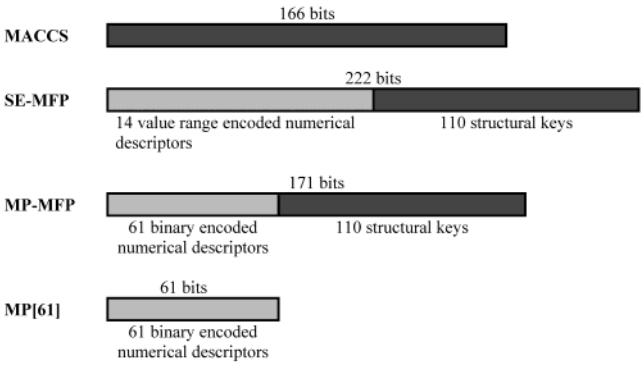

Figure 1. Fingerprints. The figure shows a schematic comparison of the fingerprints evaluated in this study. Light gray bit segments represent (either value range or binary encoded) numerical descriptors and dark gray segments structural keys.

property and structural descriptors, because keyed representations often allow chemical interpretation of bit settings or determination of bit patterns that might be characteristic for series of active compounds.[14] To exploit this type of information in similarity searching, we have previously applied scaling factors to bit positions in MACCS keys that were always or mostly set on in specific compound classes, which increased the number of correctly identified compounds for several activity classes.[15] These initial findings suggested that scaling of consensus bit positions might also have the potential to increase the performance of other keyed fingerprints in similarity searching.

Here we have combined fingerprint profiling and scaling, introduced a weighting scheme that discriminates between different levels of consensus bit settings, and applied these techniques to recently designed second generation MFPs. In systematic similarity search calculations in a database containing a variety of active compounds and an excess of randomly selected background molecules, profile scaling significantly increased the search performance of MFPs and MACCS keys over a wide range of scaling factors. One of our MFPs containing binary encoded property descriptors and structural keys outperformed other fingerprints under both nonscaled and scaled calculation conditions and displayed the largest increase in performance when scaling was applied.

METHODS

Fingerprints. In this study, we have focused on the assessment of four fingerprints, two recently generated MFPs, SE-MFP[11] and MP-MFP,[16] and two reference fingerprints, MACCS keys and MP[61], a fragment of MP-MFP consisting of 61 binary encoded property descriptors. These fingerprints are schematically shown in Figure 1. MFPs were originally designed as short fingerprint representations consisting of only a limited number of selected value range encoded molecular property descriptors and structural keys,[11] which were specifically selected based on their ability to correctly recognize different types of biologically active compounds.[7,11] In our subsequent design efforts, we continued to focus on combinations of numerical descriptors and structural keys, as they were consistently among the most powerful descriptor combinations identified by us,[17] but attempted to establish more general descriptor selection schemes that are independent of specific classes of active compounds.

First, we selected molecular property descriptors that had consistently high information content in various compound databases[18,19] but displayed little database-specific differences.[19] Fourteen such descriptors were selected (see Supporting Information) and combined with a subset of structural keys[9] (e.g., specific atom types or bond patterns, structural fragments, or functional groups; stored in a vector table format) that occurred in more than 10% and less than 90% of the compounds in an in-house collection of drug-like molecules, leading to the design of SE-MFP,[11] as shown in Figure 1. In SE-MFP, the 14 numerical descriptors are value-range encoded. For each descriptor, eight bits are reserved and descriptor values are encoded incrementally (for example, if zero to one hydrogen bond acceptor is present, bit one is set on; if two to three are present, bits one and two are set on; and so on).

Second, we made use of the concept of statistical medians that we had also applied in the design of a novel partitioning method.[20] The median is defined as the value that separates a distribution of values into two equally populated halves above and below the median.[21] For a large pool of ~150 numerical descriptors, value distributions and medians were calculated in a large in-house database containing approximately 1.34 million unique molecules[20] and the most information-rich and least correlated descriptors were determined, leading to the selection of 61 descriptors for MP-MFP (see Supporting Information),[16] shown in Figure 1. These descriptors (MP[61]) were combined with the same subset of structural keys as in SE-MFP. In MP-MFP, the property descriptors are binary encoded. This means that each descriptor is assigned one bit position and the bit is set to one if the descriptor value for a test compound is equal or larger than its stored database median or set to zero if it is smaller. Therefore, MP-MFP calculations involve the transformation of value ranges of numerical descriptors into a binary classification scheme. On a first glance, this is a low-resolution procedure. However, in MP-MFP contributions from 61 diverse descriptors are taken into account using only 61 bit positions, whereas 112 bits are required in SE-MFP to capture contributions from 14 descriptors. Thus, binary encoding permits the inclusion of many more descriptor contributions in small fingerprints.

Different from other bit string designs, bit positions set to zero in MP-MFP capture as much information as those set to one and must be taken into account when fingerprint overlap is quantified. Therefore, we have introduced a modification of the Tanimoto coefficient (Tc)[4] called averaged Tc (avTc) for similarity searching with MP-MFP, as explained previously,[16] and also MP[61]. The Tc is conventionally defined as

$$\mathrm{Tc} = \mathrm{bc}/(\mathrm{b1} + \mathrm{b2} - \mathrm{bc})$$

with b1 being the number of bits set on in molecule 1, b2 the number of bits set on in molecule 2, and bc the number of bits set on shared by both molecules under comparison. We define avTc as

$$\mathrm{avTc} = (\mathrm{Tc} + \mathrm{Tc}')/2$$

When calculating Tc′, bit positions set to zero are counted, rather than those set to one as in convetional Tc calculations. For SE-MFP and MACCS, standard Tc values were calculated.

Table 1. Compound Classes for Fingerprint Profiling and Similarity Searching[a]

class	activity	search set	profile set
BLC	β-lactamase inhibitors	7	7
PKC	protein kinase C inhibitors	7	8
ADR	antiadrenergic (β-receptor)	8	8
GLU	glucocorticoid analogues	7	7
BEN	benzodiazepine receptor ligands	11	11
CAE	carbonic anhydrase II inhibitors	11	11
H3E	H3 antagonists	10	11
TKE	tyrosine kinase inhibitors	10	10
5HT	serotonin receptor ligands	10	11
HIV	HIV protease inhibitors	9	9
COX	cyclooxygenase-2 inhibitors	8	9
ANG	angiotensin AT1 antagonists	5	5
ARO	aromatase Inhibitors	5	5
DIH	dihydrofolate reductase inhibitors	5	6
FAC	factor Xa inhibitors	7	7
MAT	matrix metalloproteinase inhibitors	6	6
VIT	vitamin D analogues	6	6
RTI	reverse transcriptase inhibitors	7	8
PPAR	PPARgamma agonists	8	8
DD2A	dopamine D2 antagonists	7	7
CRF1	CRF1 antagonists	6	6
CALA	calcium antagonists	9	9
ARI	aldose reductase inhibitors	6	6

[a] In each case, the Search Set reports the number of randomly selected active database compounds (potential hits for similarity searching), and the Profile Set reports the number of randomly selected active compounds used for fingerprint profile generation.

Database. For similarity searching, a previously assembled compound database[11] was used consisting of 23 different biological activity classes plus 5000 randomly collected background compounds. The composition of this data set is reported in Table 1. For many classes, the database contains a number of rather diverse structures having similar activity. Closely related analogues were generally not included based on visual inspection. For the calculation of fingerprint profiles, as further described in the Results section, each activity class was randomly divided in half and only the 175 active test compounds were retained in the source database. Thus, profile and search calculations were carried on different subsets of compounds with equivalent activity (essentially corresponding to learning and test sets). In a virtual screening situation, the subset of compounds used for profile calculation would be known and the test compounds would be potential hits in the source database.

Similarity Searching. Each of the 175 active test compounds was searched against the remainder of the database under systematic variation of the fingerprints, Tc/avTc threshold values (from zero to one in 0.01 increments), profile cutoff values (0.5, 0.55, 0.6 ... 0.95, 1.0), and profile scaling factors (1, 2 ... 10, 20 ... 50). Profile cutoff values and scaling factors will be further discussed below. A scaling factor of 1 corresponds to nonscaled conditions. Thus, for each of the 175 test molecules, ~62 000 similarity search calculations were carried out, optimum performance levels were determined, and averages calculated. Each calculation was scored using the following function

$$S = (C - I)/N$$

with C and I being the number of correctly and incorrectly identified compounds per activity class, respectively, and N the total number of test compounds in each class. If C was smaller than I, the score was set to zero. This scoring scheme attempts to maximize the number of correctly identified compounds having similar activity and minimize the number of false positives.

All routines required for fingerprint profiling, scaling, and similarity searching were written in SVL[22] and implemented in the Molecular Operating Environment.[23] Essentially, the only rate-limiting step for similarity searching under scaling conditions is the generation of fingerprints for database compounds. For the MFP-type designs applied here, 40−50 molecules can be processed per second on a single 2 GHz processor (including structure input). Thus, a database containing approximately a million molecules can be set up for similarity searching in less than 9 h of CPU time.

RESULTS AND DISCUSSION

Fingerprint Profiling. The concept of fingerprint profile scaling is based on the observation that compounds belonging to different biological activity classes often produce characteristic bit patterns in keyed fingerprints.[14] We have calculated profiles for all four fingerprints and the 23 activity classes studied here. Representative example profiles are shown in Figure 2. For a set of compounds, a fingerprint profile is calculated by adding all bits at each position and dividing the sum by the total number of compounds. Thus, the profile reports the relative frequency of bit settings between 0.0 (a bit position is never set on) and 1.0 (a bit is always set on). For our test cases, as summarized in Table 1, the number of compounds used for profiling of different activity classes ranged from five to 11. For MP-MFP and MP[61], where bit settings of zero and one are equivalent, as discussed above, two profiles were calculated for each class, one monitoring bits set to one, the other bits set to zero (the latter being the inverse profile of the former). Although only a relatively small number of active compounds were available for profile generation, we consistently found that different classes of compounds in our test database produced distinct bit patterns and diverse sets of consensus positions, defined as bit positions set on with high frequency. Representative examples of consensus bit patterns are shown in Figure 3. Different threshold frequencies of bit occurrence can be used as profile cutoff values in order to define consensus bit positions. The fact that different consensus patterns are produced by profiling of different activity classes suggests that fingerprint profiles capture compound class-specific information that can be exploited by scaling, as described in the following.

Scaling. If fingerprint profiles encode some class-specific information, the application of scale factors to consensus bit positions should emphasize these signature bit patterns in similarity search calculations and thereby increase the probability of identifying compounds with similar activity. The process of profile scaling is illustrated in Figure 4. In this example, the MP[61] profile of 5HT is shown with five consensus bit positions obtained for a profile cutoff value of 0.8. Furthermore, model fingerprints are shown for two active molecules A and B, the comparison of which results in a Tc value of ~0.5. When scaling factors are applied to consensus positions shared within this activity class, fingerprint overlap increases and results in a higher Tc value. Thus, scaling of consensus bit positions emphasizes class-specific

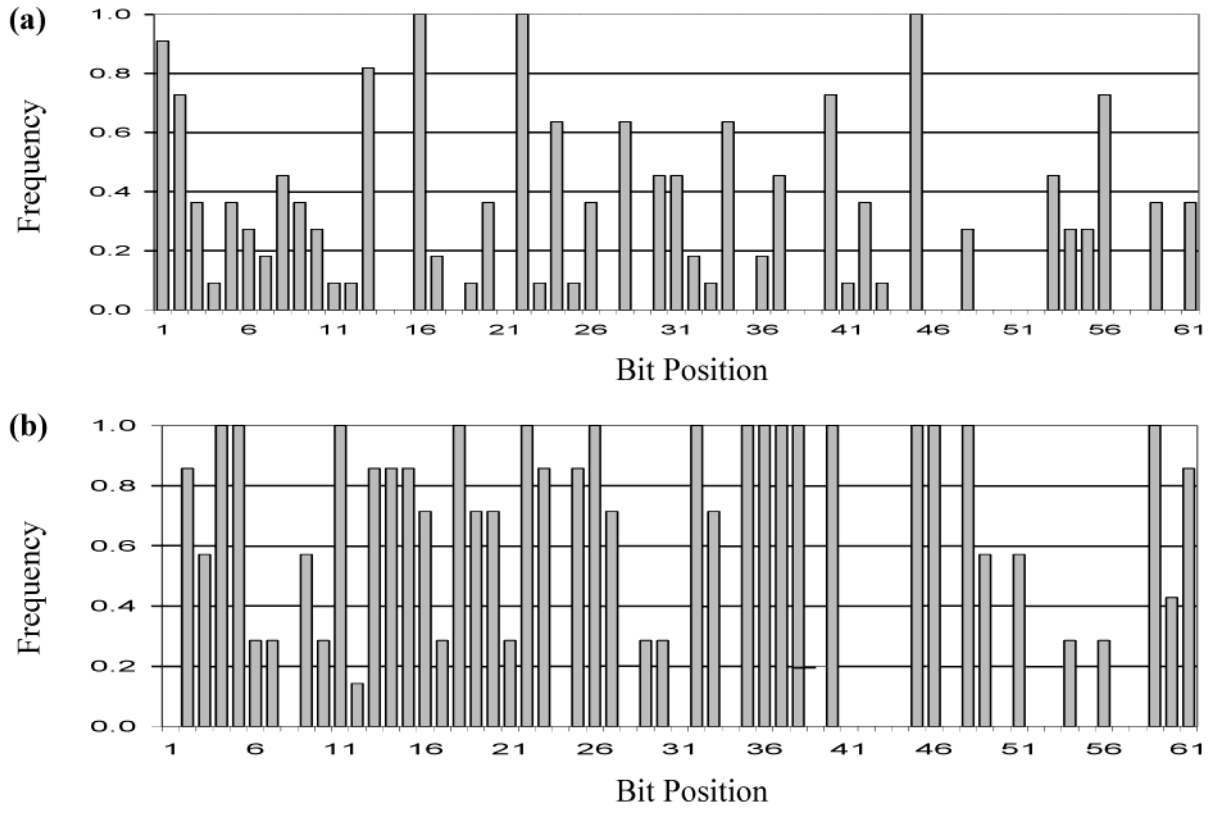

Figure 2. Fingerprint profiles. Two examples of MP[61] profiles (monitoring bit settings to one) are shown for activity classes 5HT (a) and FAC (b). Frequency stands for the relative frequency of bit occurrence.

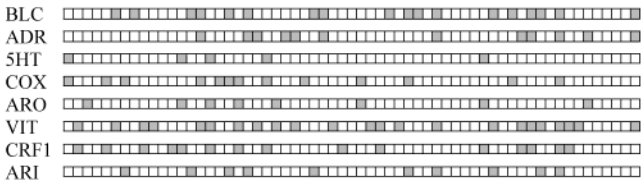

Figure 3. Consensus bit patterns. For eight activity classes, consensus bit settings are shown derived from the type of profiles shown in Figure 2. Gray shading indicates bit positions set on with a frequency or profile cutoff value of at least 0.8. For each class, a characteristic consensus bit pattern is obtained.

features. For example, in a similarity search calculation using a Tc threshold value of 0.7, the similarity of compounds A and B would have only been recognized under scaling conditions.

Figure 4 also illustrates the basic idea of weighted scaling that we introduce here. Rather than applying the same profile scaling factor (e.g., five) to all consensus bit positions above the profile cutoff of 0.8, the scaling factor is linearly weighted within the frequency interval [0.8,1.0]. This weighting procedure is more accurate than application of constant factors, since it takes differences in bit frequency among consensus positions into account. For example, applying a profile cutoff value of 0.6 and a scaling factor of 10.0 would result in effective or weighted scaling factors between 6.0 and 10.0 for the obtained consensus bit positions. For our systematic similarity search analysis, as described in the following, where we explored a wide range of profile cutoff values and scaling factors, the application of the weighted scaling procedure was considered to be an important factor to ensure frequency-related scaling of consensus bits.

Systematic Similarity Searching. A key question for this analysis has been whether profile scaling indeed systematically increases the search performance of the types of fingerprints studied here, as one might expect based on the considerations discussed above. In other words, are fingerprint profiles capable of identifying signature bit patterns with sufficient specific information to distinguish between different activity classes? To answer these questions, we designed a systematic similarity search experiment over 23 activity classes using four fingerprints. Considering the specifics of the profile scaling approach, as discussed above, three parameters substantially affect the calculations: the Tc similarity threshold value, the profile cutoff value (determining the number of consensus bit positions), and the profile scaling factor. In our calculations, these parameters were systematically varied over a wide range (see Methods), and a total of approximately 11 million similarity searches were carried out. The obtained results are summarized in Table 2 and make it possible to draw a number of conclusions.

Under nonscaled conditions, both SE-MFP and MP-MFP performed somewhat better than MACCS, with MP-MFP yielding highest scores. It correctly recognized an average of approximately 43% of all active compounds and only 0.018% false positives. For all tested fingerprints, best performance was achieved at similar Tc threshold values, within the interval 0.74−0.81. Interestingly, our short reference fingerprint MP[61] performed almost as well as MACCS, with also about 35% correct recognitions but a slightly higher false positive rate.

A key finding has been that profile scaling significantly improved the scores and performance of the fingerprints tested here. In the case of MP[61], the rate of correct

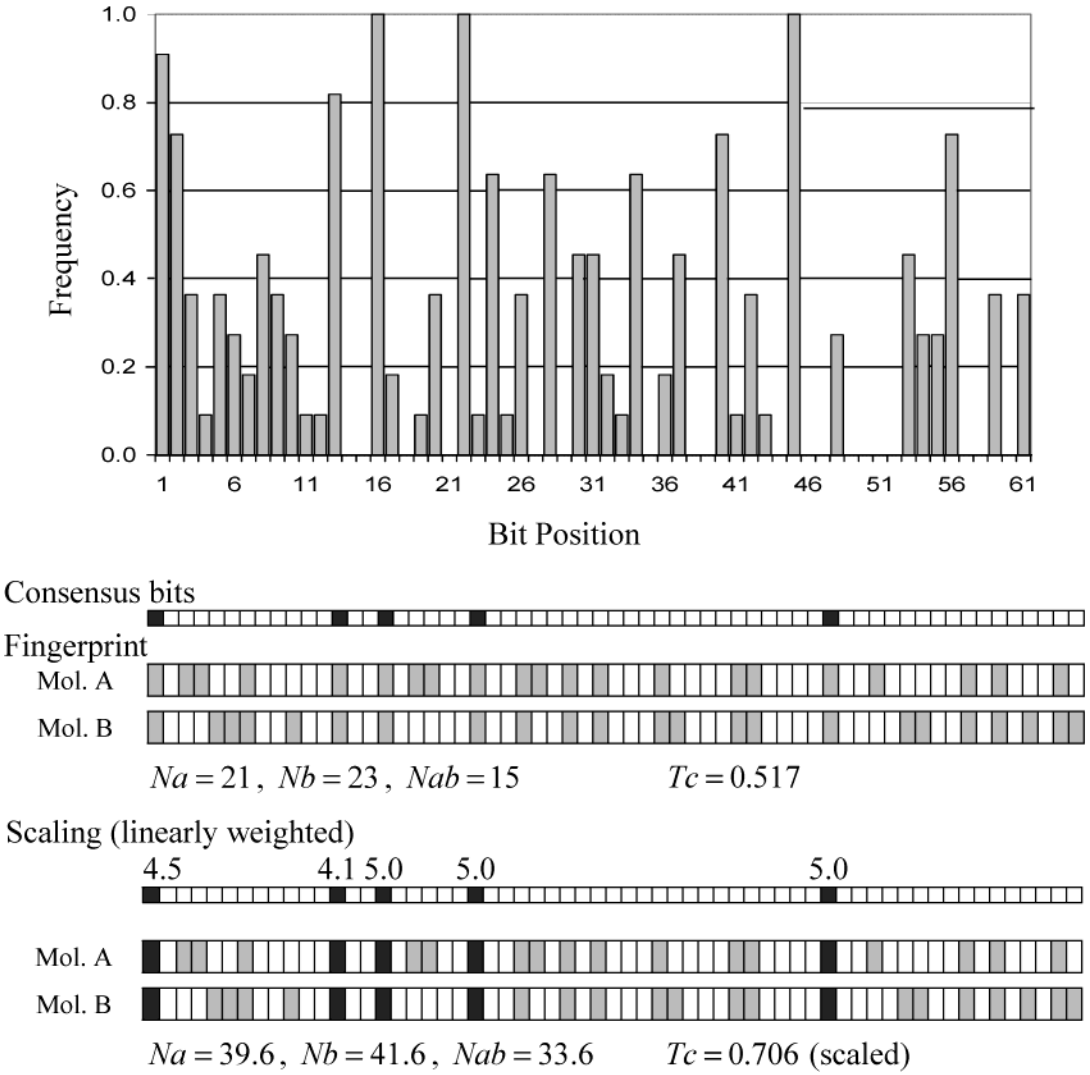

Figure 4. Profile scaling and weighting of scale factors. As an example, the MP[61] profile of activity class 5HT is shown. Consensus bits for a frequency or profile cutoff value of 0.8 are indicated in dark gray. Model fingerprints of two molecules A and B having similar activity are also shown (bits set to one are shaded light gray) and the Tc value for their comparison is reported. At the bottom, a scaling factor of 5.0 is applied and linearly weighted within the [0.8,1.0] frequency interval; i.e., bits always set on are scaled by 5.0, bits set on with a frequency of 0.9 by a factor of 4.5, and those with a frequency of 0.8 by a factor of 4.0. Under scaling conditions, the Tc value increases, thereby emphasizing the similarity of these compounds. The model Tc calculation for bit settings of one, as shown here, represents half of the avTc calculation carried out for Mp[61] and MP-MFP (see Methods).

recognitions increased by ∼3.8% under optimum scaling conditions and the false positive rate decreased (from about two false positives to less than one per calculation). MACCS keys showed a ∼12.5% increase in correct recognition and a slight increase in the false positive rate (from less than one to ∼1.5 false positives per calculation). A similar increase in correct recognitions was observed for SE-MFP but a decrease in false positives, yielding higher scores. By far the best performance under scaling conditions was seen for MP-MFP, with a ∼15.5% increase in correct recognitions and constant (and low) false positive rate. This overall performance corresponded to about five correctly identified active compounds per calculations and less than one false positive and an increase of, on average, ∼1.5 correct recognitions per calculation as a consequence of scaling.

Parameter Analysis. Interestingly, all fingerprints displayed optimum performance when a profile scale factor of 10 was applied. In addition, the fingerprints containing structural keys achieved overall best performance at a profile

Table 2. Similarity Searching in the Presence and Absence of Profile Scaling[a]

fingerprint	scaling	PSF	PCV	Score	(av)Tc	%r	%w
MACCS	yes	10	0.95	0.3616	0.94	47.86	0.030
	no			0.2920	0.78	35.30	0.015
SE-MFP	yes	10	0.95	0.3986	0.96	46.46	0.012
	no			0.3255	0.81	38.26	0.016
MP-MFP	yes	10	0.95	0.4958	0.90	58.72	0.018
	no			0.3574	0.74	43.09	0.018
MP61	yes	10	0.75	0.3314	0.90	38.81	0.012
	no			0.2572	0.75	35.06	0.040

[a] The table summarizes the results of exhaustive similarity search calculations. For each fingerprint, the Tc threshold value is reported for the highest scoring nonscaled calculations, and the profile scaling factor (PSF), profile cutoff value (PCV), and Tc threshold value are reported for the highest scoring scaled calculations. The abbreviation %r stands for %right and reports the percentage of correctly identified active compounds averaged over all activity classes, and %w means %wrong and reports the average percentage of false positives (including both active and/or background compounds).

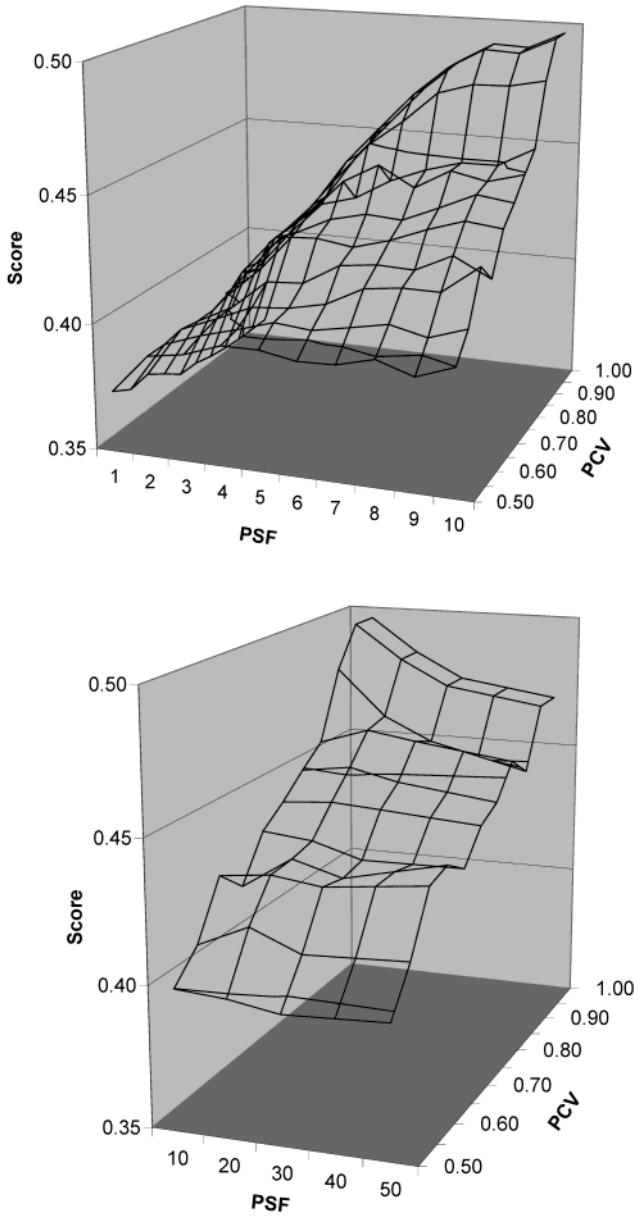

Figure 5. Relationship between scores, scale factors, and cutoff values. The 3D graphs summarize similarity search calculations on MP-MFP as an example. Averaged scores are plotted for all combinations of profile scale factors (PFS) and profile cutoff values (PCV). The top graph monitors the PFS range from one to 10 and the bottom graph the range from 10 to 50 (with a change in relative scale). The maximum score was obtained for a scale factor of 10 and cutoff value of 0.95.

cutoff level of 0.95. Thus, for these three fingerprints, the combination of a scale factor of 10 and a profile cutoff value of 0.95 was the overall preferred scaling condition. Best performance at a high profile cutoff value means that relatively few consensus bit positions need to be emphasized in order to increase the number of correct recognitions and/or decrease false-positive rates. This also implies that structures of molecules with similar activity do not need to be very similar for profile scaling to be effective (lowering the profile cutoff value generally increases the number of consensus bits).

In Figure 5, representative 3D plots are shown for profile scaling that relate scores, scale factors, and cutoff values to each other. As can be seen, scores significantly increase toward a scale factor of 10 and higher cutoff values, reach a maximum, and then gradually decrease for higher scale factors. It follows that profile scaling was overall most effective when a relatively high scaling factor was applied to consensus bits that were mostly set on in all compounds within an activity class. Analysis of consensus bit statistics provided a rationale for these findings. For example, for MP-MFP, consensus bit positions that were always set on ranged from 13 (PKC) to 57 (FAC), with an average of 37 bits over all 23 activity classes. Thus, for MP-MFP, our best performing fingerprint, an average of ~20% of all bit positions were subjected to scaling at the 0.95 cutoff level and Tc calculation. MP[61] reached best overall performance at a lower cutoff level of 0.75 because, in the absence of 110 structural keys, which contributed consensus bits to the other fingerprints, more consensus bits within the 61-bit segment were required for scaling to be effective.

In control calculations, we also applied different scaling schemes to all bit positions in fingerprint profiles (no profile cutoff criterion) and consistently observed lower scores than reported in Table 2 for optimized profile cutoff values. These finding were not unexpected. Increasing the number of consensus bits by lowering (and, ultimately, omitting) the profile cutoff value subjects bit positions to scaling that are variably set within (and also between) activity classes and affects the sensitivity of scaling calculations by increasing the "noise" and false positive rates. By contrast, signature bits that are always or mostly set on are more likely to emphasize compound class-specific features in similarity search calculations and discriminate between different activity classes under scaling conditions. This rationalizes the introduction of the profile cutoff parameter and further explains our finding that overall best performance was observed at high profile cutoff values and scaling factors.

Fingerprint profiles should in general become more accurate with increasing number of available active compounds. Therefore, to study potential profile changes due to increasing compound numbers, we also calculated fingerprint profiles of activity classes for the combined profile and search sets. A representative example is shown in Figure 6. As to be expected, some changes in bit frequencies were observed, but the profiles looked rather similar. In particular, consensus bit positions at higher profile cutoff values were little affected. Therefore, at least in these cases, between six and 11 compounds were sufficient to generate profiles that could successfully be scaled.

Finally, we also analyzed the compound class-specific predictions under scaled and nonscaled conditions. The results were encouraging. For MACCS keys and SE-MFP, profile scaling improved the search results for 22 and 21 of all activity classes, respectively (exceptions being class ARO for MACCS and classes PKC and ARI for SE-MFP). For MP-MFP and also MP[61], scaling improved the search performance of each of the 23 activity classes, with no exception. For MP-MFP, some of the observed improvements were rather dramatic. For example, for TKE and ANG, scaling improved class-specific scores from 0.16 to 0.95 and from 0.40 to 1.0, respectively. Furthermore, nearly perfect or perfect (BLC, ANG, VIT) predictions were achieved for seven activity classes under scaling conditions. For activity class VIT, scaling of SE-MFP also produced a perfect prediction. In a number of cases, different combinations of

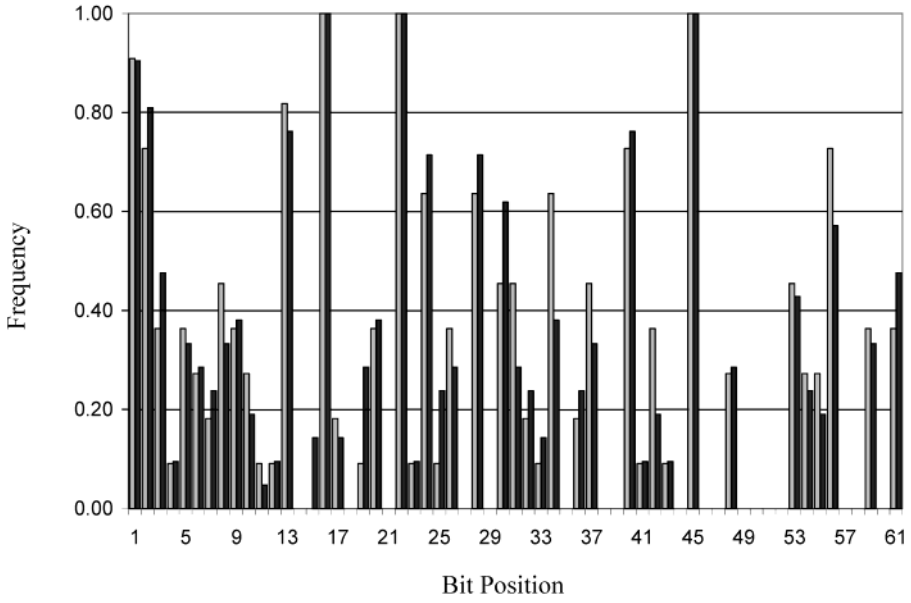

Figure 6. Profile comparison. Activity class 5HT is shown as an example. MP[61] profiles were calculated for the profile set (light gray) and the combined profile and search sets (dark gray), which essentially doubles the number of compounds used for profiling.

scaling factors and profile cutoff values produced similarly high scores for MACCS, SE-MFP, and MP-MFP. Thus, the analysis of class-specific similarity search performances further supported the usefulness of the profile scaling concept.

CONCLUSIONS

In this study, we have investigated the potential of profile scaling for similarity searching using fingerprints of moderate size and complexity. The profile scaling method is well suited for keyed fingerprint designs. It has been automated and requires little subjective intervention. Profile scaling consistently and often significantly increased the performance of fingerprints investigated herein. In our calculations, a hybrid fingerprint consisting of a combination of binary encoded numerical descriptors and structural keys performed best under both standard and scaling conditions. Thus, it should be worthwhile to further pursue this fingerprint design strategy. The observed increase in search performance under scaling conditions suggests that differences in bit frequencies, which are easily detected in fingerprint profiles, have substantial predictive value. Since the availability of a known compound set is the major requirement for profiling scaling, it was important to find out that a fairly limited number of active compounds were sufficient to generate meaningful profiles. This has implications for virtual screening because, in practice, profile scaling should be an attractive approach whenever at least a few active template compounds are available. In addition, application of profile scaling techniques might often alleviate the need for similarity search calculations using multiple templates because characteristic features are emphasized that are shared by compounds having similar activity.

Supporting Information Available: Table of numerical descriptors in SE-MFP and MP-MFP. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **2002**, *1*, 882−894.
(3) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(4) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit strings. *Combin. Chem. High Throughput Screen.* **2002**, *5*, 155−166.
(5) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(6) Matter, H.; Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(7) Xue, L.; Stahura F. L.; Godden, J. W.; Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394−401.
(8) 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.
(9) MACCS keys. MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA.
(10) James, C. A.; Weininger, D. Daylight fingerprints. Daylight theory manual. Daylight Chemical Information Systems, Inc., Irvine, CA.
(11) Xue, L.; Godden, J. W.; Bajorath, J. Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory. *SAR QSAR Environ. Res.* **2003**, *14*, 27−40.
(12) 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.
(13) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.
(14) Godden, J. W.; Stahura, F. L.; Xue, L.; Bajorath, J. Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pac. Symp. Biocomput.* **2000**, *5*, 566−575.
(15) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746−753.
(16) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, in press.

(17) 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.
(18) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.
(19) Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.
(20) Godden J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median partitioning: a novel method for the selection of representative subsets from large compound pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885−893.
(21) Meier, P. C.; Zünd, R. E. *Statistical methods in analytical chemistry;* John Wiley & Sons: New York, 2000.
(22) Sanatvy, M.; Labute, P. SVL: the scientific vector language. *Journal of the Chemical Computing Group*, URL: www.chemcomp.com/feature/svl.htm.
(23) Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada.

CI030287U