Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme

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 February 18, 2003

A new fingerprint design concept is introduced that transforms molecular property descriptors into two-state descriptors and thus permits binary encoding. This transformation is based on the calculation of statistical medians of descriptor distributions in large compound collections and alleviates the need for value range encoding of these descriptors. For binary encoded property descriptors, bit positions that are set off capture as much information as bit positions that are set on, different from conventional fingerprint representations. Accordingly, a variant of the Tanimoto coefficient has been defined for comparison of these fingerprints. Following our design idea, a prototypic fingerprint termed MP-MFP was implemented by combining 61 binary encoded property descriptors with 110 structural fragment-type descriptors. The performance of this fingerprint was evaluated in systematic similarity search calculations in a database containing 549 molecules belonging to 38 different activity classes and 5000 background molecules. In these calculations, MP-MFP correctly recognized ∼34% of all similarity relationships, with only 0.04% false positives, and performed better than previous designs and MACCS keys. The results suggest that combinations of simplified two-state property descriptors have predictive value in the analysis of molecular similarity.

INTRODUCTION

Molecular fingerprints are among the most widely used computational tools for similarity searching.[1,2] They are designed to capture structural features and other properties of molecules in a binary bit string format. Similarity search calculations proceed in "fingerprint space", which means that fingerprints are precalculated for query and database compounds and then quantitatively compared using various similarity metrics and coefficients.[1] If similarity coefficients for pairwise comparisons reach or exceed a predefined threshold value, the query and test compounds are considered as "similar". The underlying idea is that fingerprints embody characteristic bit patterns reflecting both common and distinct features in molecules and that this level of abstraction is sufficient to detect molecular similarity. Key aspects of this approach are that one-dimensional bit string comparisons are typically much easier and computationally more efficient than "direct" molecular comparisons and, in addition, that similarities and differences between molecules can be quantified and, ultimately, simply expressed as a "number". Potential caveats are that the assessment of molecular similarity depends a great deal on the specific design of fingerprints[2] and the applied similarity metrics[1,3] and that bit string comparisons have at least some intrinsic statistical limitations.[4,5]

Although essentially all current fingerprints share a linear bit string format, their design, level of sophistication, and complexity varies in part dramatically.[2] Major differences include the way bit settings are generated and the types of molecular descriptors that are encoded. In what are probably the most intuitive representations, each bit accounts for the presence or absence of a specific molecular feature or descriptor value. Prominent among these "keyed" representations are fragment-based designs such as MACCS keys,[6,7] a publicly available version of which consists of 166 fragments (and bits).[7] Even shorter, consisting of fewer than 100 bits, are so-called "mini-fingerprints", keyed designs that combine selected structural keys and value range encoded property descriptors.[8] By contrast, in "hashed" representations, diverse molecular features are mapped to overlapping bit segments.[9] Hashing produces highly characteristic molecular bit patterns but, different from keyed representations, bit positions can no longer be associated with specific features, properties, or values, which makes physical or chemical interpretation of bit patterns difficult, if not impossible. Prominent among hashed designs are Daylight fingerprints,[9] a pioneering development in this field, that consist of up to 2048 bits and capture all possible connectivity pathways through a molecule up to a defined path length.[9] Other fingerprint designs of intermediate size (i.e., ∼1000 bits) include the popular Unity fingerprints[10] that consist of various 2D or 3D descriptors and combine both keyed and hashed design elements or Barnard's fingerprints that are fragment-based,[11] following a design strategy similar to the MDL tools.[7] By far the longest currently available fingerprint representations are 3D pharmacophore fingerprints that can consist of several million bits, with each bit position accounting for the presence or absence of a specific pharmacophore pattern in a molecule[12−14] (as defined by selected atoms or groups and

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com. Correspondence should be addressed at AMRI-BRC.

[†] Albany Molecular Research, Inc.

[‡] University of Washington.

10.1021/ci030285+ CCC: $25.00 © 2003 American Chemical Society
Published on Web 05/13/2003

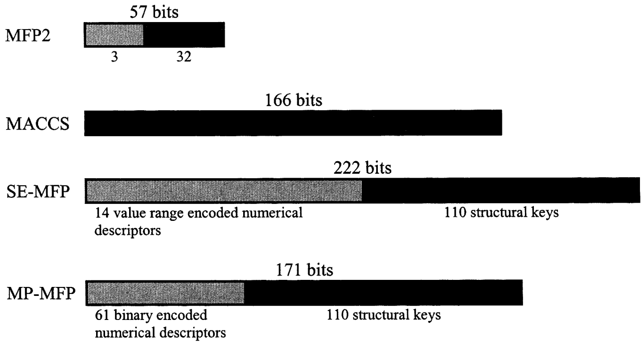

Figure 1. Schematic representation of fingerprints. Light gray bit segments are assigned to property descriptors and dark gray segments to structural keys. The length of the bit strings is approximately scaled.

respective distance ranges). Both 3-point and 4-point pharmacophore fingerprints have been introduced,[14,15] and these fingerprints are calculated based on exhaustive conformational analysis of test molecules.

In virtual screening, fingerprints are often used to search for molecules having similar bioactivity or specificity.[2] An aspect that is often overlooked, if not misunderstood, is that fingerprints are not necessarily designed or trained to recognize similar biological activity but rather to detect structural similarity. The general hypothesis on which these calculations are based is that structural similarity of molecules correlates with similar biological activity, in accord with the similar property principle.[16] Similarly, for pharmacophore fingerprints, it is assumed that the probability of molecules having similar activity increases with increasing overlap of predefined pharmacophore patterns. However, regardless of fingerprint design and complexity, these assumptions may not always be true. In fact, the more the weight that is put on the evaluation of fine structural differences in molecules, the greater is the chance that compared molecules are regarded as “different” and that existing activity relationships are missed.[17]

Here we describe the generation of a new type of fingerprint that is based on a previously unexplored design principle, the transformation of property descriptors with continuous value ranges into a binary classification scheme. This transformation is achieved by calculating statistical medians of descriptor value distributions in large compound databases that are then used as cutoff values for setting bits on or off. Since we first applied the concept of statistical medians when developing a novel partitioning method, median partitioning (MP),[18,19] and since the strategy pursued here is at least conceptually related to our previous mini-fingerprint (MFP) design concept,[8] we call this new fingerprint MP-MFP. Although the design of MP-MFP did not take any bioactivity criteria into account, it performed well in systematic virtual screening-type calculations on diverse classes of active compounds.

FINGERPRINT DESIGN STRATEGY

To rationalize the generation of MP-MFP, we first briefly describe our original MFP design approach. MFPs are short fingerprint representations consisting of only a limited number of selected value range encoded molecular property descriptors and structural keys. In Figure 1 schematic representations are shown of fingerprints discussed in this study. The generation of MFPs was originally catalyzed by our observations that combinations of only a few 1D or 2D descriptors and structural keys were sufficient to partition compounds belonging to diverse biological activity classes (e.g., different enzyme inhibitors, receptor agonists, or antagonists) with high accuracy.[20] Based on these findings, we hypothesized that such descriptor combinations might also perform well in similarity searching when encoded as short binary fingerprints, which led to the generation of our original MFP prototypes.[8] This conjecture was then confirmed by test calculations where MFPs correctly recognized more than half of all similarity relationships in a database consisting of seven diverse biological activity classes and performed better than other reference fingerprints available to us.[8,17]

Our original MFPs consisted of only 50−60 bit positions and were thus much smaller than other more widely used fingerprints. Thus, an important lesson we learned from these investigations was that increasingly complex fingerprints did not necessarily perform better than simpler designs. Moreover, these studies also indicated that 3D descriptions of molecules were not necessarily required to achieve high accuracy in compound classification similarity searching and, in addition, that structural fragments or keys were particularly powerful descriptors; findings that were well in accord with observations made by others.[21,22] In a number of subsequent investigations in our laboratory, we found that combinations of 1D/2D molecular descriptors and structural keys consistently yielded high prediction accuracy in compound classification.[23,24] Thus, we have thus far generally preferred to combine these types of descriptors when attempting to design new fingerprints.

Although MFPs have first and foremost been research tools for us, we have successfully applied these rather simple designs in the context of drug discovery including the recognition of a number of difficult or remote similarity relationships[25] and the identification of novel hits and leads in virtual screening.[26] These findings have also encouraged us to further refine these types of fingerprints. In particular, we have been interested in the incorporation of new design criteria, as described in the following.

SECOND GENERATION MFPS

Our first attempt to expand on MFP design made use of information content analysis and our adaptation of the Shannon entropy (SE) concept[27] for descriptor selection.[28,29] Rather than selecting descriptors based on their performance in compound classification, we selected descriptors that had consistently high information content in various compound databases and displayed little database-specific differences.[29] These descriptors were then combined with a set of structural keys that occurred in more than 10% and less than 90% in an in-house collection of drug-like molecules.[24] These studies led to the design of what we called SE-MFPs,[30] also shown in Figure 1. These fingerprints consist of approximately 200 bit positions and are thus larger than our original MFPs. In similarity search calculations on 21 activity classes, an SE-MFP prototype performed about 10% better than MFP2 and, as we anticipated, displayed more even performance over different activity classes.[30] Since descriptors for SE-MFP were selected based on information content only, and not

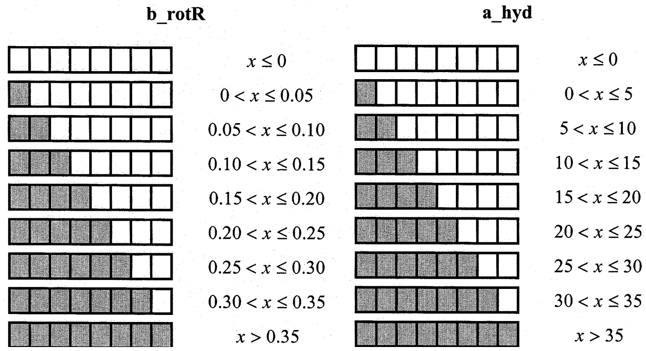

Figure 2. Value range encoding of property descriptors. Two examples are shown, "b_rotR", a descriptor accounting for the fractions of rotatable bonds in a molecule, and "a_hyd", the number of hydrophobic atoms. In each case, eight bits are used to capture value ranges populated in compound databases. Gray shading indicates bits that are set on and illustrates incremental coding of descriptor values.

compound classification, the design successfully eliminated minor bias toward some activity classes used for descriptor selection. Like original MFP designs, SE-MFPs combine value range-encoded property descriptors with bits detecting the presence or absence of structural keys. For example, the SE-MFP version shown in Figure 1 combines 14 property descriptors (using eight bits in each case to capture their value ranges) with 110 structural keys. Such value range encoding of property descriptors is illustrated in Figure 2. This encoding is cumulative in nature (i.e., values are encoded incrementally) and has the potential drawback that some of the bit positions accounting for low values are often set on in database molecules, which tends to increase fingerprint overlap in pairwise comparisons. Thus, indirectly, relatively high weight is given to a number of bit positions that are mostly or always set on.

MP-MFP DESIGN

With the design of MP-MFP, as presented herein, we have gone a step further and attempted to eliminate value range encoding, while retaining—or even enhancing—contributions of property descriptors to our fingerprints. The calculation of medians of descriptor value distributions provided us with an opportunity to do so. In statistics, a median is simply defined as the value that separates a distribution of values in two equally populated halves, above and below the median.[31] We first made use of this principle when developing the median partitioning method, designed to generate median value-based descriptor partitions for diversity selection[18] and virtual screening.[19] Following this approach, n descriptors are used in subsequent steps to divide compound collections into 2^n unique descriptor partitions with respect to their medians. Practically, this only requires the calculation of values of n descriptors for database compounds and their medians. In each step, the compound collection is divided into two subpopulations above and below the median value of a descriptor, and, thereby, twice as many partitions are generated.

Here we have applied this idea to fingerprint design. Thus, the primary goal has been to identify preferred descriptors for median-based encoding in a bit string format. Therefore, we have first calculated values for a previously described pool of more than 150 (1D, 2D, and implicit 3D) property descriptors[24] in a large compound database containing approximately 1.34 million unique molecules (collected from various medicinal chemistry vendor catalogues).[18] These value distributions were then used to calculate the information content of each descriptor (SE values; as described[28,29]) and their medians. Based on SE calculations, 129 descriptors with significant information content were preselected and used for further analysis. In our selection, we aimed to give preference to descriptors with high information content and have avoided the inclusion of strongly correlated descriptors, i.e., descriptors that, directly or indirectly, account for very similar features or effects. To balance these objectives, we calculated a matrix of pairwise correlation coefficients for the preselected 129 descriptors. Based on our previous experience with descriptor correlation effects,[18] a correlation coefficient threshold value of 0.8 was applied to identify strongly correlated pairs. When a descriptor pair with a coefficient above this threshold value was detected, the descriptor with lower information content was removed from the matrix. Once all of these correlations were eliminated, a total of 61 molecular descriptors remained that provided the basis for the generation of MP-MFP. These descriptors and their medians in our source database are reported in Table 1. Approximately half of these descriptors belong to a particular class that map various physicochemical properties on molecular surface areas approximated from 2D representations of molecules.[32] Therefore, these descriptors are best rationalized as implicit 3D descriptors. This finding was not surprising because, as complex formulations, these descriptors are information-rich,[29] and, by design, they are also not correlated to each other.[32] Thus, among the property descriptors for fingerprint design selected here, relatively high weight was given to the evaluation of molecular surface properties. Also, different from our previous fingerprint designs, MP-MFP puts much more weight on 3D information.

Based on these selections, MP-MFP was assembled as follows. For each of the 61 property descriptors, one bit position was assigned, and the median of the descriptor was applied as the cutoff value for the bit to be either set on (i.e., 1) for a value equal or above the median or, alternatively, set off (0) for a value below the median. These 61 bits positions were combined with the set of 111 structural keys that were present in more than 10% and less than 90% of our collection of drug-like molecules,[24] similar to SE-MFP, as described above. Thus, MP-MFP consists of a total of 171 bit positions, as shown in Figure 1. Compared to SE-MFP, it replaces value range encoding of property descriptors by an on/off (or "above/below") format and captures 61 (rather than 14) single descriptors. Calculation of MP-MFP for test molecules is simple. Values of MP-MFP encoded property descriptors are calculated for each molecule and compared to the saved database medians in order to set each of the corresponding bit positions. In addition, the presence (1) or absence (0) of the encoded structural fragments is monitored. The key element of MP-MFP design is that molecular descriptors values with continuous value ranges and measurable information content are transformed into a binary classification scheme.

MP-MFP was generated using SVL code[34] and implemented in the Molecular Operating Environment[35] for similarity searching. All routines required for systematic

Table 1. Molecular Descriptors Selected for MP-MFP[a]

descriptor	median	definition
Topological Descriptors[36−38]		
balabanJ	1.51	Balaban averaged distance sum connectivity
petitjean	0.48	matrix (diameter − radius)/diameter
chi0v_C	12.58	carbon valence connectivity index (order 0)
VDistEq	3.59	vertex distance equality index
Physicochemical Descriptors		
logP(o/w)	0.003	log of the octanol/water partition coefficient
Atom and Bond Counts		
a_acc	4.00	number of H-bond acceptor atoms
a_don	1.00	number of H-bond donor atoms
a_base	1.00	number of basic atoms
a_nBr	1.00	number of Br atoms
a_nCl	1.00	number of Cl atoms
a_nF	1.00	number of F atoms
a_nI	1.00	number of I atoms
a_nN	3.00	number of N atoms
a_nO	3.00	number of O atoms
a_nP	1.00	number of P atoms
a_nS	1.00	number of S atoms
b_1rotN	9.00	number of rotatable single bonds
b_1rotR	0.17	fraction of rotatable single bonds
b_double	2.00	number of double bonds
b_triple	1.00	number of triple bonds
a_ICM	1.63	atom information content (mean)
van der Waals Surface Descriptors		
TPSA	78.46	polar VDW surface area calculated from connection tables
vsa_acid	0.47	VDW surface areas of acidic atoms
vsa_base	0.36	VDW surface areas of basic atoms
vsa_don	5.68	VDW donor surface area
vsa_pol	0.85	VDW polar surface area
vsa_other	44.01	VDW other surface area
Partial Charge Descriptors[39]		
PEOE_RPC-	0.18	relative negative partial charge
PEOE_VSA_FHYD	0.84	fractional hydrophobic vdw surface area
PEOE_VSA_FNEG	0.50	fractional negative vdw surface area
Complex Surface Area Descriptors[32]		
PEOE_VSA+0	68.28	sum of v_i where q_i is in the range [0.00,0.05)
PEOE_VSA+1	66.34	sum of v_i where q_i is in the range [0.05,0.10)
PEOE_VSA+2	17.24	sum of v_i where q_i is in the range [0.10,0.15)
PEOE_VSA+3	1.82	sum of v_i where q_i is in the range [0.15,0.20)
PEOE_VSA+4	12.95	sum of v_i where q_i is in the range [0.20,0.25)
PEOE_VSA+5	2.01	sum of v_i where q_i is in the range [0.25,0.30)
PEOE_VSA+6	1.69	sum of v_i where q_i is greater than 0.3
PEOE_VSA-0	73.53	sum of v_i where q_i is in the range [−0.05,0.00)
PEOE_VSA-1	69.36	sum of v_i where q_i is in the range [−0.10,−0.05)
PEOE_VSA-2	0.90	sum of v_i where q_i is in the range [−0.15,−0.10)
PEOE_VSA-3	1.39	sum of v_i where q_i is in the range [−0.20,−0.15)
PEOE_VSA-4	0.14	sum of v_i where q_i is in the range [−0.25,−0.20)
PEOE_VSA-5	13.70	sum of v_i where q_i is in the range [−0.30,−0.25)
PEOE_VSA-6	5.01	sum of v_i where q_i is less than −0.30
SMR_VSA0	47.72	sum of v_i such that R_i is in [0,0.11]
SMR_VSA1	27.79	sum of v_i such that R_i is in (0.11,0.26]
SMR_VSA2	18.01	sum of v_i such that R_i is in (0.26,0.35]
SMR_VSA3	14.80	sum of v_i such that R_i is in (0.35,0.39]
SMR_VSA4	5.51	sum of v_i such that R_i is in (0.39,0.44]
SMR_VSA5	174.70	sum of v_i such that R_i is in (0.44,0.485]
SMR_VSA6	26.03	sum of v_i such that R_i is in (0.485,0.56]
SlogP_VSA0	22.00	sum of v_i such that $L_i < = -0.4$
SlogP_VSA1	26.85	sum of v_i such that L_i is in (−0.4,−0.2]
SlogP_VSA2	23.86	sum of v_i such that L_i is in (−0.2,0]
SlogP_VSA3	20.93	sum of v_i such that L_i is in (0,0.1]
SlogP_VSA4	9.56	sum of v_i such that L_i is in (0.1,0.15]
SlogP_VSA5	18.87	sum of v_i such that L_i is in (0.15,0.20]
SlogP_VSA6	0.39	sum of v_i such that L_i is in (0.20,0.25]
SlogP_VSA7	141.10	sum of v_i such that L_i is in (0.25,0.30]
SlogP_VSA8	17.64	sum of v_i such that L_i is in (0.30,0.40]
SlogP_VSA9	84.79	sum of v_i such that $L_i > 0.40$

[a] Here v_i is the van der Waals surface area of atom *i*, and q_i represents the partial charge of atom *i*.[39] L_i is the contribution to logP(o/w) of atom *i*,[40] and R_i is the contribution to molar refractivity of atom *i*.[40] For each descriptor, the database median value is reported.

performance evaluation, as described below, were also written in SVL.

SIMILARITY ASSESSMENT

Based on its design, calculation of the conventional Tanimoto coefficient (Tc)[1] was not sufficient for MP-MFP comparisons, for the following reasons. The Tanimoto coefficient is defined as

$$Tc = bc/(b1 + b2 - bc)$$

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. Thus, Tc only takes into account bit positions that are set on (1). However, in MP-MFP, property descriptor bit positions that are on or off capture equivalent information (a value above or below the reference median). Thus, simply put, bit positions set to 0 must be taken into account as much as position set to 1 when fingerprint overlap is calculated. Therefore, we devised a modification of Tc that we call averaged Tc (avTc) in order to address these requirements

$$avTc = (Tc + Tc')/2$$

When calculating Tc′, bit positions set to 0 are counted, rather than those set to 1. The averaged Tc was calculated for MP-MFP although it also contains conventional bit positions assigned to structural keys (essentially implying that the absence of structural fragments, and not only their presence, is an indication of molecular similarity).

PERFORMANCE EVALUATION

It was difficult to predict how MP-MFP would perform in similarity searching, given its novel bit organization and the modified similarity metric required for comparison. It remained to be determined whether the binary transformation of continuous descriptor value ranges would not have too low resolution to produce reasonable search results (even when combined with structural keys). On the other hand, the design of MP-MFP permitted the incorporation of a significant number of different property descriptors into a short fingerprint, which could compensate for low resolution at single bit positions. Clearly, we needed a suitable test case in order to answer these questions and evaluate MP-MFP performance relative to other fingerprints.

Therefore, we assembled a database consisting of 549 molecules belonging to 38 different activity classes, as reported in Table 2. The number of compounds per activity class ranged from 10 to 22. Some of these classes were collected from the literature as described,[23] and others were taken from the Synthline database.[35] To further increase the difficulty of the search problem, 5000 compounds were randomly collected from our in-house source database and added as “background” molecules (i.e., no activity assignment), yielding a total of 5549 database compounds.

To assess overall fingerprint performance, the following scoring function was applied[8]

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

with C and I being the number of correctly and incorrectly identified compounds per activity class, respectively, and N the total number of compounds in each class. If C was smaller than I, the score was set to zero. Scores were individually calculated for compounds in each activity class and then averaged over all 38 classes.

Table 2. Biological Activity Classes

class designation	activity	no. of compds
COX	cyclooxygenase-2 (Cox-2) inhibitors	17
TKE	tyrosine kinase inhibitors	20
HIV	HIV protease inhibitors	18
H3E	H3 antagonists	21
BEN	benzodiazepine receptor ligands	22
5HT	serotonin receptor ligands (5-HT)	21
CAE	carbonic anhydrase II inhibitors	22
BLC	β-lactamase inhibitors	14
PKC	protein kinase C inhibitors	15
ESTR	estrogen antagonists	11
ACE	antihypertensive (ACE inhibitor)	17
ADR	antiadrenergic (β-receptor)	16
GLU	glucocorticoid analogues	14
ANG	angiotensin AT1 antagonists	10
ARO	aromatase inhibitors	10
TOP	DNA topoisomerase I inhibitors	10
DIH	dihydrofolate reductase inhibitors	11
FAC	factor Xa inhibitors	14
FAR	farnesyl transferase inhibitors	10
MAT	matrix metalloproteinase inhibitors	12
VIT	vitamin D analogues	12
THRI	thrombin inhibitors	15
RTI	reverse transcriptase inhibitors	15
PPAR	PPARgamma agonists	16
PDE4	phosphodiesterase IV inhibitors	11
PAFA	PAF antagonists	10
NEPI	neprilysin inhibitors	12
NAHE	Na+/H+ exchange inhibitors	12
HH2A	histamine H2 antagonists	13
FIBA	fibrinogen gpIIb/IIIa antagonists	17
ENDA	endothelin ETA antagonists	15
DD2A	dopamine D2 antagonists	14
DOA	delta opioid agonists	10
CRF1	CRF1 antagonists	12
CCR5	CCR5 antagonists	16
CALA	calcium antagonists	18
ARI	aldose reductase inhibitors	12
HT1D	5 HT1D agonists	14

Each of the 549 active molecules in our test database was searched against the remaining compounds over the entire Tc or avTc range [0,1] in 0.01 increments, thus requiring a total of 101 calculations per test molecule. In these calculations, detected background molecules were considered false positives. In each case, search results yielding the highest score and their corresponding Tc or avTc value were saved for averaging. This made it also possible to determine optimum Tc threshold values for each of the following trial fingerprints: MP-MFP, MFP2, SE-MFP, and MACCS. As an additional reference, we also evaluated MP[61], a subset of MP-MFP only containing the 61 binary encoded property descriptors. For MP-MFP and MP[61], avTc values were calculated and conventional Tc values for the other fingerprints. The results are summarized in Table 3.

The tested fingerprints correctly recognized between ∼22% and ∼34% of all similarity relationships and produced low false positive rates of less than 0.1%. MACCS, one of the gold standards in the field, produced on average 26% correct hits. However, the performance of the small MFP2 was only about 4% lower. MP[61] is of comparably small size but exclusively consists of binary encoded property descriptors and performed slightly better than MFP2, con-

Table 3. Performance Evaluation[a]

fingerprint	score	Tc/avTc	correct (%)	incorrect (%)
MACCS	0.21	0.77	26.39	0.02
MFP2	0.16	0.86	22.69	0.06
SE-MFP	0.23	0.81	28.29	0.02
MP-MFP	0.25	0.72	34.31	0.04
MP[61]	0.17	0.76	24.22	0.04

[a] Results are reported for systematic similarity search calculations in our test database consisting of 549 molecules in 38 activity classes plus 5000 background molecules. Tc or avTc (for MP-MFP and MP[61]) cutoff values that yielded the highest scores are shown. "Correct" reports the percentage of compounds having similar activity that were correctly identified, and "incorrect" reports the percentage of false positive recognitions including background molecules.

firming the predictive ability of these binary descriptor settings. SE-MFP, containing 14 value range encoded descriptors, performed about 2% better than MACCS, at a similar Tc optimum. However, the best performing fingerprint in our test calculations was MP-MFP that outperformed MACCS by about 8% and correctly recognized ~34% of all active compounds with 0.04% false positives, at an optimum avTc of 0.72. Comparison with MP[61] showed that the addition of selected structural keys increased overall search performance by about 10%, indicating that the combination of binary encoded property and structural descriptors was a preferred design.

When we monitored class-specific performance of the fingerprints during our calculations, some trends could be observed. In a number of cases, most fingerprints performed well, whereas in some others, none of the fingerprints was capable of recognizing a significant number of activity relationships. These observations illustrated that evaluation of similarity search tools depends at least to some extent on the particular test cases studied, which emphasizes the need for as many different tests as possible. Exhaustive search calculations over 38 diverse biological activity classes in the presence of a large excess of randomly selected background molecules were a rather challenging test case, and, from this point of view, we considered the results encouraging, in particular, for MP-MFP.

CONCLUDING REMARKS

In this study, we have introduced binary encoding of molecular property descriptors as an approach to fingerprint design by applying the simple concept of statistical medians. Following our MFP design idea, a hybrid fingerprint containing binary encoded property and structural descriptors was constructed and compared to previously generated fingerprints. An attractive feature of simple fingerprint prototypes such as SE-MFP or MP-MFP is that they can easily be modified to incorporate additional descriptors or change descriptor combinations. Also, bit settings in these fingerprints are straightforward to interpret. Similar to descriptor selection based on information content, median-based encoding has the advantage that no biologically active compounds are needed to select suitable descriptors or train fingerprints, thus avoiding bias toward certain activity classes. Replacing cumulative value range encoding of property descriptors by binary encoding has at least three advantages. First, it avoids the situation that some bit positions are mostly set on, which potentially biases Tc calculations. Second, bits that are set off also contribute to similarity assessment. Third, many more descriptor contributions can be embedded in fingerprints of small to moderate size, without substantially increasing the complexity of their design. Both information content- and median-based descriptor selections for similarity searching rely on the statistical relevance of property distributions in large compound databases and, in addition, on the assumption that these distributions contain sufficient information for the detection of molecular similarity. The results reported in this study support the validity of this concept. Median-based descriptor settings and MP-MFP design have demonstrated that combinations of binary low resolution descriptors are discriminatory and capable of correlating molecular properties with biological activity. In combination with structural descriptors, these two-state property descriptors yield effective similarity search tools, as revealed by our systematic test calculations. It is also hoped that MP-MFP and related fingerprint designs will be effective in drug discovery situations, as has been the case for other MFP designs.[26,41]

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. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(3) 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.

(4) Flower, D. R. On the properties of bit string based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(5) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.

(6) 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.

(7) MACCS keys. MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA.

(8) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.

(9) James, C. A.; Weininger, D. Daylight theory manual, Daylight Chemical Information Systems, Inc., Irvine, CA.

(10) UNITY. Chemical Information Software, Tripos, Inc., St. Louis, MO.

(11) BCI. Barnard Chemical Information Ltd., Sheffield, UK.

(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) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: Description and application to α_1-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(14) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(15) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.

(16) *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(17) Xue, L.; Godden, J.; Bajorath, J. Evaluation of descriptors and minifingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.

(18) 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.

(19) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.

(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) 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.

(22) 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.

(23) 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.

(24) Xue, L.; Bajorath, J. Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757−764.

(25) 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.

(26) Stahura, F. L.; Xue, L.; Godden; J. W.; Bajorath, J. Methods for compound selection focused on hits and application in drug discovery. *J. Mol. Graph. Model.* **2002**, *20*, 439−446.

(27) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1963.

(28) 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.

(29) 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.

(30) 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.

(31) Meier, P. C.; Zünd, R. E. *Statistical methods in analytical chemistry;* John Wiley & Sons: New York, 2000.

(32) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model*. **2000**, *18*, 464−477.

(33) Sanatvy, M.; Labute, P. SVL: The scientific vector language. *J. Chem. Computing Group* URL: www.chemcomp.com/feature/svl.htm.

(34) Molecular Operating Environment (MOE), version 2001.01, Chemical Computing Group Inc., Montreal, Quebec, Canada.

(35) *Synthline Drug Database (from Drugs of the Future)*; Prous Science: Barcelona, Spain.

(36) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−422.

(37) Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

(38) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological indices: their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.

(39) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(40) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(41) A referee suggested that we comment on the “real world” performance of these fingerprints. MFP-MP has not yet been tested in drug discovery projects. However, MFP-type designs have identified novel hits against six of eight (proprietary) drug targets that have recently been subjected to virtual screening in our laboratory. A similar case study is presented in ref 26.

CI030285+