Molecular Similarity Analysis and Virtual Screening by Mapping of Consensus Positions in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality

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

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI),
AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011,
21 Corporate Circle, Albany, New York 12212-5098, and Department of Biological Structure,
University of Washington, Seattle, Washington 98195

Received August 15, 2003

A novel compound classification algorithm is described that operates in binary molecular descriptor spaces and groups active compounds together in a computationally highly efficient manner. The method involves the transformation of continuous descriptor value ranges into a binary format, subsequent definition of simplified descriptor spaces, identification of consensus positions of specific compound sets in these spaces, and iterative adjustments of the dimensionality of the descriptor spaces in order to discriminate compounds sharing similar activity from others. We term this approach Dynamic Mapping of Consensus positions (DMC) because the definition of reference spaces is tuned toward specific compound classes and their dimensionality is increased as the analysis proceeds. When applied to virtual screening, sets of bait compounds are added to a large screening database to identify hidden active molecules. In these calculations, molecules that map to consensus positions after elimination of most of the database compounds are considered hit candidates. In a benchmark study on five biological activity classes, hits for randomly assembled sets of bait molecules were correctly identified in 95% of virtual screening calculations in a source database containing more than 1.3 million molecules, thus providing a measure of the sensitivity of the DMC technique.

INTRODUCTION

The concept of molecular similarity has greatly influenced the development of compound classification and database mining methods in pharmaceutical research.[1] The ability to evaluate the degree of molecular similarity beyond structural equivalence or correspondence has provided a basis for the correlation of molecular properties and biological activity of test compounds.[1,2] Such comparisons and correlations have by and large been made possible through the definition of chemical reference spaces that in turn depend on the use of various chemical descriptors to capture structural features and molecular properties.[2,3] Thus, rather than carrying out direct molecular comparisons, by eye or computer, metrics have been developed that compare molecules in theoretical chemistry spaces, which ultimately makes it possible to analyze very large numbers of compounds. Essentially, all of the approaches that are currently used to assess molecular similarity on a large scale make use of descriptors to define chemical spaces.[2] Among others, these methodologies include hierarchical[4] and nonhierarchical[5] clustering techniques, cell-based[6,7] and statistical[8,9] partitioning, multidimensional[10,11] or binary[12] QSAR models, and bit string methods such as 2D molecular fingerprints[13,14] or 3D pharmacophore fingerprints.[15,16]

The definition of chemical reference spaces often differs significantly dependent on the specific requirements and characteristics of each methodology. For example, search spaces defined by fingerprints might not only be very different from clustering or partitioning spaces, but also from each other, ranging from narrowly defined spaces consisting of only a few descriptors[13] to millions of possible pharmacophore patterns.[16] Chemical space transformation and dimension reduction methods[17−21] play an important role in many cases, for example, partitioning algorithms, the successful application of which often—but not always—depends on the ability to generate low-dimensional or simplified space representations.[22]

For molecular classification or similarity search calculations, a key question is how to find preferred descriptors for the generation of suitable chemical references spaces. This question is a priori difficult to answer and, consequently, in addition to chemical intuition, machine learning techniques such as genetic algorithms[18,23] or neural nets[24] and information theoretic approaches[25] have been employed to automate and/or rationalize descriptor selection. However, once a descriptor reference space has been defined for a specific application, it is usually kept fixed, and there is little opportunity to modify this reference space in the course of the analysis.

Molecular similarity assessment, as discussed herein, also provides the methodological basis for virtual screening calculations[2,26] that utilize compounds with desired activity as templates (as opposed to known 3D structures of protein

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.

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

[‡] 21 Corporate Circle.

[§] University of Washington.

10.1021/ci0302963 CCC: $27.50 © 2004 American Chemical Society
Published on Web 11/08/2003

targets). Small molecule-based virtual screening is critically dependent on the ability to successfully correlate biological activity with various (and often complex) molecular properties, since the major challenge is here to identify diverse structural motifs displaying similar activity.[26] Another challenge for virtual screening is the ability to process increasingly large source databases often consisting of millions of molecules. Thus, molecular similarity calculations suitable for virtual screening must be computationally efficient, yet sensitive enough to find novel hits.

Here we introduce a novel molecular similarity method that is suitable for large-scale virtual screening. The approach relies on finding consensus positions for sets of active compounds in simplified yet high-dimensional descriptor spaces. It separates active from inactive compounds by redefining consensus positions in chemical spaces of increasing dimensionality. Simplified descriptor spaces are generated by converting descriptors with continuous value ranges into a binary format. This is accomplished by calculation of statistical medians for descriptor distributions in compound source databases. With regard to its practical requirements, DMC is much simpler than many other small molecule-based virtual screening techniques and does not depend on the application of machine learning techniques for variable selection. This flexible mapping technique consistently produced hits for different bioactivity classes in virtual screening calculations. It is anticipated that this methodology can be readily applied for hit identification to many therapeutically relevant compound classes.

METHODS

Descriptor Pool. For our current analysis, a previously described[9] set of ~130 1D, 2D, and implicit 3D molecular property descriptors with continuous value ranges was used. Values of these descriptors in our compound source database (see below) were calculated with MOE,[27] and descriptors that had little or no information content in this database, as determined by entropy calculations,[25] were eliminated. Thus, a total of 104 descriptors formed the pool for DMC calculations. Results of compound classification and virtual screening calculations are in general affected by the composition of descriptor sets used. However, in our experience, results of DMC calculations and overall performance are little influenced by the addition of more descriptors to the relatively extensive basis set used here.

Activity Classes and Database Compounds. As our source database, an in-house compound collection was used that contained ~1.34 million molecules collected from various medicinal chemistry vendor catalogues.[9] As test cases for DMC analysis, five therapeutically relevant activity classes were selected, as described previously:[28] benzodiazepines (abbreviated BEN; 59 compounds in total), cyclooxygenase-2 inhibitors (COX; 31 compounds), histamine H3 antagonists (H3; 52 compounds), serotonin receptor ligands (SER; 71 compounds), and tyrosine kinase inhibitors (TK; 35 compounds). Figure 1 shows representative examples of active compounds highlighting their structural diversity (even within activity classes).

Binary Descriptor Space Transformation. Continuous value ranges of descriptors in our pool were converted into a binary format by calculation of statistical medians[29] of their database distributions. The median is defined as the value that divides a population of values into two halves above and below the median. We first made use of this concept when introducing the median partitioning algorithm for diversity selection.[9] After calculating medians for all descriptors in our source databases, each compound was assigned a one for a specific descriptor if its value was larger or equal to the median or a zero if it was smaller than the median. Based on this formalism, selection of n descriptors defines an n-dimensional reference space and each database compound is assigned an n-dimensional vector, represented as a bit string, where the value for each dimension is one or zero. Thus, dimension reduction is not involved it the generation of the chemical spaces, but compound representation along each dimension is simplified by application of a binary model.

Mapping of Consensus Positions. For a compound set of interest, for example, a class of active compounds, a consensus position in chemical space can be defined by identifying descriptors having identical bit settings for all compounds of the set. We initially determine the maximum number of our 104 descriptors having identical bit settings, thereby defining a descriptor space where all compounds in the set map to exactly the same position. Since this procedure is a summation over a descriptor-by-compound matrix, it is extremely fast. Excluding the initial one-time calculation of descriptor values for molecules in the source database, mapping of database compounds to consensus positions requires only minutes on a standard PC even if the source database is very large. Essentially, the only time limiting step is the speed of disk access. Programs for mapping of consensus positions and dimension extension, as described below, were written in C.

Relaxation of Consensus Criteria and Dimension Extension. The specific position in chemical space that is occupied by our set of active compounds might also be shared by many other database compounds, dependent on the dimensionality and "resolution" of the initially defined reference space. Since our goal is to find only those database compounds that are most similar to our activity set, the dimensionality of the descriptor space is increased, in subsequent steps, to distinguish more and more database compounds from our active molecules. This is achieved by defining consensus positions that no longer require identical descriptor settings for all test compounds. Thus, one or more compounds in our activity set are allowed to deviate in one or more descriptor settings from the rest of the compounds. We therefore calculate complete descriptor bit strings for all compounds in our class and divide the sum of bits set to one by the total number of compounds. Following our nomenclature, the first dimension extension is achieved for bit ratios "$<=$ 0.1 or $>=$0.9", the second extension for "$<=$0.2 or $>=$0.8" and so on. This means that the first dimension extension permits 10% variability in binary descriptor settings and the second step 20%. For example, if our activity class includes 10 compounds, category "$<=$0.1 or $>=$0.9" means that additional descriptors will be accepted if one compound has a different bit setting than the other nine, either a zero versus nine ones or vice versa. Accordingly, category "$<=$0.2 or $>=$0.8" means that additional descriptors will be accepted if bit settings for two

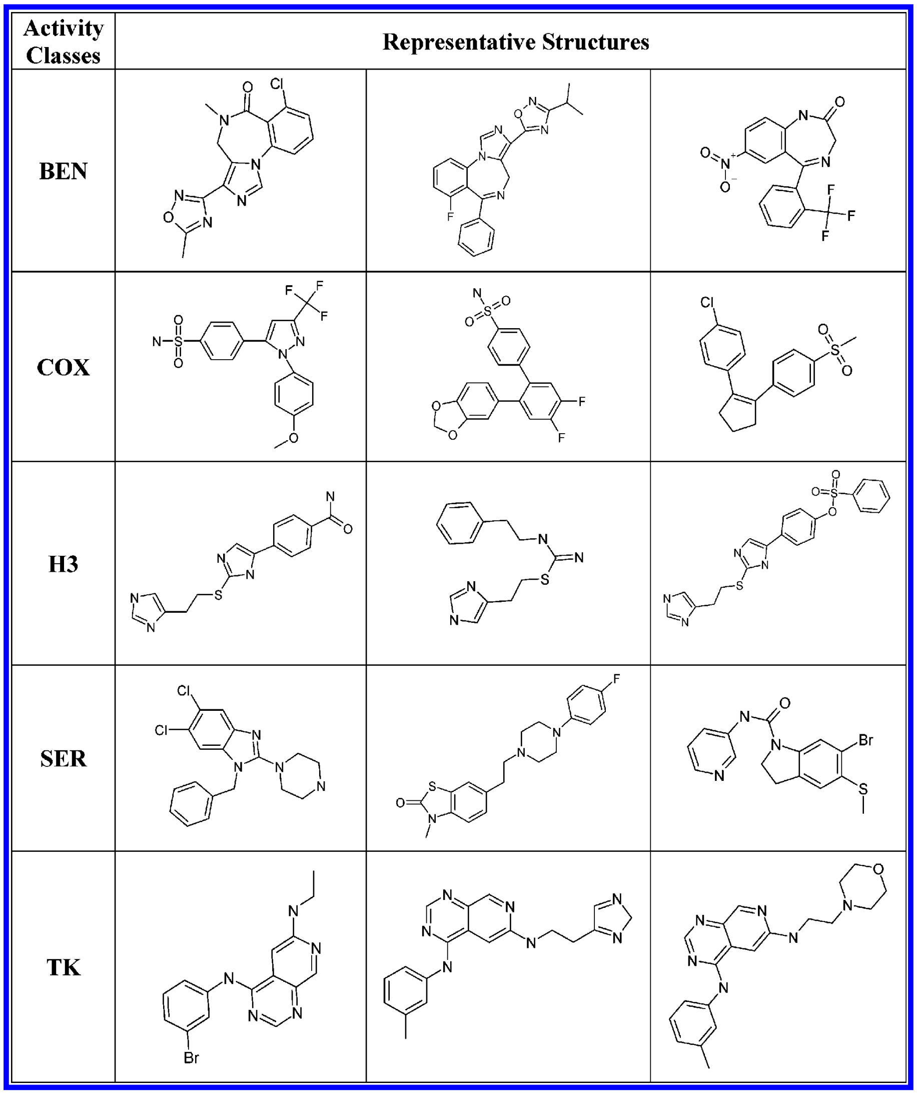

Figure 1. Activity classes. Representative examples of compounds belonging to different activity classes are shown.

of 10 compounds differ from the rest. For cases “>=0.9” and “>= 0.8”, the descriptor consensus at this position is set to one, whereas it is set to zero for “<=0.1” and “<=0.2”. Figure 2 illustrates the principle of dimension extension, and Figure 3 provides a summary of DMC calculations.

Virtual Screening Calculations. To benchmark DMC methodology for virtual screening, activity classes were randomly divided into a “bait” set and active database compounds (potential hits) that were “hidden” in the source databases. DMC analysis was carried out for the bait compounds, and the number of database compounds sharing consensus positions was reduced in subsequent steps of dimension extension until only a small number was retained. It was then determined whether these molecules contained active database compounds (correctly identified hits). “Hit rates” were calculated by dividing the number of active database compounds by the total number of selected database molecules, and “recovery rates” by dividing the number of correctly identified hits by the total number of active database compounds (potential hits). To assess the effects of random separation of bait and active database compounds and take structural differences within an activity class into account, DMC calculations for each activity class were carried out using 20 different random sets of baits and corresponding active database compounds.

RESULTS AND DISCUSSION

Concept and Characteristic Features of DMC. We set out to develop a methodology to assign molecules having similar activity to consensus positions in chemical space and identify database compounds that closely map to these positions. The underlying idea is that equivalent positions in chemical space correlate with molecular similarity. Rather than defining a fixed chemical space framework, we intended to specifically design chemical reference spaces for given sets of bioactive compounds and, in addition, provide the flexibility to modify and extend these chemical spaces during similarity analysis. This was thought to increase the sensitivity of descriptor combinations for the correlation of molecular properties and specific activity. In addition, we wanted to ensure that the algorithm could be applied to efficiently process very large compound collections in accord with requirements of virtual screening. Therefore, we focused on the definition of simplified and easy to calculate chemical space representations, which was facilitated by applying the concept of statistical medians and transforming molecular descriptors into a binary classification scheme. Combinations

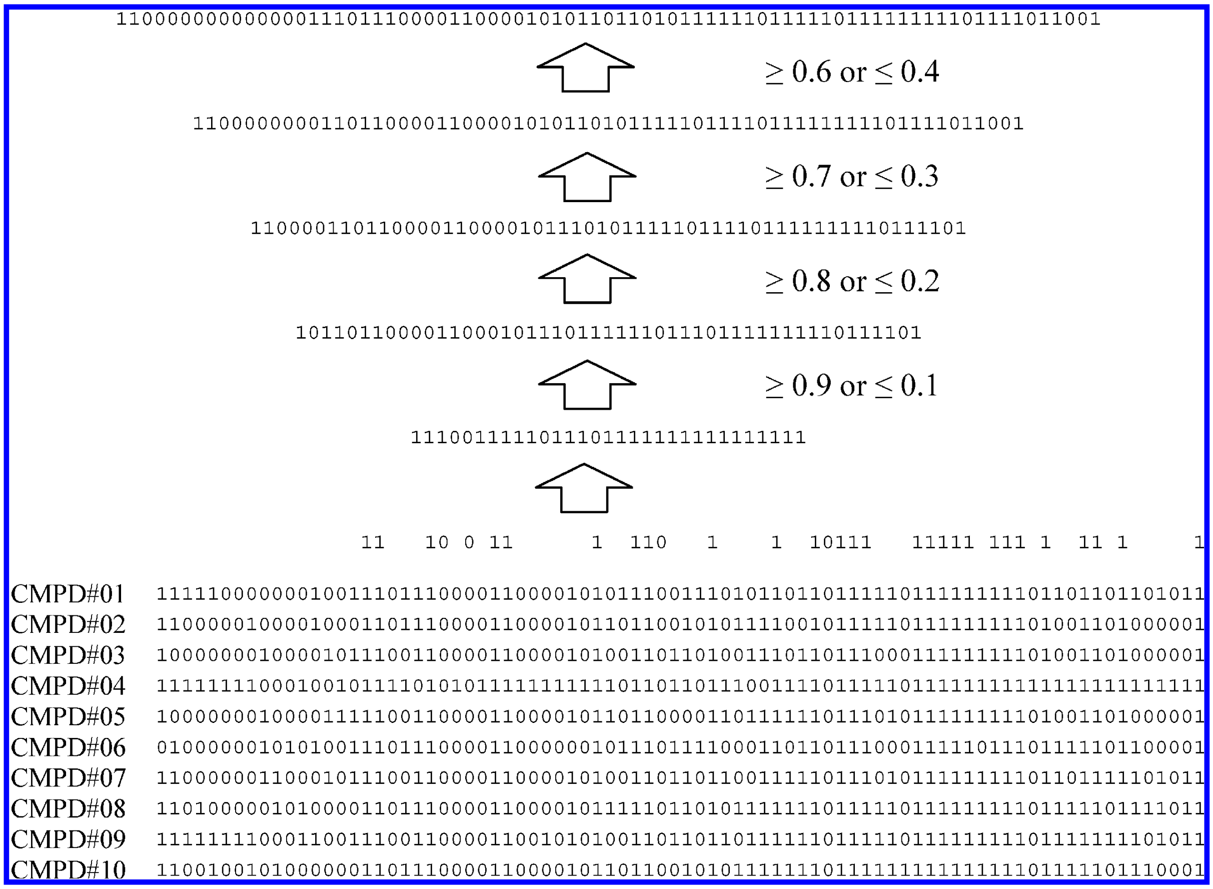
1100000000000001110111000011000010101101101011111011111011111111101111011001
≥ 0.6 or ≤ 0.4
110000000011011000011000010101101011111011110111111111101111011001
≥ 0.7 or ≤ 0.3
110000110110000110000101110101111101111011111111110111101
≥ 0.8 or ≤ 0.2
10110110000110001011101111110111011111111110111101
≥ 0.9 or ≤ 0.1
1110011111011101111111111111111
11 10 0 11 1 110 1 1 10111 11111 111 1 11 1 1
CMPD#01 111110000000100111011100001100001010111001110101101101111101111111110110110101011
CMPD#02 110000010000100011011100001100001011011001010111100101111101111111110100110100001
CMPD#03 100000001000010111001100001100001010011011010011101101110001111111110100110100001
CMPD#04 111111110001001011110101011111111111011011011100111011110111111111111111111111
CMPD#05 100000001000011111001100001100001011011000011011111101110101111111110100110100001
CMPD#06 010000001010100111011100001100000010111011110001101101110001111011101111101100001
CMPD#07 110000001100010111001100001100001010011011011001111011101011111111110110111101011
CMPD#08 110100000101000011011100001100001011111011010111111101111011111111110111110111011
CMPD#09 111111110001100111001100001100101010011011011011111101111011111111110111111101011
CMPD#10 110010010100000011011100001100001011011001010111111101111111111111110111110111000 1

Figure 2. Dimension extension in DMC. DMC analysis is illustrated for 10 compounds belonging to the same activity class. Initially, the maximum number in a large pool of binary-transformed descriptors is determined that map each compound to exactly the same position. This descriptor combination defines the lowest-dimensional reference space for the analysis. During subsequent mapping steps, the dimensionality of the compound class-dependent reference space is increased in order to discriminate more and more database compounds from active molecules. This can be rationalized as creating a finer grid for similarity evaluation. The increase in dimensionality is achieved by defining new consensus positions that do no longer require identical descriptor coordinates for each active compound, as described in the text. Thus, for each descriptor, one or more active compounds are allowed to deviate from the position of the remaining compounds, which increases the number of descriptors that fulfill the consensus criterion.

of binary transformed descriptors spaces can be represented as bit strings, and we have previously demonstrated that so derived bit string representations can have significant predictive power in conventional similarity searching.[30] Therefore, the approach we call dynamic mapping of consensus positions (or DMC) combines elements of bit string and partitioning techniques, as each signature bit string represents a unique segment of chemical space. Key aspects of the methodology are the generation of compound class-dependent and dimensionally variable chemical spaces for database analysis, as illustrated in Figure 2.

Mapping of Compound Class-Specific Consensus Positions. Table 1 summarizes the results of mapping calculations for 31 to 71 compounds belonging to five activity classes when added to ~1.3 million database compounds. The number of descriptors producing identical bit patterns for all compounds in a class ranges from nine for SER to 38 for COX. Thus, different compound sets produce different chemical space solutions. In each case, the initially determined number of identically set descriptors constitutes the lowest-dimensional chemistry space for mapping of consensus positions. At this level of resolution, the number of other database compounds sharing consensus positions ranges from 1041 for COX to 91 151, illustrating that a finer descriptor grid must be applied for more stringent molecular similarity evaluation.

Dimension Extension. A higher resolution grid is obtained by dimension extension. Table 1 also reports the effects of dimension extension. During this process compounds sharing similar activity stay close to each other, whereas database compounds that differ from them move further away. A first round of dimension extension (category "<=0.1 or >=0.9"; see Methods) already leads to a significant increase in the number of accepted descriptors (ranging from a factor of 1.3 for COX to 5.2 for SER) and, at the same time, to a significant decrease in the number of database compounds sharing consensus positions of activity classes (ranging from a factor of 3.4 for BEN to 21.9 for H3). This decrease in database compound numbers was equally, if not more significant for the second step. The statistics also revealed that subsets of compound activity classes can produce rather specific descriptor settings. For example, SER compounds initially shared only nine (of 104) identical descriptor settings. However, permitting 10% bit variability during the first dimension extension (i.e., for seven of 71 compounds), increased the number of consensus descriptors to 47. Overall the data show that definition of consensus positions in higher-dimensional spaces indeed distinguished an increasing amount of database compounds from sets of active molecules. In other words, so defined consensus positions became increasingly compound class-specific. The five activity classes also displayed different descriptor prefer-

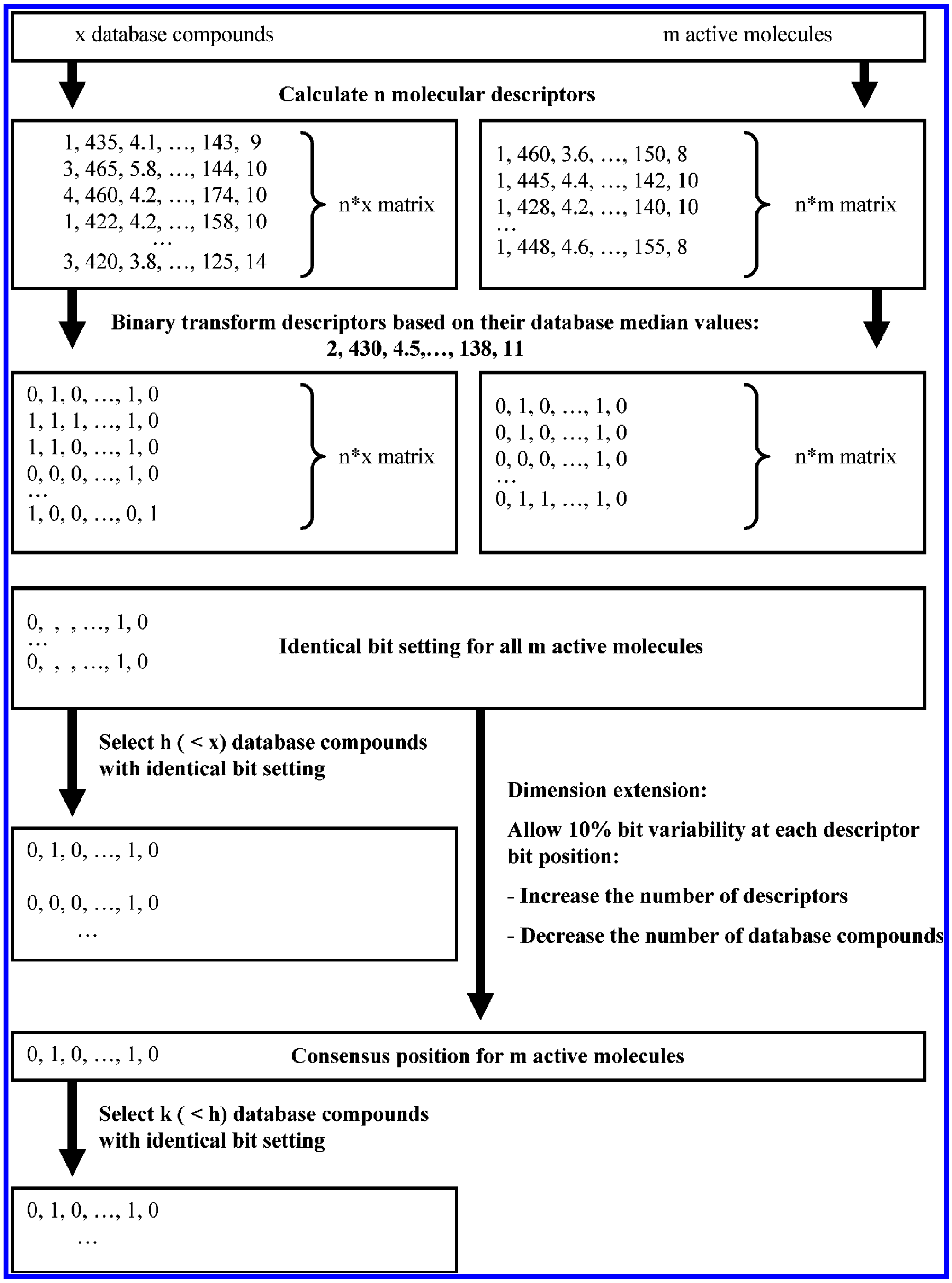

Figure 3. DMC scheme. A summary of DMC calculations is shown, as described in the text. Major steps of this process include initial binary transformation of descriptors, identification of consensus positions for activity classes in binary descriptor spaces, mapping of database compounds to consensus positions, and dimension extension of descriptor spaces.

ences. For example, at extension levels where fewer than 100 database compounds mapped to consensus positions (from "<=0.1 or >=0.9" for COX and H3 to "<=0.3 or >=0.7"; Table 1), between 55 and 71 descriptors defined consensus positions, but only five descriptors were shared by all activity classes.

When 30% bit variability was permitted for definition of consensus positions ("<=0.3 or >=0.7"), between 71 and 87 (of 104) descriptors were included in these calculations, and all or almost all of ~1.3 million database compounds did no longer match any consensus position. This illustrates that, in this case, consensus bit patterns for activity classes became essentially impossible to match, if their definition permitted 30% bit variability or more (and thus included a large number of descriptors). It follows that the most reasonable dimensionality range for the test cases studied here was achieved by one or two steps of dimension extension of the originally defined descriptor spaces. Smaller extension intervals (e.g., "<=0.2 or >=0.8" followed by "<=0.25 or >=0.75") could also be defined. As shown below, however, for our virtual screening applications, the current extension steps were very appropriate (in particular, since the bait sets were smaller than the initially mapped activity classes). Taken together, the results confirmed that consensus positions could be successfully determined for bioactivity classes at increasing levels of chemical resolution

Table 1. Mapping of Consensus Positions[a]

	≤0.0 or ≥1.0	≤0.1 or ≥0.9	≤0.2 or ≥0.8	≤0.3 or ≥0.7	≤0.4 or ≥0.6
		BEN: 59 Compounds			
Consensus DS	22	38	56	71	85
database compounds	91151	26617	9627	35	0
		COX: 31 Compounds			
Consensus DS	38	55	77	85	95
database compounds	1041	84	1	0	0
		H3: 52 Compounds			
Consensus DS	22	61	69	82	92
database compounds	1252	57	2	0	0
		SER: 71 Compounds			
Consensus DS	9	47	70	87	95
database compounds	28441	2278	63	0	0
		TK: 35 Compounds			
Consensus DS	20	48	69	83	94
database compounds	4106	343	33	5	0

[a] The table summarizes the mapping statistics for identifying consensus positions for all five activity classes. At the beginning, each compound within a class occupied exactly the same position. During dimension extension, consensus criteria were relaxed as described in the text. “Consensus DS” reports the total number of descriptors (from a pool of 104) that defined consensus positions at different mapping levels and “database compounds” the number of other source database molecules that mapped to the same consensus position. Consistent with our nomenclature, “≤0.0 or ≥1.0” refers to the initial mapping of consensus positions prior to dimension extension.

and used to evaluate the similarity to other database compounds; the underlying principle of DMC analysis.

Virtual Screening Analysis. For the application of DMC to virtual screening, bait sets consisting of 10 molecules were randomly selected from the five bioactivity classes, and the remaining molecules were added to the large source database. In this case, compound subsets available for identification of consensus positions were much smaller than the entire activity class, and differences in their molecular composition were expected to influence the definition of reference spaces and mapping of compounds to consensus positions. Therefore, we randomly selected 20 different bait subsets for each class (which in turn also changed the composition of the sets of active database compounds). We attempted to simulate practical virtual screening applications whereby the large number of database compounds had to be reduced until a sufficiently small number of putatively similar candidate compounds were obtained (for testing), for example, approximately 100 or fewer. During subsequent extension and mapping steps, it was not necessary to remove database compounds from our calculation but only to update their descriptor bit signatures. The results of 100 virtual screening runs are summarized in Table 2.

Consensus Positions. The results reported in Table 2 confirmed that different bait sets produced different descriptor solutions, consensus positions, and similarity matches. This was expected because single bit differences between compounds within bait sets initially exclude descriptors from the definition of consensus positions. Also, relatively small differences in descriptor values close to medians can change binary settings in a series of compounds (boundary effect). However, within each activity class, there was a distinct trend to select similar descriptor subsets, irrespective of bait set composition. At every extension level, at least half of the descriptors, and often ~75%, were in common to different runs. The general tendency of the bait sets to match larger numbers of descriptors than the entire activity class can be readily explained by a decreased probability of bit mismatches in the smaller bait sets.

Search Performance. Overall, the obtained search results were encouraging. Virtual screening calculations succeeded for all five activity classes, and it was generally possible to sufficiently reduce the number of database compounds during only one or two rounds of dimension extension. In 95 of 100 calculations over five activity classes a number of hits were correctly identified. In some cases such as COX, initially defined descriptor spaces were already sufficiently predictive to produce desired virtual screening outcomes (for example, 17 candidate compounds including five hits or 53 candidates containing 8 hits). Reducing the number of database compounds was overall most difficult for benzodiazepines where five of 20 independent calculations failed. However, even in this case, three other calculations produced only correct hits as candidates (corresponding to 100% hit rate). In general, best results were obtained for H3 after dimension extension, where the majority of candidate compounds were indeed correctly identified hits (with a number of calculations yielding 100% hit rate), and results obtained for COX and TK were of comparable quality. In these cases, variations in the number of obtained candidate compounds and hit rates were also smaller than for activity classes BEN and SER.

Search Stringency. The predictive value and characteristics of DMC calculations can be evaluated by comparing recovery rates and hit rates. Recovery rates were also highest for activity class H3, with an average of ~24% at the “<= 0.1 or >=0.9” level, where average hit rates were ~56%. These numbers reflect the trend that hit rates were in general higher than recovery rates for the test cases studied here. On the basis of our calculations, recovery rates of 10−20% (e.g., three to five hits for 25 active compounds in ~1.3 million database molecules) could be expected for selection of ~100 of fewer database compounds. These recovery rates often remained more or less constant when even fewer candidate compounds were selected, which ultimately produced overall highest hit rates. Taken together, these observations indicate that DMC calculations were more prone to false negatives than false positives. This implies that even

Table 2. Virtual Screening Trials[a]

≤0.0 or ≥1.0					≤0.1 or ≥0.9					≤0.2 or ≥0.8				
ADC	DC	HR (%)	RR	DS	ADC	DC	HR (%)	RR	DS	ADC	DC	HR (%)	RR	DS
a. BEN: 10 Bait + 49 Active Database Compounds														
39	33439	0.1	0.80	33	16	3782	0.4	0.33	53	4	188	2.1	0.08	72
33	27641	0.1	0.67	41	19	16297	0.1	0.39	51	4	7	36.4	0.08	67
36	40680	0.1	0.73	45	24	8026	0.3	0.49	49	3	235	1.3	0.06	72
39	39706	0.1	0.80	35	17	5115	0.3	0.35	46	4	108	3.6	0.08	57
22	3292	0.7	0.45	38	8	2433	0.3	0.16	54	1	101	1.0	0.02	65
43	60262	0.1	0.88	25	23	3356	0.7	0.47	41	5	5	50.0	0.10	55
21	12948	0.2	0.43	41	7	1026	0.7	0.14	63	2	25	7.4	0.04	74
35	45848	0.1	0.71	27	25	6809	0.4	0.51	38	0	650	0.0	0.00	62
43	43973	0.1	0.88	25	12	4910	0.2	0.24	53	8	1184	0.7	0.16	63
27	29187	0.1	0.55	25	13	5555	0.2	0.27	52	5	249	2.0	0.10	62
42	48338	0.1	0.86	32	21	15344	0.1	0.43	44	7	2582	0.3	0.14	66
37	29939	0.1	0.76	34	22	10590	0.2	0.45	44	7	0	100.0	0.14	57
24	10404	0.2	0.49	37	13	168	7.2	0.27	56	6	2	75.0	0.12	68
27	17454	0.2	0.55	40	11	260	4.1	0.22	53	2	0	100.0	0.04	79
17	17528	0.1	0.35	43	12	9103	0.1	0.24	53	5	538	0.9	0.10	63
31	12988	0.2	0.63	26	20	6075	0.3	0.41	37	16	3544	0.4	0.33	46
36	32166	0.1	0.73	40	20	16297	0.1	0.41	51	4	1098	0.4	0.08	71
40	58563	0.1	0.82	26	18	4112	0.4	0.37	42	3	7	30.0	0.06	64
37	32220	0.1	0.76	25	30	22362	0.1	0.61	37	3	0	100.0	0.06	60
27	15374	0.2	0.55	47	12	2168	0.6	0.24	54	0	3	0.0	0.00	71
b. COX: 10 Bait + 21 Active Database Compounds														
9	63	12.5	0.43	62	5	0	100.0	0.24	75	1	0	100.0	0.05	83
15	140	9.7	0.71	50	3	18	14.3	0.14	62	1	0	100.0	0.05	75
8	45	15.1	0.38	51	3	2	60.0	0.14	65	0	0	0.0	0.00	77
5	12	29.4	0.24	50	1	1	50.0	0.05	72	0	1	0.0	0.00	81
7	44	13.7	0.33	57	1	4	20.0	0.05	73	0	0	0.0	0.00	82
10	362	2.7	0.48	49	5	19	20.8	0.24	69	4	0	100.0	0.19	78
5	16	23.8	0.24	52	1	1	50.0	0.05	69	1	0	100.0	0.05	77
12	423	2.8	0.57	48	6	9	40.0	0.29	65	1	0	100.0	0.05	80
12	641	1.8	0.57	47	3	3	50.0	0.14	66	1	0	100.0	0.05	81
11	191	5.4	0.52	44	5	2	71.4	0.24	65	2	0	100.0	0.10	73
7	69	9.2	0.33	52	3	13	18.8	0.14	65	1	0	100.0	0.05	81
7	70	9.1	0.33	53	3	0	100.0	0.14	78	3	0	100.0	0.14	83
9	218	4.0	0.43	48	5	7	41.7	0.24	61	2	0	100.0	0.10	73
14	101	12.2	0.67	47	4	0	100.0	0.19	65	2	0	100.0	0.10	77
8	25	24.2	0.38	50	4	3	57.1	0.19	67	1	0	100.0	0.05	89
13	305	4.1	0.62	43	5	1	83.3	0.24	61	2	0	100.0	0.10	85
14	355	3.8	0.67	44	7	55	11.3	0.33	65	1	0	100.0	0.05	80
4	20	16.7	0.19	67	2	0	100.0	0.10	82	1	0	100.0	0.05	85
18	712	2.5	0.86	44	10	139	6.7	0.48	57	1	0	100.0	0.05	80
2	8	20.0	0.10	60	0	1	0.0	0.00	72	0	0	0.0	0.00	82
c. H3: 10 Bait + 42 Active Database Compounds														
27	54	33.3	0.64	57	12	0	100.0	0.29	65	9	0	100.0	0.21	73
33	414	7.4	0.79	24	15	0	100.0	0.36	68	3	0	100.0	0.07	82
24	99	19.5	0.57	60	15	46	24.6	0.36	67	8	0	100.0	0.19	74
30	32	48.4	0.71	57	16	2	88.9	0.38	68	2	0	100.0	0.05	84
24	41	36.9	0.57	60	8	0	100.0	0.19	75	3	0	100.0	0.07	83
32	184	14.8	0.76	36	23	44	34.3	0.55	62	12	16	42.9	0.29	68
18	80	18.4	0.43	63	12	7	63.2	0.29	68	6	0	100.0	0.14	79
27	18	60.0	0.64	64	8	0	100.0	0.19	77	2	0	100.0	0.05	82
24	31	43.6	0.57	63	20	1	95.2	0.48	72	5	0	100.0	0.12	81
39	166	19.0	0.93	55	21	68	23.6	0.50	61	2	0	100.0	0.05	75
30	45	40.0	0.71	58	10	2	83.3	0.24	68	1	0	100.0	0.02	81
34	93	26.8	0.81	59	8	12	40.0	0.19	69	4	3	57.1	0.10	75
21	78	21.2	0.50	60	4	0	100.0	0.10	73	0	0	0.0	0.00	82
35	181	16.2	0.83	35	15	6	71.4	0.36	64	2	0	100.0	0.05	79
20	0	100.0	0.48	71	13	0	100.0	0.31	76	3	0	100.0	0.07	82
17	297	5.4	0.40	26	4	1	80.0	0.10	68	0	0	0.0	0.00	83
21	20	51.2	0.50	53	13	0	100.0	0.31	76	3	0	100.0	0.07	83
26	162	13.8	0.62	58	8	16	33.3	0.19	68	0	0	0.0	0.00	78
32	138	18.8	0.76	37	13	25	34.2	0.31	63	2	14	12.5	0.05	72
29	144	16.8	0.69	46	4	0	100.0	0.10	75	1	0	100.0	0.02	85

Table 2 (Continued)

≤0.0 or ≥1.0					≤0.1 or ≥0.9					≤0.2 or ≥0.8				
ADC	DC	HR (%)	RR	DS	ADC	DC	HR (%)	RR	DS	ADC	DC	HR (%)	RR	DS
d. SER: 10 Bait + 61 Active Database Compounds														
27	2075	1.3	0.44	51	6	92	6.1	0.10	72	3	62	4.6	0.05	79
30	1288	2.3	0.49	41	23	472	4.6	0.38	61	2	7	22.2	0.03	77
24	1942	1.2	0.39	59	4	305	1.3	0.07	72	3	96	3.0	0.05	81
40	3873	1.0	0.66	21	9	265	3.3	0.15	58	1	5	16.7	0.02	72
12	1111	1.1	0.20	53	8	536	1.5	0.13	73	1	12	7.7	0.02	81
38	4837	0.8	0.62	37	18	552	3.2	0.30	57	3	82	3.5	0.05	77
28	1839	1.5	0.46	62	15	197	7.1	0.25	75	4	61	6.2	0.07	83
44	6525	0.7	0.72	33	20	1809	1.1	0.33	62	0	5	0.0	0.00	76
33	2938	1.1	0.54	44	17	400	4.1	0.28	66	12	3	80.0	0.20	80
47	4633	1.0	0.77	22	24	296	7.5	0.39	36	3	2	60.0	0.05	51
21	290	6.8	0.34	31	11	36	23.4	0.18	51	1	4	20.0	0.02	77
37	1851	2.0	0.61	35	8	87	8.4	0.13	52	0	0	0.0	0.00	75
19	261	6.8	0.31	61	13	55	19.1	0.21	74	3	9	25.0	0.05	86
48	5813	0.8	0.79	33	16	163	8.9	0.26	62	3	13	18.8	0.05	79
28	1421	1.9	0.46	52	20	367	5.2	0.33	68	8	9	47.1	0.13	80
25	1392	1.8	0.41	40	14	460	3.0	0.23	55	3	7	30.0	0.05	74
30	653	4.4	0.49	25	11	32	25.6	0.18	48	0	1	0.0	0.00	81
46	6268	0.7	0.75	30	14	166	7.8	0.23	63	0	2	0.0	0.00	78
42	3014	1.4	0.69	37	8	59	11.9	0.13	61	2	12	14.3	0.03	83
31	1780	1.7	0.51	46	16	33	10.7	0.26	68	14	41	25.5	0.23	79
e. TK: 10 Bait + 25 Active Database Compounds														
16	866	1.8	0.64	27	5	138	3.5	0.20	42	2	2	50.0	0.08	62
13	479	2.6	0.52	29	5	95	5.0	0.20	54	1	40	2.4	0.04	86
10	483	2.0	0.40	40	8	298	2.6	0.32	56	2	11	15.4	0.08	77
10	456	2.1	0.40	29	5	182	2.7	0.20	70	1	1	50.0	0.04	77
20	861	2.3	0.80	27	5	54	8.5	0.20	62	3	31	8.8	0.12	73
17	685	2.4	0.68	26	9	242	3.6	0.36	49	4	76	5.0	0.16	71
9	357	2.5	0.36	61	0	46	0.0	0.00	71	0	6	0.0	0.00	82
8	437	1.8	0.32	57	4	179	2.2	0.16	74	3	68	4.2	0.12	85
15	869	1.7	0.60	27	3	85	3.4	0.12	69	2	15	11.8	0.08	85
21	2267	0.9	0.84	33	6	138	4.2	0.24	61	3	1	75.0	0.12	67
20	1371	1.4	0.80	23	6	27	18.2	0.24	51	2	0	100.0	0.08	67
17	497	3.3	0.68	31	5	50	9.1	0.20	53	2	17	10.5	0.08	79
12	1022	1.2	0.48	57	6	99	5.7	0.24	78	3	33	8.3	0.12	86
16	1481	1.1	0.64	36	7	147	4.5	0.28	57	3	11	21.4	0.12	75
12	1006	1.2	0.48	44	3	19	13.6	0.12	68	2	3	40.0	0.08	76
16	1036	1.5	0.64	27	8	145	5.2	0.32	70	3	43	6.5	0.12	80
10	459	2.1	0.40	34	6	157	3.7	0.24	62	3	1	75.0	0.12	77
8	422	1.9	0.32	54	8	291	2.7	0.32	62	1	54	1.8	0.04	78
10	380	2.6	0.40	49	6	128	4.5	0.24	73	1	6	14.3	0.04	85
10	514	1.9	0.40	45	7	239	2.8	0.28	66	4	35	10.3	0.16	78

[a] The tables report virtual screening calculations for each activity class based on 20 different randomly selected sets of bait compounds. For clarity, only two steps of dimension extension are reported. The following abbreviations are used: ADC, active database compounds (potential hits); DC, other database compounds; HR, hit rate; RR, recovery rate; DS, number of descriptors for definition of consensus position (corresponding to the dimensionality of the chemical reference space).

simplified descriptor reference spaces produced a rather high level of chemical resolution, due to their high dimensionality.

Comparison with RMP. The same activity class data set studied here was previously used to benchmark recursive median partitioning (RMP),[28] the adaptation of the median partitioning algorithm for virtual screening, thus permitting direct comparison with DMC. Table 3 reports achieved hit and recovery rates. The RMP calculations produced average hit rates of 14% (with a maximum of 21% for SER) and average recovery rates of 19%, whereas DMC produced average hit rates of 37% (with a maximum of 74% for H3) and average recovery rates of 18%. As shown in Table 3, DMC hit rates were higher for four of five activity classes (except SER) than RMP hit rates, and DMC recovery rates were higher for three of five classes. Thus, while recovery

Table 3. Comparison of DMC and RMP[a]

	DMC		RMP	
activity class	HR (%)	RR (%)	HR (%)	RR (%)
BEN	26	9	12	18
COX	50	18	3	11
H3	74	29	3	5
SER	9	23	21	22
TK	26	9	13	40

[a] DMC and RMP calculations were carried out on the same activity classes and in the same compound source database. RMP hit rates were taken from ref 28, and recovery rates were calculated from these data.[28] DMC rates are reported as average values of the 20 trials for the overall best scoring dimension extension level, as shown in Table 2. HR stands for hit rate and RR for recovery rate.

rates were overall quite similar in magnitude, DMC calculations produced significantly higher hit rates, which we attribute to the fact that DMC analysis operates in higher-dimensional (and more resolved) binary descriptor space representations than RMP.

Conclusions. In this study, we have introduced a new method for coordinate-based analysis of molecular similarity in chemical reference spaces. This methodology combines elements of bit string and partitioning methods, and its novel features include the generation of compound-class specific descriptor spaces and dimension extension. Molecular similarity relationships are detected through mapping of consensus positions of active compounds. Various types of molecular property descriptors can be used for DMC analysis, and there are no limitations on their numbers. In addition, the low computational complexity of this approach permits rapid mining of very large compound databases. In the test cases studied here, consensus positions could always be mapped, and different classes of compounds produced different descriptor solutions. In virtual screening-type calculations, DMC consistently identified compound having similar activity. Best results were obtained if database compounds were reduced to a rather small number of candidates for testing (e.g., 10−100). Although descriptor spaces for DMC analysis were simplified through median-based binary transformation, they became highly discriminatory through redefinition of consensus positions and dimension extension and consistently permitted the identification of similar molecules in the presence of large numbers of background compounds.

ACKNOWLEDGMENT

The authors thank Douglas B. Kitchen for helpful discussions.

REFERENCES AND NOTES

(1) *Concepts and applications of molecular similarity;* Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.

(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) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(4) 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.

(5) Willett, P.; Wintermann, V.; Bawden, D. Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(6) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Design* **1998**, *9*, 339−353.

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

(8) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(9) 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.

(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) Duca, J. S.; Hopfinger, A. J. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(12) Labute, P. Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, *4*, 444−455.

(13) 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.

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

(15) 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.

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

(17) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(18) 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.

(19) Xie, D.; Tropsha, A.; Schlick, T. An efficient projection protocol for chemical databases: single value decomposition combined with truncated Newton minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 167−177.

(20) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.

(21) Agrafiotis, D. K.; Lobanov, V. S. Multidimensional scaling of combinatorial libraries without explicit enumeration. *J. Comput. Chem.* **2001**, *22*, 1712−1722.

(22) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707−715.

(23) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 165−179.

(24) Sadowski, J. Optimization of chemical libraries by neural network methods. *Curr. Opin. Chem. Biol.* **2000**, *4*, 280−282.

(25) 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.

(26) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **2002**, *1*, 882−894.

(27) MOE (Molecular Operating Environment), version 2001.01, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.

(28) 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.

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

(30) 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*, 1151−1157.

CI0302963