Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations

Florence L. Stahura, Jeffrey W. Godden, Ling Xue, and Jürgen Bajorath*

Computer-Aided Drug Discovery, New Chemical Entities, Inc., 18804 North Creek Parkway,
Bothell, Washington 98011, and New Chemical Entities and Department of Biological Structure,
University of Washington, Seattle, Washington 98195

Received April 24, 2000

Molecular descriptors were identified by Shannon entropy analysis that correctly distinguished, in binary QSAR calculations, between naturally occurring molecules and synthetic compounds. The Shannon entropy concept was first used in digital communication theory and has only very recently been applied to descriptor analysis. Binary QSAR methodology was originally developed to correlate structural features and properties of compounds with a binary formulation of biological activity (i.e., active or inactive) and has here been adapted to correlate molecular features with chemical source (i.e., natural or synthetic). We have identified a number of molecular descriptors with significantly different Shannon entropy and/or "entropic separation" in natural and synthetic compound databases. Different combinations of such descriptors and variably distributed structural keys were applied to learning sets consisting of natural and synthetic molecules and used to derive predictive binary QSAR models. These models were then applied to predict the source of compounds in different test sets consisting of randomly collected natural and synthetic molecules, or, alternatively, sets of natural and synthetic molecules with specific biological activities. On average, greater than 80% prediction accuracy was achieved with our best models. For the test case consisting of molecules with specific activities, greater than 90% accuracy was achieved. From our analysis, some chemical features were identified that systematically differ in many naturally occurring versus synthetic molecules.

INTRODUCTION

It is a common conjecture that many natural products (NP) and synthetic compounds have distinctly different chemical characteristics. In some cases, for example certain classes of antibiotics or highly complex metabolites, it is relatively easy to recognize a compound isolated from natural sources. In many others, however, chemical differences between natural and synthetic molecules are subtle and much less obvious, if at all present. Relatively little effort has thus far been spent to systematically compare characteristics of natural products and synthetic molecules. Only recently some differences were identified by statistical analysis of structural fragments in natural and synthetic compounds.[1] These include, for example, different distributions of halogen atoms and nitrogen or oxygen containing groups. Amides and halogens are more frequently found in synthetic compounds, while naturally occurring molecules are often richer in oxygen (e.g., alcohol or ester groups). Furthermore, natural products were found to have, on average, higher molecular weights than synthetic compounds and a higher degree of steric complexity.[1]

We were interested in exploring systematic differences between natural products and synthetic compounds from a chemoinformatics point of view. We reasoned that, if such differences exist, we might be able to capture them by analyzing large collections of molecules from natural and synthetic sources. If successful, it should then be possible to go a step further, derive predictive models to distinguish between natural and synthetic molecules, and, ultimately, better understand some general chemical differences between compounds from natural and synthetic sources. Conceptually, this approach is somewhat similar to investigations designed to recognize drug-like features in compounds or distinguish between drugs and nondrugs.[2,3]

Rather than attempting to identify differences between natural and synthetic molecules by substructure search methods and/or direct statistical analysis, we initially set out to explore systematic differences in the values and distributions of molecular descriptors[4−6] that were calculated for compounds in large natural and synthetic compound databases. Such molecular descriptors typically account for physicochemical properties, molecular topology, connectivity, conformational parameters, or structural keys (fragments). Descriptor-based representations of molecular structure and properties are often used to analyze structure−activity relationships and molecular similarity or diversity.[7−9]

We intended to identify molecular descriptors that are sensitive to class-specific properties of natural and synthetic compounds as a basis for the generation of predictive models to distinguish between these molecules. Thus, the identification of descriptors that have systematically different values in natural and synthetic molecules is an important step in our approach. A chemical descriptor whose value varies little within compounds in a database has little power to distinguish these compounds. However, if its value differs significantly from those calculated for another compound collection, it may well be used to distinguish between these

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

10.1021/ci0003303 CCC: $19.00
Published on Web 09/02/2000

classes of compounds. For example, a descriptor that is sensitive to rare earth elements has little discriminatory power when used within a combinatorial library but is understandably useful for analyzing a library of anti-cancer agents.

A major difficulty for this type of analysis is that distributions of molecular descriptor values in compound databases cannot be directly translated into variability because units and value ranges of descriptors differ. Therefore, descriptor variability was analyzed by application of an entropic formulation, termed "Shannon entropy" (SE),[10] which was originally applied in digital communication technology.[10] In our implementation, this entropic metric, calculated from a previously described histogram method and binning algorithm,[11] is a robust estimator of descriptor variance. The method makes it possible to compare descriptors of very different nature because it considers each descriptor's relative information content. In this context, SE combines both the extrinsic variability of the descriptor across compounds in a chemical database and the intrinsic variability that results from the nature of the descriptor itself. The latter aspect carries the notion of the "bandwidth potential" of a descriptor. For example, a numerical descriptor counting the number of rotatable bonds in a molecule has significantly higher intrinsic variability than a descriptor that detects the presence or absence of a particular structural motif.

In the approach described herein, SE-based characterization of descriptors is followed by their application in binary QSAR analysis. Binary QSAR is a statistical method originally developed to identify molecules with desired biological specificity by virtual screening of compound databases.[12,13] It correlates structural features and properties of molecules, represented by molecular descriptors, with a binary formulation of activity ("active" versus "inactive") and can thus be used to derive predictive models on the basis of screening data.[13] It has also been successfully applied to study structure–activity relations of estrogen receptor[14] ligands and inhibitors of carbonic anhydrase.[15] Here we have adapted the method to correlate molecular properties, as expressed by values of descriptors selected from SE calculations, with chemical origin (i.e., "natural" or "synthetic"), rather than activity.

Both the Shannon entropy and binary QSAR concepts were originally developed for applications quite different from those reported here. However, we demonstrate that combining these conceptually different approaches has made it possible to effectively discriminate between natural products and synthetic compounds. High prediction accuracy was achieved by more than one combination of relatively few descriptors. In addition, chemically intuitive intrinsic differences between natural products and synthetic molecules could be deduced from successfully applied models.

MATERIALS AND METHODS

As sources for our analysis, the ACD[16] (synthetic compounds) and Chapman and Hall[17] (CH; naturally occurring molecules) databases were selected. Values of a total of 98 different descriptors that could be calculated from 2D representations of molecules were generated for 199,420 ACD and 116,364 CH compounds using MOE.[18] Table 1 lists all descriptors used to generate predictive models and those specifically discussed. In addition, the set of 166 MACCS structural keys[6,19] was included in our analysis. The variability, V, of each MACCS key in ACD and CH was calculated as follows:

$$V = abs[(K_{ACD}/N_{ACD}) - (K_{CH}/N_{CH})] \qquad (1)$$

K_{ACD} and K_{CH} are the number of times a specific MACCS key occurs in ACD and CH, respectively, and $\mathrm{N}_{\mathrm{ACD}}$ and N_{CH} are the total number of compounds in the ACD and in CH databases ("abs" stands for the absolute value).

The variability of numerical descriptors that can adopt a wide range of values was determined by calculation of Shannon entropy values. Shannon entropy is defined as

$$SE = -\sum p_i \log_2 p_i \qquad (2)$$

In this formulation, p is the probability of observing a particular descriptor value. p is calculated from the number of compounds with a descriptor value that falls within a specific histogram bin, or "count" (c), for a specific data interval i. Thus, p is calculated as

$$p_i = c_i/\sum c_i \qquad (3)$$

Equation 2 contains a logarithm to the base 2, which corresponds to a scale factor and permits the resulting SE to be considered as the number of binary bits necessary to capture the information contained within the descriptor variation. In this fashion, SE values for different data sets can be directly compared, provided a uniform binning scheme can be defined. This is the case when data sets are represented in histograms where the data range is divided into the same number of bins. As long as the number of data intervals between the minimum and maximum value is held constant, SE values are independent of the size of the interval. Such a binning algorithm was previously implemented[11] and used here to calculate SE values of the 98 numerical descriptors for all ACD and CH compounds.

"Entropic separation" of descriptors in ACD and CH was calculated by adapting a statistical approach commonly applied to normal distributions and variances. Instead of variance, SE values were used, and in place of the data average (an estimator of the central tendency) the distribution mode was used. Thus, the absolute distance between the bin numbers of the most populated bins in histograms (the mode) of descriptor distributions in ACD and CH was divided by the average of half of their SE values (analogous to half of the variance).

Binary QSAR methodology and its original applications have been described in detail elsewhere.[12–15] Binary QSAR employs the Bayes' Theorem[20] to correlate properties of molecules with a probability to adopt, or be within, two states, "active" or "inactive". Based on a learning set consisting of "active" and "inactive" compounds, calculated values of specified molecular descriptors are related to a binary formulation of activity (i.e., "active" = 1 and "inactive" = 0). Binary QSAR then estimates the probability density of molecules in a test set to be active or inactive by assigning scores between 0 and 1 to each molecule on the basis of calculated descriptor values.[12] This is done by principal component analysis[21] of molecular descriptor space

Table 1. Definition of Molecular Desriptors

descriptor	definition
	(a) Physicochemical Properties and Charge Descriptors[a]
Fcharge	total charge of the molecule (sum of formal charges)
PC+	sum of the positive q_i
PC−	sum of the negative q_i
RPC+	the largest positive q_i divided by the sum of the positive q_i
RPC−	the smallest negative q_i divided by the sum of the negative q_i
PEOE_RPC+	the largest positive p_i divided by the sum of the positive p_i
PEOE_RPC−	the smallest negative p_i divided by the sum of the negative p_i
PEOE_VSA-5	sum of v_i where p_i is in the range [−0.30,−0.25]
PEOE_VSA+1	sum of v_i where p_i is in the range [0.05,0.10]
PEOE_VSA+2	sum of v_i where p_i is in the range [0.10,0.15]
PEOE_VSA+3	sum of v_i where p_i is in the range [0.15,0.20]
PEOE_VSA+4	sum of v_i where p_i is in the range [0.20,0.25]
apol	sum of the atomic polarizability (including implicit hydrogens)
bpol	sum of the absolute value of the difference between atomic polarizability of all bonded atoms in the molecule
	(b) Atom, Bond, and Electron Pair Count Descriptors
a_ICM	entropy of the distribution of elements in the molecule
a_nBr	number of bromine atoms
a_nP	number of phosphorus atoms
a_nI	number of iodine atoms
a_nH	number of hydrogen atoms
b_1rotR	fraction of single nonring bonds
b_single	number of single nonaromatic bonds
b_double	number of double non aromatic bonds
b_triple	number of triple bonds
VadjEq	entropy of the distribution of values in the adjacency matrix
VadjMa	log (2 times the number of bonds)
b_ar	number of aromatic bonds
	(c) Connectivity Indices, Kier Kappa Shape Indices,[22] Graph Distance Matrix Descriptors, and Surface Area Descriptors[b]
KierA3	third alpha modified shape index
petitjean	value of (diameter − radius)/diameter; (shape descriptor)[23]
radius	if r_i is the largest distance matrix entry in row I of A, then the radius is defined as the smallest of r_i[23]
chi1v	sum of the inverse square roots of $v_i v_j$ for all bonded heavy atoms i and j
chi0v_C	sum of the inverse square roots of v_i for the carbon atoms
VdistMa	if m is the sum of the distance matrix entries then VDistMa is defined to the sum of a_{ij}* log a_{ij}/m − log m over all i and j (logarithms are taken in base 2)
VdistEq	entropy of the distribution of values in the distance matrix
vsa_acid	approximately to the sum of the VDW surface area of acidic atoms
vsa_hyd	approximately to the sum of the VDW surface area of hydrophobic atoms
vsa_pol	approximately to the sum of the VDW surface area of polar atoms
vsa_base	approximately to the sum of the VDW surface area of basic atoms
vsa_other	approximately to the sum of the VDW surface area of other atoms

[a] q_i is the partial charge of atom i in a molecule and p_i represents here the partial charge of atom i calculated according to the PEOE method[24] and v_i is the van der Waals surface area of atom i. [b] d_i is the number of heavy atoms bonded to atom i. $v_i = (p_i - h_i)/(z_i - p_i - 1)$ where, in this case, p_i is the number of s and p valence electrons of atom i, h_i is the number of hydrogen bonded to atom i, and z_i is the atomic number of atom i. n is the number of atoms in the non-hydrogen graph of the molecule, m is the number of bonds, and a is the sum of $(r_i/r_c - 1)$. r_i is the covalent radius of atom i and r_c is the covalent radius of a carbon atom. The graph distance matrix of a molecule with n atoms is defined as the n by n matrix, A, where a_{ij} is the length of the shortest path in graph between atoms i and j. The descriptors represent values derived from the graph distance matrix of the non-hydrogen molecular graph of a molecule.

to obtain a decorrelated and normalized set of descriptors and derive a probability density function. Each descriptor combination defines a specific probability function used as a model to predict the active or inactive state of test compounds. Since the probability function produces continues values between 0 and 1, a cutoff value must be defined to discriminate between active and inactive states.

Binary models were developed based on different descriptor combinations using the QuaSAR−Model module[13] of MOE.[18] Furthermore, combinations of preferred molecular descriptors were explored by factorial analysis.[9] A probability cutoff value of 0.5 was used to classify compounds as natural (>0.5) or synthetic (<0.5). In all binary QSAR calculations, the number of principal components derived from descriptor combinations was variable and not limited. A smoothing factor of 0.25 was applied to each probability function.[12] Training sets for binary QSAR models consisted of 500 or 10,000 compounds. In each case, half of the compounds were randomly selected from ACD and half from CH. Both learning sets gave very similar binary models, suggesting that their size was not critical. Six binary QSAR models (M1-M6) were generated on the basis of different descriptor sets. Using these models, predictions were carried out for six test databases (D1-D6). D1-D3 each consisted of 500 different randomly selected compounds (250 from ACD and 250 from CH), D4 of 1000 compounds (500 ACD and 500 CH), and D5 of 500 (400 ACD and 100 CH). In contrast to randomly assembled test sets, D6 consisted of sets of natural or synthetic compounds each having a specific biological activity, as further described in the Results section.

The performance of different models was evaluated by calculating overall prediction accuracy:

$$PA = (CN + CS)/NT \tag{4}$$

CN is the number of correctly identified natural products, CS is the number of correctly identified synthetic compounds,

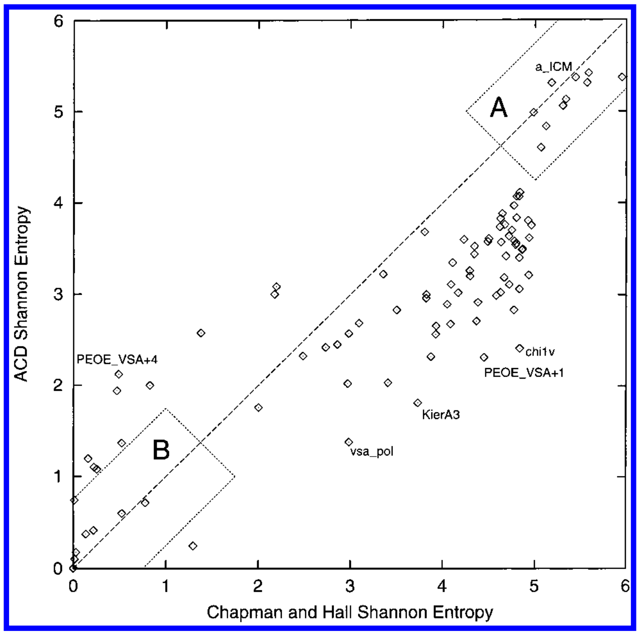

Figure 1. Shannon entropy comparison. SE values of descriptors calculated for ACD compounds are plotted against corresponding values for the Chapman and Hall database. Region (**A**) contains descriptors with highest Shannon entropy and region (**B**) those with low entropy (and thus little variability). "Off-diagonal" descriptors have the greatest difference in variability between the two databases and are labeled. Descriptors in region (**A**) are as follows: b_rotR, b_1rotR, VAdjEq, a_ICM, PEOE_RPC-, VDistEq, VDistMa, PEOE_RPC+, VAdjMa, balabanJ, and in (**B**): PC+, RPC-, FCharge, PC-, RPC+, a_nI, a_nP, a_nF, a_nBr, b_triple (for definitions, see Table 1).

and NT is the total number of compounds in the database. If all compounds were correctly identified, a prediction accuracy of 1 (or 100%) would be achieved, but wrong predictions reduce the score. Due to the binary classification scheme, any "missed" natural molecule is a "false positive" synthetic one and vice versa.

RESULTS AND DISCUSSION

Shannon Entropy Analysis. Calculation of SE values for descriptors in ACD and CH revealed significant differences in their variability, with absolute values ranging from close to zero (no variability) to almost six (high variability). Figure 1 shows a comparison of descriptor entropy values calculated for ACD and CH compounds. Some descriptors have low entropy values in both natural and synthetic compounds (area **B** in Figure 1), whereas others have consistently high entropy (area **A**). For example, "b_triple", the number of triple bonds in a molecule, has little variation in both ACD and CH and would thus not be suitable to distinguish between such compounds. By contrast, some PEOE-type charge descriptors (see also Table 1) are highly variable in both databases and may therefore be better candidates. Equally interesting are descriptors in off-diagonal regions of the Shannon entropy plot. These descriptors show strong variation in one database but relatively little in the other. For example, "vsa_pol", a descriptor of polar surface area, is about twice as variable in CH than in ACD, while the opposite is observed for "PEOE_VSA+4" (Figure 1).

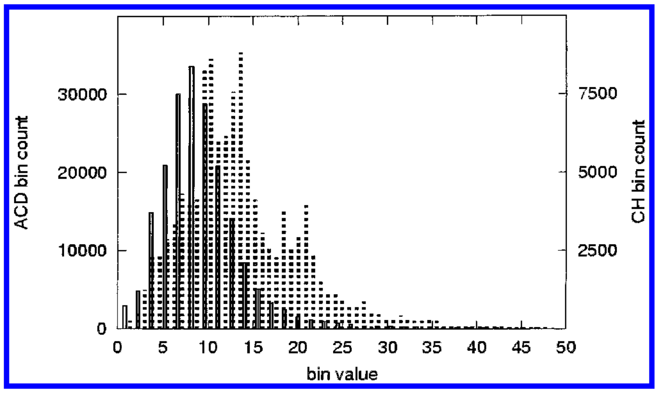

Figure 2. Descriptor distributions. The distribution of descriptor "chi0v_C", used as an example, in ACD (solid line/left scale) is shown overlaid with the corresponding histogram for CH (dotted line/right scale). In both cases, the central tendencies are taken from the mode or the bin with the highest occupancy. This histogram provides the basis for the calculation of "entropic separation" (see Methods).

Table 2. Descriptors with the Largest Entropic Separations in ACD and CH Databases[a]

descriptor	entropic separation	SE (CH/ACD)
a_ICM	8.14	5.4/5.3
bpol	5.08	4.8/3.1
chi0v_C	4.71	4.9/3.6
b_double	4.52	2.9/2.5
chi1v	4.42	4.8/2.4
a_nH	4.08	4.7/3.2
b_single	3.93	4.9/3.2
b_ar	3.86	2.2/3.0
vsa_hyd	3.84	4.8/3.5
apol	3.83	4.8/3.6

[a] "SE" reports the absolute Shannon entropy value of descriptors in the ACD and Chapman and Hall databases. Descriptor chi0v_C is used as an example to illustrate the calculations of Shannon entropy and entropic separation. First, SE values are calculated independently for each compound database. For chi0v_C the histogram bin counts from ACD starting from the left in Figure 2 are as follows: 2933, 4903, 14877, 21023, 29964, ... 0, 1. These raw counts are divided by the sum of the bin counts to produce the probabilities (p_i): 0.025, 0.042, 0.128, 0.181, 0.258, ... 0.000, 0.000. These probabilities are summed in using equation $-\sum p_i \log_2 p_i$ producing the SE values reported in the right-hand column above. Next, to calculate the entropic separation, the bin containing the mode (i.e., the most frequently observed value) is identified for each database. For chi0v_C the bin number for the mode of the ACD distribution is 6, and for the CH distribution it is 16 (see Figure 2**)**, thus the bin separation between the modes is 10. To express this separation as a ratio of the conjoint variability of the data (i.e., the entropic separation), this number is divided by the average of half of the Shannon entropies calculated above. For chi0v_C this value is 10/(((4.9/2)+(3.6/2))/2) = 4.71 see above).

Entropic Separation. Descriptors with high entropy in both databases and differential (off-diagonal) entropy values have good potential to be capable of discriminating between natural and synthetic compounds, provided the distribution of their values differs in a database-specific manner. To identify such descriptors, we have analyzed the separation of descriptor distributions in histograms and calculated entropic separations. As an example, the distributions of "chi0v_C" (a surface area descriptor; see Table 1) are shown in Figure 2. If the distributions of this descriptor across the ACD or CH databases are superimposed, the spread in both databases is approximately the same, whereas the two peaks are resolved from one another. Entropic separation is then calculated as described before. The 10 descriptors with largest entropic separation are reported in Table 2. It is notable that the top descriptor employs an entropy term in its own calculation. This descriptor, "a_ICM", captures the entropy

Table 3. Variability of Structural Keys[a]

key no.	K_{ACD}/N_{ACD}	K_{CH}/N_{CH}	V	definition
139	0.27	0.74	0.47	OH group
161	0.71	0.26	0.45	nitrogen
127	0.18	0.62	0.44	more than one nonring oxygen attached to ring
116	0.18	0.60	0.42	three bonds between a CH_3 and CH_2
156	0.66	0.25	0.42	three or four coordinated atoms bound to a nitrogen
152	0.44	0.85	0.41	three or four coord. carbons bound to two oxygens and a carbon
99	0.21	0.62	0.41	C=C double bond
105	0.28	0.68	0.41	three ring atoms bonded to a central ring atom
108	0.19	0.58	0.39	four bonds between a CH_3 and CH_2
143	0.40	0.78	0.39	nonring oxygen attached to a ring

[a] "V" means variability and identifies the MACCS keys that are most variable between the ACD and CH databases. It is defined as the absolute value of ($K_{ACD}/N_{ACD} - K_{CH}/N_{CH}$), where K_{ACD} is the number of times a specific MACCS key is set on in the ACD database and K_{CH} is the corresponding value for CH. N_{ACD} and N_{CH} are the total number of compounds in the ACD and CH databases, respectively (199,420 and 116,364). Only two of the 10 most variable MACCS keys (#156 and #161) occur with higher frequency in ACD than CH.

Table 4. Binary QSAR Models

binary QSAR model	descriptors	descriptor selection criteria
M1	petitjean, PEOE_VSA_+2, b_double, PEOE_VSA_−5, PEOE_VSA_+3, radius, vsa_other	descriptors with **average Shannon entropy** in both ACD and CH (**close to diagonal** in Figure 1)
M2	PC+, PC-, RPC+, RPC-, Fcharge, a_nI, a_nP, a_nBr, b_triple, vsa_acid, vsa_base	descriptors with **low Shannon entropy** in ACD and CH (**region B** in Figure 1)
M3	b_1rotR, VadjEq, a_ICM, PEOE_RPC-, VdistEq, VdistMa, PEOE_RPC+, VAdjMa	descriptors with **high Shannon entropy** in ACD and CH (**region A** in Figure 1)
M4	139, 161, 127, 116, 156, 152, 99, 105	**most variable MACCS keys** in the ACD and CH databases (ranked from the most to the least variable)
M5	a_ICM, bpol, chi0v_C, b_double, chi1v, a_nH, b_single, b_ar	descriptors with the **largest entropic separation**
M6	139, 127, 99, a_ICM, a_nH, b_ar, b_single	best model identified by **complete factorial analysis** of combinations of descriptors used in models 4 and 5 (see Results section)

of the element distribution within a molecule, and the range of possible values is already distilled to its information content. It has high SE values in both databases and by far the largest entropic separation. However, Table 2 also shows that large entropic separation does not necessarily require the presence of high SE values. For example, "b_double", which counts the number of double bonds in a molecule, shows only medium variability in both ACD and CH, yet it is one of the descriptors displaying the largest entropic separation.

Variable Structural Keys. Since structural keys are binary in nature (i.e., present or absent), rather than numerical, SE values are, in this case, not a meaningful measure of their distribution. Therefore, we have determined the variability of 166 MACCS keys by calculating differences in their frequency of occurrence in ACD and CH. Significant differences in the distribution of a number of structural keys in natural and synthetic molecules were detected in these calculations. Table 3 lists the 10 keys with greatest difference in their frequency of occurrence. Structural keys representing the hydroxyl group and nitrogen atoms show greatest differences in natural and synthetic molecules, which is well in accord with previously discussed results of Henkel et al.[1] We also find significant differences for keys that (indirectly) address the degree of chemical saturation (e.g., double bonds) or condensation (e.g., fused ring atoms).

Descriptor Selection. With our variability analysis, we aimed to identify sets of molecular descriptors that are most likely to capture, directly or indirectly, natural and synthetic compound class-specific features. On the basis of our findings, we concluded that numerical descriptors with significant variability and/or high entropic separation in ACD and CH and, in addition, structural keys with markedly different distributions would be most promising to derive predictive models that are capable of distinguish between natural and synthetic compound. Using these guidelines, we selected specific sets of descriptors for model building, as discussed in the following.

Binary QSAR Models. Initially, we computed five alternative models using different descriptor combinations and our training set consisting of randomly selected natural and synthetic molecules. These models, M1-M5, are shown in Table 4. Later, based on the results of our test calculations, an additional model was computed, M6 in Table 4**,** which is discussed below. Models M1 and M2 were generated using descriptors that, following our approach, should not be powerful discriminators of natural and synthetic compounds, since they were based on descriptors with either average (M1) or low (M2) Shannon entropy. By contrast, models M3-M5 were generated using preferred descriptor combinations. M3 was based on descriptors with high Shannon entropy, M4 only on most variable structural keys, and M5 on those descriptors with largest entropic separation.

Model Evaluation and Predictive Performance. The binary QSAR models M1 to M5 were initially tested using two different databases, termed D1 and D6. D1 consisted of 250 ACD and 250 CH compounds that were randomly collected, and D6 consisted of 15 classes of synthetic and natural compounds (a total of 245 compounds) each of which having a specific biological activity (mostly enzyme inhibitors). The composition of D6 is reported in Table 5. Therefore, whereas test case D1 consisted of compounds with

Table 5. Composition of Test Database 6 (D6)[a]

	biological activity	no. of compds
	Synthetic Compound Class	
BA_BEN	benzodiazepine receptor ligands	22
BA_CAE	carbonic anhydrase II inhibitors	22
BA_H3E	H3 antagonists	21
BA_TKE	tyrosine kinase inhibitors	20
BA_5HT	serotonin receptor ligands	21
BA_COX	cyclooxygenase-2 inhibitors	17
	Natural Products	
NP_5LP	5-lipoxygenase inhibitors	17
NP_ACE	angiotensin converting enzyme inhibitors	9
NP_CAT	Acyl-CoA: cholesterol acyltransferase inhibitors	20
NP_BLC	β-lactamase inhibitors	14
NP_PPD	phosphodiesterase inhibitors	14
NP_PA2	phospholipase 2 inhibitors	12
NP_PKC	protein kinase C inhibitors	15
NP_RVT	reverse transcriptase inhibitors	14
NP_TMB	thrombin inhibitors	7

[a] Synthetic compound classes were taken from the literature as described previously,[9] and classes of natural molecules with specific biological activity were assembled from the Chapman and Hall database.

Table 6. Prediction Accuracy Achieved by Application of Binary QSAR Models[a]

binary QSAR model	CN	CS	PA (%)
(A) Results Obtained for Test Database D1 (500 Molecules)			
M1	164	168	66.4
M2	215	47	52.5
M3	192	189	76.2
M4	205	203	81.6
M5	180	190	74.0
(B) Results for D6 (245 Compounds; See Table 5)			
M1	64	103	68.2
M2	112	18	53.0
M3	90	118	85.0
M4	104	119	91.0
M5	82	122	83.2

[a] "CN" is the number of correctly identified natural molecules and, "CS" is the number of correctly identified compounds from ACD. "PA" reports the overall prediction accuracy.

unspecified (random) characteristics, D6 was composed of compounds with target-specific activity (and thus more drug-like properties), yet different chemical origin. We generated D1 and D6 in this way to provide two conceptually different test cases.

The performance of models M1 to M5 in classifying compounds in D1 and D6 is reported in Table 6. Overall the models behaved as predicted from our descriptor variability analysis. M2 performs worst with a prediction accuracy of about 53%, which is close to what one would expect, in this case, from random predictions (i.e., 50%). M1, consisting of descriptors with average or medium entropy, performs somewhat better with about 67% prediction accuracy. By contrast, high prediction accuracy with, on average, greater than 80% was obtained with models M3, M4, and M5, which were designed using descriptors with highest variability or largest entropic separation. Each of these models consisted of only seven or eight descriptors. M4 was the best model with an average prediction accuracy of approximately 86%. This model was derived using the eight most variable structural keys and no numerical descriptor. These keys included, among others, those accounting for the nitrogen atoms, oxygen atoms attached to rings, and hydroxyl groups. This finding was consistent with differences in the frequency of occurrence of these keys in ACD and CH, as discussed above.

Table 7. Performance of Preferred Binary QSAR Model M6[a]

database	S	N	CN	CS	PA (%)
D1	250	250	199	206	81.0
D2	250	250	191	174	73.0
D3	250	250	199	231	86.0
D4	500	500	335	400	73.5
D5	400	100	81	363	88.8
D6	123	122	107	120	92.6

[a] "N" and "S" are the number of natural and synthetic molecules in each test database, respectively. "CN" is the number of correctly identified natural molecules, "CS" is the number of correctly identified synthetic compounds, and "PA" is the achieved overall prediction accuracy.

On the basis of these results, we went a step further, combined the descriptors of models M4 (structural keys) and M5 (largest entropic separation), and explored different combinations of these descriptors by factorial analysis for their predictive performance in D1 and D6. In these calculations, a model with further increased performance was identified (M6 in Table 4**)**. It consisted of seven descriptors, three structural keys, and four numerical descriptors and achieved a prediction accuracy of 81% in D1 and 93% in D6. Then, to investigate the influence of database-specific effects, M6 was tested on four additional randomly assembled natural/synthetic compound databases with different composition (D2 to D5). The results of all calculations are reported in Table 7. Observed prediction accuracy was between 73% and 93%, with an average of 83% for a total of 3245 test compounds. Thus, the predictive value of the model was high.

We also analyzed compounds that were incorrectly predicted in our calculations and a number of ACD compounds were identified that were either natural molecules or derivatives of natural products. Figure 3 shows some examples of such compounds (that were part of D1 and "incorrectly" predicted by M6). One of these compounds is a natural product (from the green algae *Botryococcus braunii*), another an unsaturated fatty acid, and the compounds are derivatives of molecules found in the Chapman and Hall database. Model M6 predicted these and other compounds to be "natural", but, in our analysis, they were considered incorrect because their source was ACD, i.e., "synthetic". Thus, overall prediction accuracy was probably even higher than suggested above, consistent with our finding that prediction accuracy was very high in database D6 that consisted of selected sets of natural and synthetic molecules, rather than randomly sampled ACD and CH compounds.

Class-Specific Descriptors. Several conclusions can be drawn from our findings. Relatively few descriptors, identified by variability analysis, were sufficient to distinguish between molecules from natural and synthetic sources, and, furthermore, different descriptor combinations yielded similar predictive performance. Our preferred descriptor combination consisted of three structural keys and four numerical descriptors and provided the basis of model M6 (Table 4). It illustrates that relatively simple descriptors, "a_ICM" being the exception, were sufficient to discriminate between natural

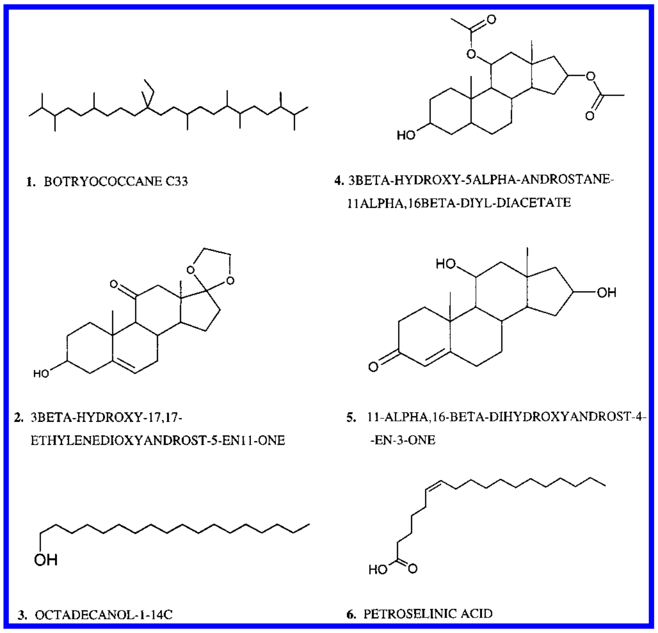
1. BOTRYOCOCCANE C33
HO
4. 3BETA-HYDROXY-5ALPHA-ANDROSTANE-11ALPHA,16BETA-DIYL-DIACETATE
HO
2. 3BETA-HYDROXY-17,17-ETHYLENEDIOXYANDROST-5-EN11-ONE
HO
OH
5. 11-ALPHA,16-BETA-DIHYDROXYANDROST-4--EN-3-ONE
OH
3. OCTADECANOL-1-14C
HO
6. PETROSELINIC ACID

Figure 3. "Incorrectly" identified compounds. These six molecules are part of test database D1 and were randomly collected from ACD and thus considered "synthetic" in our analysis. When evaluated with model M6, they produced a binary QSAR score of 0.97 or greater (i.e., "natural"). All these compounds are from natural sources or represent analogues of natural products. The examples shown here can be divided into two chemical types, one having a rigid steroid core structure and the other containing long (and highly flexible) aliphatic chains.

and synthetic molecules, if they were differentially set or distributed in ACD and CH. This is the case for structural keys accounting for hydroxyl groups (#139) and oxygen atoms attached to rings (#127). Moreover, values of other descriptors in our preferred sets inversely correlate and we can therefore rationalize why they have discriminatory power. Inversely correlated descriptors include, for example, the number of hydrogen atoms ("a_nH") or single bonds ("b_single") versus double bonds (key #99) or aromatic bonds ("b_ar").

Distinct Chemical Features. By analyzing descriptor distributions and calculating entropic separations, we have been able to identify chemical characteristics that distinguish many natural and synthetic molecules. In addition to the differences in the distribution of nitrogen and oxygen containing groups, as discussed above, we detected substantial differences in the degree of chemical saturation, condensation, and aromatic character. For example, of the 10 descriptors with highest entropic separation, shown in Table 2, nine have a greater central value (or mode) for natural than for synthetic compounds. "b_ar" is the only top 10 descriptor with reverse tendency, having a consistently higher value in synthetic compounds. Thus, synthetic compounds generally have more aromatic character than natural products that are, on average, composed of lower weight atoms.

CONCLUSIONS

We present a new approach to effectively classify synthetic and natural molecules on the basis of structural features and chemical properties. The method adopts and combines two concepts, Shannon entropy and binary QSAR, that have not been used previously in the same context and that have originally been developed for applications very different from those reported here. The obtained results confirm our initial hypothesis that the identification of differentially distributed and/or inversely correlated molecular descriptors is critical for the predictive value of binary QSAR models. As proposed, descriptors with low or average entropy or variability gave poor models, but those with high entropy and, in particular, large entropic separations were capable predicting whether a given compound was obtained from natural or synthetic sources. In addition, analysis of preferred descriptor combinations has made it possible to highlight some systematic chemical differences between natural and synthetic compounds.

REFERENCES AND NOTES

(1) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 643−647.

(2) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(3) Ajay; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem*. **1998**, *41*, 3314−3324.

(4) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(5) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D molecular descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(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) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(8) 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.

(9) 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*, 669−704.

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

(11) 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.

(12) Labute, P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, *7*, 444−455.

(13) Binary QSAR function is part of the "QuaSAR−Model" module of MOE (Molecular Operating Environment), Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3; described in Labute, P. Binary QSAR: A new technology for HTS and UHTS data analysis; electronic publication: http://www.chemcomp.com/feature/htsbqsar.htm.

(14) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.

(15) Gao H.; Bajorath J. Comparison of binary and 2D QSAR analysis using inhibitors of human carbonic anhydrase II as a test case. *Mol. Divers.* **1999**, *4*, 115−130.

(16) ACD (Available Chemicals Directory); available from MDL Information Systems Inc.: 14600 Catalina Street, San Leandro, CA 94577.

(17) Chapman and Hall, Dictionary of Natural Products, CD-ROM 1999 version; CRC Press LLC: NW Corporate Blvd, Boca Raton, FL 33431.

(18) MOE (Molecular Operating Environment), version 1999.05; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.

(19) MACCS keys; MDL Information Systems, Inc.: 14600 Catalina Street, San Leandro, CA 94557.

(20) Feller, W. *An Introduction to Probability Theory and its Applications*; Wiley & Sons Inc.: New York, 1950; Vol. 1.

(21) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.

(22) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.
(23) 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.
(24) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

CI0003303