Mining High-Throughput Screening Data of Combinatorial Libraries: Development of a Filter to Distinguish Hits from Nonhits

Andreas Teckentrup,[†] Hans Briem,[‡] and Johann Gasteiger*

Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Strasse 65,
D-88397 Biberach, Germany, and Computer-Chemie-Centrum, Institute for Organic Chemistry,
University of Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received October 16, 2003

Kohonen neural networks generate projections of large data sets defined in high-dimensional space. The resulting self-organizing maps can be used in many applications in the drug discovery process, such as to analyze combinatorial libraries for their similarity or diversity and to select descriptors for structure−activity relationships. The ability to investigate thousands of compounds in parallel also allows one to conduct a study based on single-dose experiments of high-throughput screening campaigns, which are known to have a greater uncertainty than IC_{50} or K_i values. This is demonstrated here for a data set of 5513 compounds from one combinatorial library. Furthermore, a method was developed that uses self-organizing maps not only as an indicator of structure−activity relationships, but as the basis of a classification system allowing predictive modeling of combinatorial libraries.

INTRODUCTION

High-throughput screening and combinatorial chemistry have become cornerstones of the early drug development process. The amount of data generated in this process requires new approaches to data analysis in order to extract relevant information from the observations.

In the hit-to-lead phase, strong emphasis is put on the synthesis of focused combinatorial libraries to explore the chemical environment of hit compounds found in medium-sized or large screening libraries. Nowadays, this process is not only driven by the intention to increase biological activity but also to improve pharmacokinetic properties. Such a process can be supported by in silico screening methods which try to predict whether virtual compounds are likely hits for a given target.

We will demonstrate how high-throughput screening results of a library obtained by parallel synthesis can be analyzed to extract structural characteristics of the hits and to develop a filter that allows the virtual screening of additional compounds. In the course of this process, first a relationship between biological activity and the structures of the compounds synthesized has to be established. It was one of our goals that the derived correlation between structure and activity should not only have a good predictive power but should also be amenable to physical or chemical interpretation.

Furthermore, we strongly believe that an initial investigation of a data set should be performed with an unsupervised learning method in order to allow for an unbiased view on the data set, thus preventing premature assumptions on structure−activity relationships. In previous applications,[1] we found self-organizing neural networks, such as that introduced by Kohonen[2,3] to be excellent methods for the unsupervised investigation of a data set.

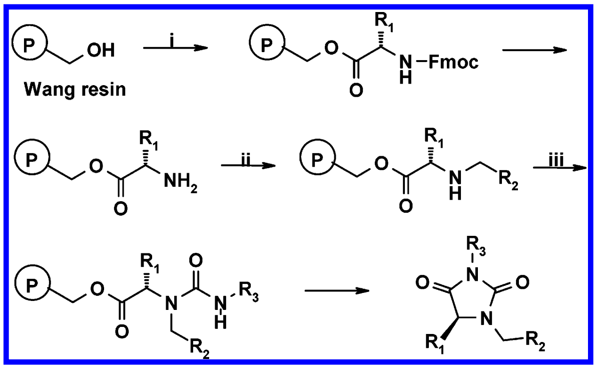

Figure 1. Composition of the combinatorial hydantoin library. Eighteen Fmoc-protected amino acids (i), 24 aldehydes (ii), and 24 isocyanates (iii) were combined to create 10 368 hydantoins.

Data Set. The data set used in this study comprises a series of hydantoins obtained in a parallel synthesis campaign from amino acids, aldehydes, and isocyanates, as depicted in Figure 1.

A set of 18 amino acids was reacted in a parallel manner with 24 aldehydes and then with 24 isocyanates, yielding 10 368 hydantoins. This compound library was part of a high-throughput screening campaign for an antagonistic activity toward a cardiovascular target, which provided percent-of-control (% ctrl) values in a one-point measurement for each compound submitted. Compounds with % ctrl values below 50% were classified as hits. In the particular assay investigated here, only 5513 compounds of the entire combinatorial library had been tested, 185 of which were identified as hits.

The hydantoin data set was selected for two reasons. First, with 5513 screened compounds, the data set was big enough to test the general performance of self-organizing networks when dealing with large amounts of data. Second, the number of hit compounds within the library was sufficiently large

* Corresponding author phone: +49 9131 85 26570; fax: +49 9131 85 26566; e-mail: gasteiger@chemie.uni-erlangen.de.
[†] Boehringer Ingelheim Pharma GmbH & Co. KG.
[‡] Current address: Schering AG, CDCC/Computational Chemistry, D-13342 Berlin, Germany.

10.1021/ci034223v CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/28/2004

Table 1. Data Sets Used in This Study

	data set I	data set II	data set III
screened compounds	5513	3685	1828
hits	185	118	67
nonhits	5328	3567	1761
hit fraction	3.4%	3.2%	3.7%

to ensure that structure−activity relationships, if present, could be found.

In the following, the entire set of tested compounds is denoted as data set I. To set up classification models which should be able to predict the antagonistic activity in terms of hits and nonhits, data set I was randomly split to form training and test data sets. In this process, care was taken that the ratio of hits to nonhits was about the same in the training and the test data sets. The training data set, denoted as data set II, comprises 3685 compounds, including 118 hits. The test data set, data set III, consists of 1828 compounds, with 67 hits. Table 1 summarizes the composition of the data sets used in this study.

METHODS

Structure Representation. To derive structure−activity relationships, molecular descriptors are required to encode molecular properties of compounds in a mathematical manner. Usually, the structures and associated compounds are represented by m-dimensional vectors $X(x_1, x_2, ..., x_m)$. These structure representations have to satisfy three basic conditions:

The translation of structural information into the vector X has to be unique.

The vector X has to have a fixed length m for all compounds in the data set.

The molecular descriptor X has to be invariant toward translation and rotation.

In our study, we used the well-known Daylight 2D fingerprints[4] and compared their performance with distance-based descriptors. Daylight 2D fingerprints belong to the class of binary descriptors, which originally were developed for database applications, such as similarity searches.[5] In addition, two distance-based methods were applied: autocorrelation vectors and radial distribution functions. The use of autocorrelation functions to encode molecular constitution was first introduced by Moreau and Broto,[6] a concept which was extended later on to molecular conformations and surfaces.[7] Topological autocorrelation functions $A(d)$ encoding the constitution of a molecule link the properties p_i and p_j of two atoms i and j by the distance-based δ-function, which becomes 1 if the distance d is equal to the number of separating bonds d_{ij} between i and j.

$$A(d) = \frac{1}{2}\sum_{\substack{i,j\\ i\neq j}} p_i p_j \delta(d - d_{ij}) \qquad (1)$$

The summation is performed over all atoms of a molecule. As already mentioned, autocorrelation vectors can also be used to encode three-dimensional objects, such as molecular conformations or even molecular surfaces. In these cases, d and d_{ij} are distances in 3D space, between either atoms or points defining the molecular surface. Since d is now continuous, an additional binning of d is required to translate the function $A(d)$ into a vector $A(d_n)$. Here, the sum is weighted by the number, L_n, of property pairs p_i and p_j assigned to bin n.

Table 2. Molecular Representations

representation	vector length	distance range
autocorrelation of molecular constitutions	16	0−15 bonds
autocorrelation of molecular conformations	24	0.0−23.0Å
atomic radial distribution functions	225	0.1−22.5Å
autocorrelation of molecular surfaces	25	0.0−24.0Å

$$A(d_n) = \frac{1}{2L_n}\sum_{\substack{i,j\\ i\neq j}} p_i p_j \qquad (2)$$

Atomic radial distribution functions have been applied extensively in studies of the correlation between molecular structures and infrared spectra.[8] They are strongly related to autocorrelation functions. Here, the properties p_i and p_j are linked by a Gaussian function. In contrast to the autocorrelation vectors, atomic radial distribution functions $g(r)$ are only used for encoding molecular conformations, not constitutions.

$$g(r) = \frac{1}{2}\sum_{\substack{i,j\\ i\neq j}} p_i p_j e^{-B(r-r_{ij})^2} \qquad (3)$$

To turn the function $g(r)$ into a vector, the values of $g(r)$ are only calculated at equidistantly distributed distances r_k with $k = 1, 2, ..., m$. If the parameter B, which controls the width of the Gaussian term, increases to infinity, it turns into the δ-function.

$$\lim_{B\to\infty} e^{-B(r-r_{ij})^2} = \delta(r - r_{ij}) \qquad (4)$$

Therefore, the autocorrelation functions encoding molecular conformations can be seen as a special case of atomic radial distribution functions.

To represent the hydantoin data set, autocorrelation vectors and atomic radial distribution functions were generated. Hydrogen atoms were only considered in the calculation of surface autocorrelation vectors. The parameter B of atomic radial distribution functions was set to 100. As shown in Table 2, these various representations differ in the length of the resulting vectors and in the distance range covered by d or r.

Prior to building autocorrelation vectors and atomic radial distribution functions, the following atomic properties were calculated by empirical methods. Descriptors to encode molecular constitution and conformation:

identity I, $p_i = 1$ for all atoms i
σ charge q_σ[9]
π charge q_π[10]
total charge q_{tot}
σ electronegativity χ_σ[11]
π electronegativity χ_π[10]
lone-pair electronegativity χ_{LP}[10]
atomic polarizability α[12]

These chemical descriptors ensured that polarity, electrostatic effects (charge values), and hydrogen bonding properties (electronegativity values) were taken care of. The polarizability descriptors are somehow related to size and

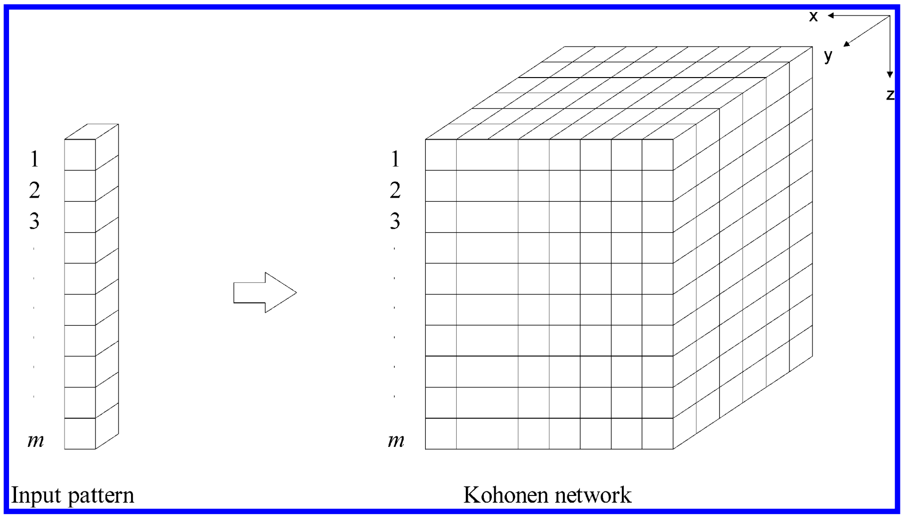

Figure 2. Block architecture of Kohonen neural networks. The vertically arranged neurons must have the same dimensionality as the input pattern of the training data set.

hydrophobic effects. Descriptors to calculate surface autocorrelation vectors:

Coulombic electrostatic potential ESP
hydrogen bonding potential HBP[13]
hydrophobic potential HYP[14]

Thus, all in all, 27 different distance-based representations were applied. All physicochemical properties were calculated by PETRA.[15] The 3D structures of molecules were generated by CORINA.[16] Autocorrelation vectors and atomic radial distribution function were formed by AUTOCORR,[17] SURFACE,[18] and ARC.[19]

In addition, three Daylight 2D fingerprint representations with lengths of 256, 512, and 1024 bits, respectively, encoding molecular fragments of up to seven bonds were generated. The binary fingerprints were used as a reference the distance-based descriptors had to compete with.

Self-Organizing Neural Networks. The aim of self-organizing neural networks is to create a low-dimensional map of a high-dimensional landscape in the way a cartographer does.[20] With the help of the resulting map, relationships among the data can be explored simply by visual inspection, which is obviously not feasible within a high-dimensional space. In other words, a projection of a set of data points defined in an m-dimensional space into an l-dimensional space

$$X(x_1, x_2, ..., x_m) \rightarrow Y(y_1, y_2, ..., y_l) \qquad (5)$$

with $m > l$ is performed in order to reduce the complexity of a given problem. Usually, the target space is two-dimensional, $l = 2$. During the projection, the topology of the input space X is preserved, which means that objects adjacent in m-dimensional space X are also neighbors in the l-dimensional output space Y. This key feature is achieved by a specific network topology (Figure 2) and its learning algorithm. A Kohonen network can be interpreted as a block of single boxes, each box representing a so-called weight of the network. Neurons W_j, with

$$W_j(t) = (w_{j1}(t), w_{j2}(t), ..., w_{jm}(t)) \qquad (6)$$

are built up by m of those weights and are arranged along the z axis to form a rectangular plane in x and y directions.

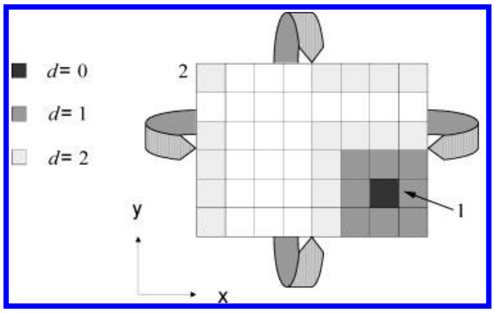

Figure 3. Since a toroidal surface of a Kohonen network has no real edges, the neuron 1 belongs to the second neighborhood area of neuron 2. Therefore, the topological distance between them is $d = 2$.

Throughout this study, only networks featuring a toroidal topology were applied. As illustrated in Figure 3, this means that neurons of opposite edges of the xy plane are considered to be next to each other. Joining these edges would then lead to a torus.

Kohonen networks learn in a time-dependent, iterative manner. At the beginning of the learning phase, when the number of processed cycles is $t = 0$, the weights $W_j(t)$ of the network are randomly initialized with real numbers. Two additional parameters affect the course of learning by controlling how the weights of the neurons are adapted: the learning rate $\eta(t)$ as a function of time and the learning span $\phi(d)$ as a function of topological neuron distances (eq 8). Several forms of learning rate and learning span function have been described,[21] but this topic will not be addressed in detail here.

One single training cycle consists of several steps:

1. A pattern $X(t)$ of the training data set is randomly selected and presented to the self-organizing network.

2. The activation of all neurons is calculated, and the neuron with the highest activation, the winning neuron, c, is determined. In this competition, the magnitude of activation is a function of the Euclidian distance between the pattern $X(t)$ and the neurons $W_j(t)$. The lower the distance, the higher the activation.

$$c \leftarrow \min\|X(t) - W_j(t)\| = \min \{\sum_{i=1}^{M}[x_i(t) - w_{ji}(t)]^2\}^{1/2} \qquad (7)$$

3. All neurons of the network are adapted in order to make them more similar to the input pattern $X(t)$, in which the magnitude of the adaptation decreases with number of processed cycles, t, and an increasing topological distance d_{jc} to the winning neuron c within the network.

$$w_{\mathrm{ji}}(t+1) = w_{\mathrm{ji}}(t) + \eta(t)\phi(d_{\mathrm{jc}})[x_{\mathrm{i}}(t) - w_{\mathrm{ji}}(t)] \qquad (8)$$

4. The stopping criteria defined before starting the training are checked. If the conditions are satisfied, learning is stopped; otherwise, it loops back to step 1. In the simplest case, a fixed number of cycles will be processed.

When the learning is finished, the information of the training data is stored in the neurons of the network. To achieve the projection of the m-dimensional data into the two-dimensional plane $Y(y_1, y_2)$, the entire training data set is sent again through the network. Each data pattern is then assigned to its winning neuron c and characterized by the coordinates y_1 and y_2. Patterns which are similar in the high-dimensional space of X should also be neighboring in two-dimensional space. They should have the same or at least similar coordinates, y_1 and y_2. The resulting projection can then easily be visualized if one or more additional properties of the training patterns are known. This can be either a real number or a discrete category. Each neuron of the network is colored according to its occupancy by data patterns. Several ways of color-coding are possible, for example, coloring by the mean, the highest, or the lowest property value occurring in a particular neuron.

In this study, we tried to find a relationship between structural properties of compounds and their biological activity toward a certain target. This means that patterns consisted of encoded structural information as already mentioned above, and the resulting self-organizing networks were then colored according to biological activity. By doing so, the strength of the relationship between the chosen structure representation and the biological activity can easily be assessed by visual inspection. Assuming a similar binding mode for all compounds, the most highly actives are expected to be grouped together within the self-organizing map.

RESULTS

Visual Inspection of Several Descriptor Behaviors. To determine which of the 30 structure representations is most closely related to the antagonistic activity in data set I, two-dimensional Kohonen networks were trained for each representation. Following the similarity principle, which states that similar compounds should have similar properties, the hit compounds should group together in a Kohonen 2D projection if the corresponding descriptors used in the learning process were relevant for activity.

The self-organizing networks comprised 45 × 60 neurons and were trained in 100 000 cycles using the Kohonen map generator KMAP.[22] To visualize the distribution of the hydantoin library compounds within the network, each neuron occupied by one or more hit compounds is represented in dark gray, whereas neurons containing only nonhits are colored light gray. Empty neurons remain white. The degree of correlation between structural representation and biological activity is derived by visual inspection of the maps.

Figure 4 shows the Kohonen maps obtained from distance-based molecular representations encoding the atomic properties I, q_π, q_σ, q_{tot}, χ_π, χ_σ, χ_{LP}, and α, respectively. All three sets of molecular descriptors, autocorrelation vectors of constitution and of conformation as well as atomic radial distribution function behaved similarly. None of the Kohonen maps shows the ideal picture of one coherent group of neurons comprising all hit compounds. Identity, polarizability, and electronegativities χ_π, χ_σ, and χ_{LP} gave the poorest results. Hits are spread over almost the entire Kohonen maps; no pronounced grouping can be recognized. Somehow, better results were obtained by using atomic charges, which led to several clusters of neurons with hits. The best result was obtained with autocorrelation vectors of molecular conformation and σ charges.

The Kohonen projections of surface autocorrelation vectors are shown in Figure 5. Here, the quality of the maps varies strongly, depending on the respective surface property. Using the electrostatic potential ESP, the hit compounds are distributed over the entire network. In contrast, several small groups of hit neurons can be recognized in the Kohonen map of the hydrophobic potential HYP. The best result was obtained using the hydrogen-bonding potential HBP. Hit compounds are projected into a small, compact region of neurons, indicating that a relationship between this structure representation and the biological activity exists.

Figure 6 shows the Kohonen maps generated with Daylight 2D fingerprints. Independent of the length of the bit vectors, no grouping of hit compounds can be found. These maps are of a similar poor quality as those obtained by distance-based representations of identity, polarizability, and electronegativities. In the particular study described here, these descriptors do not correlate with biological activity.

It might be anticipated that combinations of different types of molecular descriptors would even improve the quality of the projection, as compared to the Kohonen maps of HBP. Therefore, we constructed new 75-dimensional descriptors by concatenating the surface autocorrelation vectors of ESP, HYP, and HBP. The resulting Kohonen map is shown in Figure 7. Due to the combination with the poorly performing descriptors ESP and HYP, the separation power of HBP gets lost. In contrast to supervised methods, which weight the single components of the descriptor vector and try to consider only relevant descriptors, the Kohonen method treats each descriptor component equally. A combination of well and poorly performing descriptor vectors is not recommended when applying Kohonen networks. Therefore, this study is solely focused on single-type descriptor vectors.

From 2D Projection to Classification Models. Although the results of projection experiments presented here are unambiguous in finding the best molecular representation explaining the biological activity of the hydantoin library, a general shortcoming of the method is that only a qualitative assessment of the results is possible, which still depend on the individual observer. To address this problem, a classification method based on Kohonen neural networks was established. In our study, the aim of such a classification system is to predict whether a compound totally unknown to the system belongs to the class of hits or the class of nonhits. In contrast to other methods that include the class information in the derivation of the models, for example, binary QSAR,[23] Kohonen networks are set up in an un-

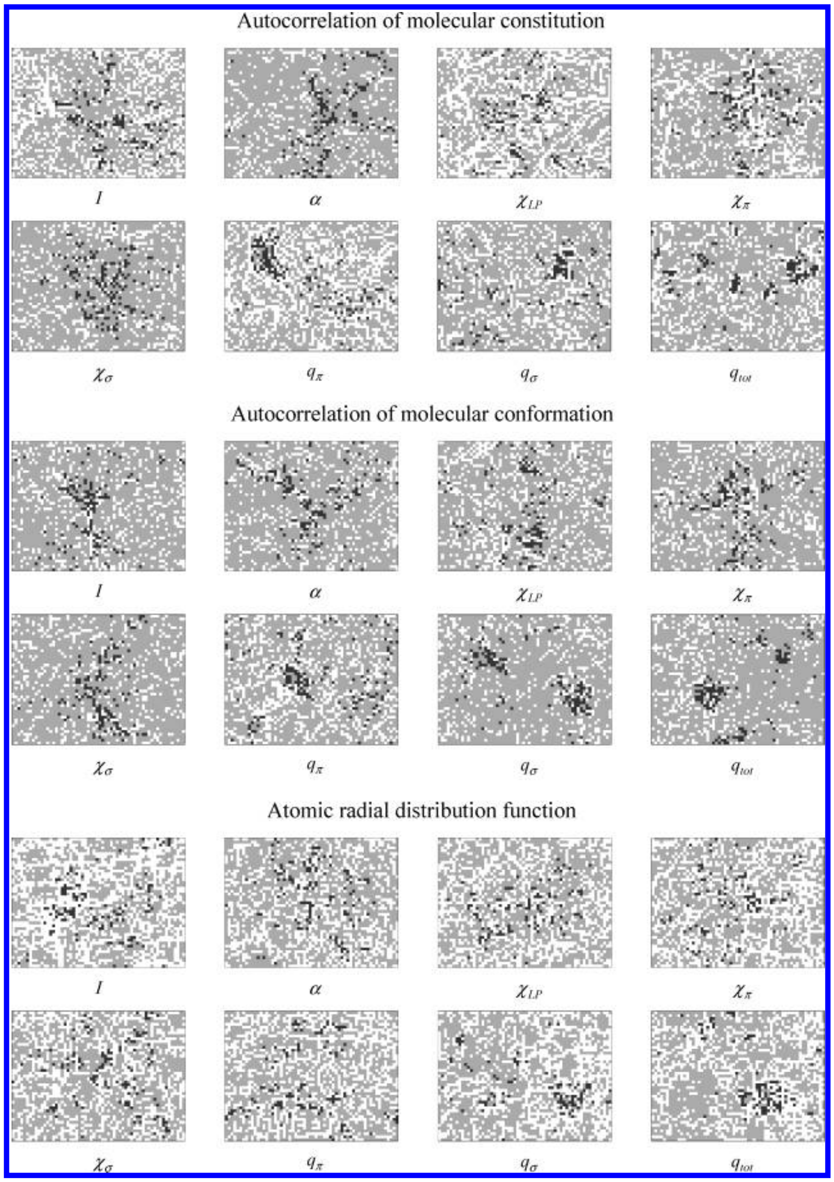

Figure 4. Kohonen maps created from distance-based molecular descriptors encoding the atomic properties I, q_π, q_σ, q_{tot}, χ_π, χ_σ, χ_{LP}, and α. Neurons containing one or more hit compounds are shaded dark gray, neurons occupied exclusively by nonhit compounds are light gray and empty neurons are indicated white.

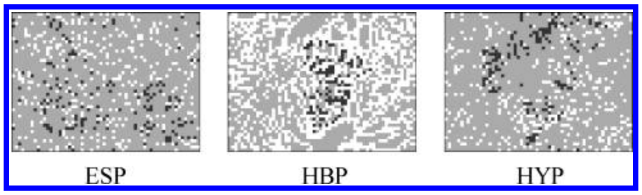

Figure 5. Kohonen maps derived from surface autocorrelation vectors of Coulombic electrostatic potential, ESP; hydrogen-bonding potential, HBP; and hydrophobic potential, HYP.

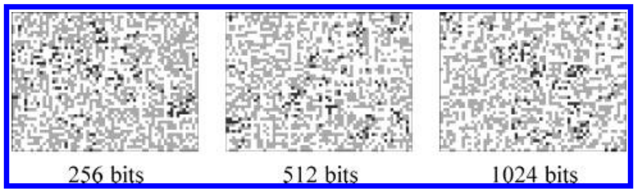

Figure 6. Kohonen maps of Daylight 2D fingerprint representations with bit lengths of 256, 512, and 1024, respectively.

supervised learning procedure. Since no activity data affects the result of the Kohonen projection, any kind of over-fitting pretending structure−activity relationships is avoided.

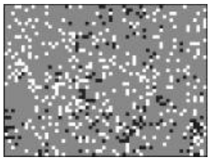

Figure 7. Kohonen map derived from a descriptor composed of surface autocorrelation vectors of Coulombic electrostatic potential, ESP; hydrogen-bonding potential, HBP; and hydrophobic potential, HYP.

To classify the compounds, self-organizing maps have to be trained to establish classification rules from the occupation of the neurons. Four different rules, which are illustrated in Figure 8, were investigated:

Rule 1: Compounds falling into neurons that contain more hit than nonhit compounds are classified as hits. If a neuron comprises the same number of hits and nonhits, compounds are predicted to be nonhits.

Rule 2: If a compound is assigned to a neuron that contains more hits than nonhits or a neuron next to it, the compound is classified as a hit.

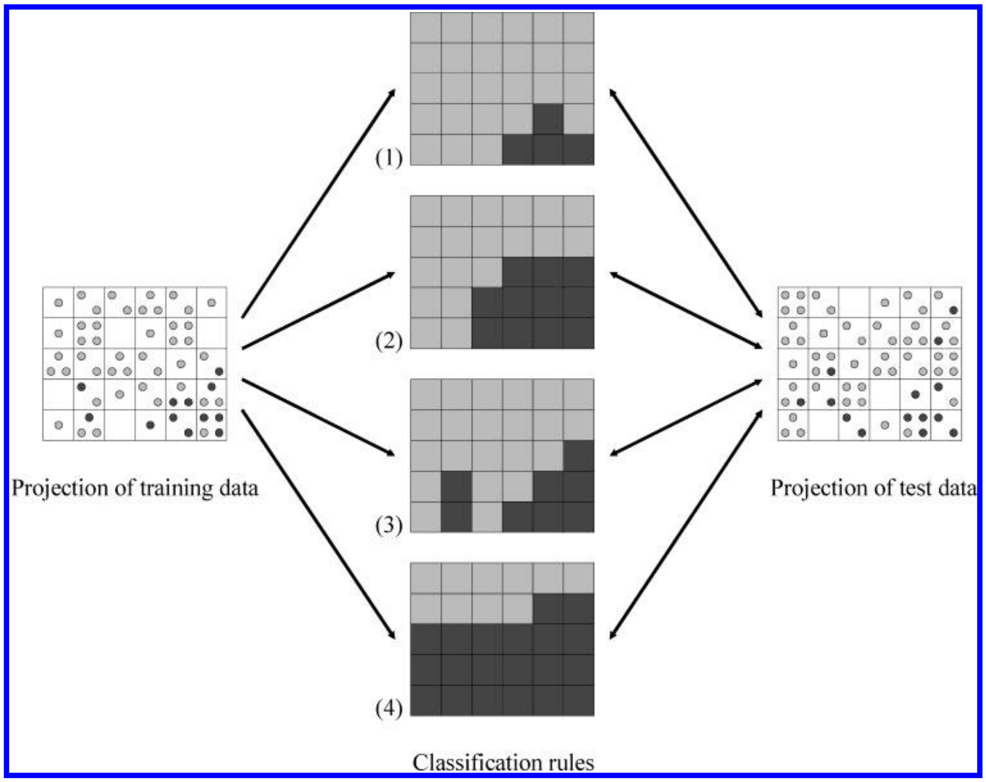

Figure 8. Classification rules 1−4 illustrated for a hypothetical distribution of hits and nonhits in a small area of the Kohonen networks. The rules were derived from the projection of the training data set. Compounds falling in neurons shaded dark gray are classified as hits, whereas compounds assigned to light gray neurons are classified as nonhits. The quality of the classification model can be assessed by the predictions of the test data set.

Rule 3: If a neuron contains at least one hit compound, every compound assigned to that neuron is predicted to be a hit.

Rule 4: In extension to rule 3, compounds are also classified as hits if they fall into neurons next to one which is occupied by at least one hit compound of the training data set.

Each of the molecular representations introduced above, which have been investigated by visual comparison of Kohonen maps, was subsequently used for deriving classification systems. Thirty self-organizing networks of 48 × 38 neurons each were trained in 100 000 learning cycles using data set II containing only two-thirds of the investigated compounds of the hydantoin library. From the resulting 2D projections, the four classification rules have been applied to create classification models. The performance of the models to predict the membership of the hit or nonhit class of compounds not included in the training data was tested using the remaining part, data set III, of the hydantoin library. Data set III was sent through the networks, and the winning neurons of the test compounds were determined. Table 3 summarizes the results of applying the four classification rules.

Since classification rule 1 is the most rigorous one, it performs the poorest in predicting class membership. Almost all nonhits are recognized with each descriptor, whereas the number of perceived hits is small. The quality of the predictions based on rules 2 and 3 are comparable. The barrier to categorize a compound as a hit is lower, as compared to rule 1. As a consequence, the quantity of predicting hits increases and that of nonhits decreases. This trend is even more pronounced with rule 4, which is designed

Table 3. Accuracy of the Derived Classification Models Depending on Molecular Representation and the Applied Classification Rule[a]

representation	property	rule 1	rule 2	rule 3	rule 4
autocorrelation of molecular constitutions	I	10/99	34/94	43/96	82/76
	α	3/99	42/91	22/96	79/73
	χ_{LP}	19/99	46/91	52/95	76/79
	χ_π	10/99	39/93	46/94	79/75
	χ_σ	4/99	22/89	28/94	78/70
	q_π	21/99	55/96	46/88	88/73
	q_σ	19/99	43/93	55/96	81/80
	q_{tot}	19/99	61/92	55/96	87/83
autocorrelation of molecular conformations	I	27/100	61/90	37/96	84/77
	α	16/99	49/88	37/95	81/76
	χ_{LP}	18/99	43/90	52/95	78/75
	χ_π	16/99	42/89	48/95	78/75
	χ_σ	15/100	57/94	37/95	79/77
	q_π	21/99	48/92	46/92	81/72
	q_σ	22/99	69/92	51/96	88/81
	q_{tot}	16/99	70/93	48/96	90/84
atomic radial distribution functions	I	10/99	42/93	57/94	91/78
	α	19/99	42/91	48/93	88/72
	χ_{LP}	13/99	42/95	70/91	87/76
	χ_π	12/99	37/93	57/94	82/75
	χ_σ	27/100	45/94	63/93	88/76
	q_π	18/99	40/97	58/87	82/75
	q_σ	28/99	64/91	69/94	91/82
	q_{tot}	24/99	60/93	60/95	90/83
autocorrelation of molecular surfaces	ESP	6/99	21/91	25/94	72/64
	HBP	21/99	63/96	66/96	96/92
	HYP	9/99	33/92	36/95	87/75
Daylight 2D fingerprints	256-dim.	9/100	30/95	45/89	81/72
	512-dim.	13/99	33/95	54/90	85/73
	1024-dim.	15/100	31/97	46/89	84/71

[a] In percent correctly predicted hits vs percent correctly predicted nonhits.

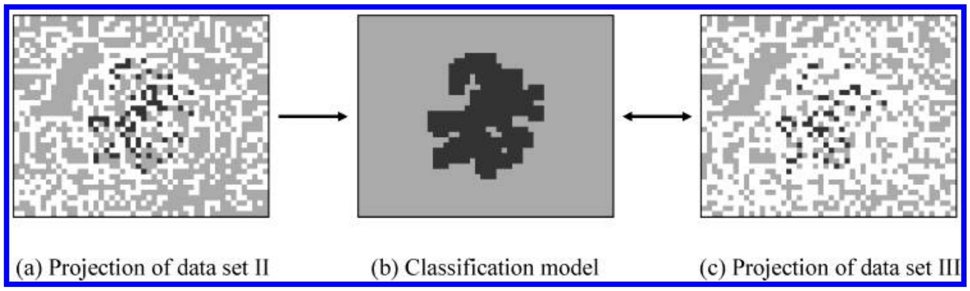

Figure 9. Classification model derived from the projection of data set II represented by surface autocorrelation vectors of hydrogen bonding potential. The model and the Kohonen map of data set III, which serves as a test data set, have almost the same shape, which results in the high accuracy of the model.

to identify hit compounds at the expense of nonhit compounds.

Comparing the performance of the single molecular representation over all classification rules, the picture established by visual analysis of the self-organizing maps is confirmed. In the combination of distance-based representations with atomic properties, identity, polarizability, electronegativities χ_π, χ_σ, χ_{LP}, and π-charges gave the poorest results. The Daylight 2D fingerprints provide only poor results. Extending the bit length of the fingerprints does not lead to a significantly better classification accuracy. This indicates that bit collisions—instances in which different molecular paths share the same bits—have at most minor impact on the results. The autocorrelation vectors of the electrostatic and the hydrophobicity potential on the molecular surfaces perform similar to the Daylight fingerprints. The best result could be obtained with surface autocorrelation vectors of the hydrogen bonding potential. Applying classification rule 4, 64 of 67 (96%) of the hits in data set III could be identified. In addition, only 142 (8.1%) of 1761 nonhits were misclassified.

This classification model retrieved by surface autocorrelation vectors is illustrated in Figure 9. The Kohonen map (a) is obtained by projecting data set II into two-dimensional space. Applying rule 4 yields the classification model (b). Here, the reason for the good classification power of this model becomes obvious: The dark gray hit neurons occupy only a small, compact area of the entire network, and thus, the number of false positives remains small. Kohonen map (c) shows the projection of the test data set III. Obviously, the hit compounds are assigned to neurons of the hit area of the classification model. Furthermore, the similar shape of hit distribution in maps (a) and (c) can be interpreted as a sign for a good structural homogeneity and a comparatively high quality of the biological data.

The reason classification rule 4 performs significantly better than the others becomes obvious by inspecting the classification maps of the surface autocorrelation vectors of HBP shown in Figure 10. Two parameters control the classification accuracy: the size of the neighborhood considered in the classification and the condition to be satisfied to classify a compound as a hit. If the activity class is derived only from occupation of the neuron a compound is assigned to, one important feature of Kohonen networks is neglected: neighboring neurons affect each other. Moreover, the classification maps generated by applying rules 1 and 3 contain a large number of empty neurons. The activity class of compounds falling into these neurons cannot be predicted. Here, additionally considering the first shell around occupied neurons appears to be a reasonable compromise with respect to the network size. By applying rules 1 and 2, the predicted class is derived from the majority of the activity classes present in the neurons. This procedure does not reflect the imbalance of hits and nonhits in the data set investigated. Even by taking into account the first shell of neighboring neurons, the number of neurons predicting compounds to be hits is still too small to finally identify hits in the test data set.

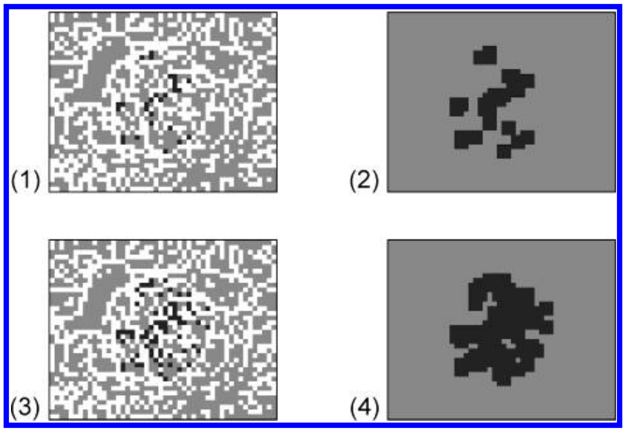

Figure 10. Classification maps of the hydrogen bonding potential HBP derived by applying classification rules 1−4.

Robustness with Respect to Hit Thresholds. The concentration of test compounds plays a significant role in the design of a high-throughput screening experiment. Often, due to solubility problems, a high screening concentration cannot be achieved for every compound in the library. On the other hand, if the screening concentration for a library is reduced, the hit criterion has to be shifted toward higher % ctrl values, to identify the same number of hits. With such an approach, there is a risk that also the number of false positives can increase, as compared to the results at a higher screening concentration.

In an application of the classification model derived so far, the question was addressed to what extent the hit criterion can be raised without loosing the correlation between the structure represented as autocorrelation vectors of the hydrogen bonding potential and the antagonistic activity.

The original screening concentration of the hydantoin library was 25 μg/mL. In addition, for 5377 compounds of data set I, the screening was also performed at a concentration of only 5 μg/mL. As a basis for investigating the impact of the hit threshold, the Kohonen neural network derived with data set I was used. After determining the winning neurons of the 5377 hydantoin compounds, classification models at different hit criteria were derived following classification rule

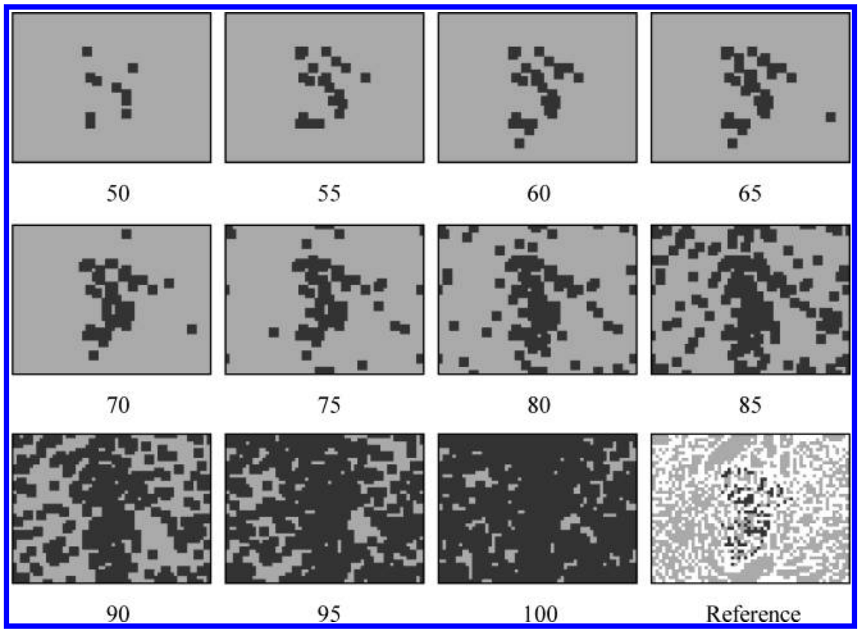

Figure 11. Classification rules of data set I described by surface autocorrelation vectors of hydrogen-bonding potential. The hit threshold at a screening concentration of 5 μg/mL is increased from 50 to 100% ctrl. For larger hit thresholds, the gray shaded hit neurons finally dominate the classification system, as compared to the reference map at 25 μg/mL screening concentration.

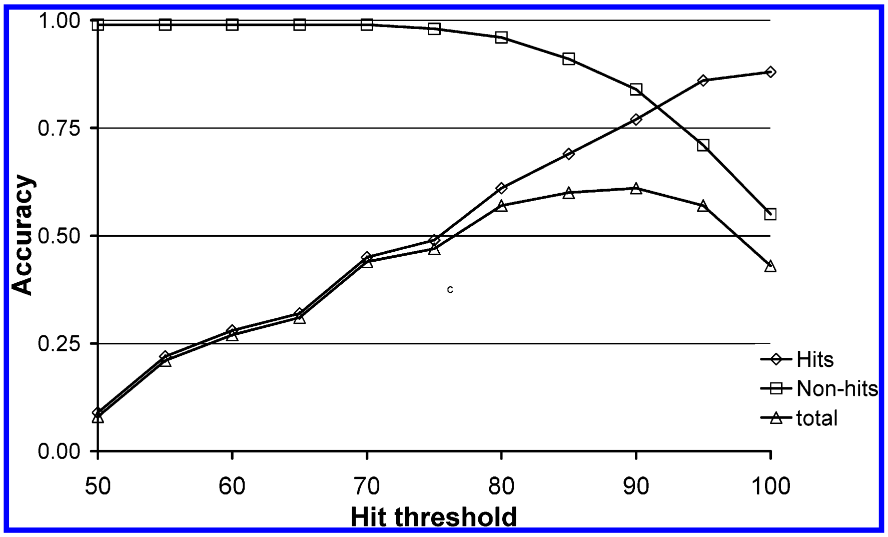

Figure 12. Quality of the classification models derived at 5 μg/mL screening concentration for predicting activity at a screening concentration of 25 μg/mL.

4. As shown in Figure 11, the threshold separating hits from nonhits was increased in steps of 5% ctrl up to 100% ctrl. The comparison of the new classification models with the original projection of data set I shows that up to a threshold of 60% ctrl, the arrangement of hit neurons is reproduced. At a level of 65% ctrl, the first group of hit neurons appears that cannot be found in the reference projection at 25 μg/mL. By further increasing the threshold, the number of neurons classifying compounds as hits also increases. Parallel to this observation, the structure−activity relationship recognizable in the central grouping of hit neurons is getting weaker.

The classification models shown in Figure 11 were applied to data set I in order to quantify the amount of correctly predicted screening results at 25 μg/mL. The dependency of the prediction quality on the hit criterion is plotted in Figure 12. Due to the increasing number of neurons classifying compounds as hits, the number of correctly predicted hits grows with an increasing hit threshold. Up to a threshold of 70% ctrl, 99% of the nonhits are identified. This value then decreases with a further increase of the threshold. The curve of the total accuracy shows a plateau between 80% ctrl and 95% ctrl. The maximum is located at a hit criterion of 90% ctrl with 61% correct predictions.

This study confirms that the results of a high-throughput screening at a higher concentration cannot be easily compared with results at a lower concentration. In our case, a hit threshold between 70% ctrl and 80% ctrl for a screening campaign at 5 μg/mL seems to be ideal in order not to loose the structure−activity relationship present in the data generated at a screening concentration of 25 μg/mL. In this case, between 45 and 69% of hits are identified at both concentrations using the derived classification models. For these thresholds, the number of false positive hits varies between 1 and 9%.

CONCLUSIONS

Kohonen's self-organizing network can provide insight into the relationships between molecular structures, represented via high-dimensional descriptors, and an associated biological activity in an unsupervised manner. By a simple visual inspection of the projections of high-dimensional data into a two-dimensional plane, molecular descriptors can be identified that have a nonlinear linkage to biological activity. Thus, appropriate physicochemical properties can be chosen to calculate autocorrelation vectors or atomic radial distribution functions for encoding molecular structures.

In this study, we were able to apply the technique of self-organized learning to a combinatorial library of hydantoins. We identified the hydrogen-bonding potential on the molecular surface as the molecular representation which best describes the relationship between compounds and biological activity. On the basis of this selection, we could set up a classification model for predicting antagonistic activity with high accuracy.

We applied the model to the question whether a variation of the compound concentration in high-throughput screening together with a shift of the hit threshold has an impact on the established structure–activity relationships.

The derived model can be used to address additional problems. One such application is the virtual screening of compounds in order to design focused libraries. Furthermore, by retesting all compounds that fall into hit neurons, the identification of false negatives within the hydantoin library is conceivable. The approach outlined here has greatly reduced the number of compounds that have to be retested.

ACKNOWLEDGMENT

This work was performed as part of the SOL project, which is funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) under Grant No. 0311680. The authors thank Dr. Herbert Köppen and Dr. Stefan Peters from Boehringer Ingelheim Pharma GmbH & Co. KG for helpful discussions and their contributions to this work.

REFERENCES AND NOTES

(1) Anzali S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J. Sadowski, J.; Teckentrup, A.; Wagener, M. The Use of Self-Organizing Neural Networks in Drug Design. In *3D QSAR in Drug Design*; Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, Netherlands, 1998; Vol. 2, pp 273–299.

(2) Kohonen, T. Self-organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59–69.

(3) Kohonen, T. *Self-Organization and Associative Memory*, Springer-Verlag: Berlin, 1989.

(4) Daylight Chemical Information Systems, Inc., 27401 Los Altos, Mission Viejo, CA 92691, http://www.daylight.com.

(5) Barnard, J. M. Structure Representation. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, III, H. F.; Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; pp 2818–2826.

(6) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures. *Nouv. J. Chim.* **1980**, *4*, 757–764.

(7) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteriod Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.

(8) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3D Stucture of a Molecule in its IR Spectrum. *Fresenius' J. Anal. Chem.* **1997**, *359*, 50–55.

(9) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.

(10) Gasteiger, J.; Saller, H. Calculation of Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem., Int. Ed. Eng.* **1985**, *24*, 687–689; *Angew. Chem.* **1985**, *97*, 699–701.

(11) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity—An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541–2544.

(12) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Application to Studies of X-ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559–564.

(13) Vedani, A.; Huhta, D. W. A New Force Field for Modeling Metalloproteins. *J. Am. Chem. Soc.* **1990**, *112*, 4759–4762.

(14) Heiden, W.; Moeckel, G.; Brickmann, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 503–514.

(15) Marsili, M.; Saller, H.; Hutchings, M. G.; Fröhlich, A.; Gasteiger, J. PETRA; University of Erlangen-Nuremberg, 1995.

(16) Sadowski, J.; Schwab, C. H.; Gasteiger, J. CORINA, version 2.4; University of Erlangen-Nuremberg, 1998; available from Molecular Networks GmbH, Erlangen; http://www.mol-net.de.

(17) Sadowski, J.; Gasteiger, J. AUTOCORR, version 1.0; University of Erlangen-Nuremberg, 1994; available from Molecular Networks GmbH, Erlangen; http://www.mol-net.de.

(18) Sadowski, J.; Gasteiger, J. SURFACE, version 1.0; University of Erlangen-Nuremberg, 1994; available from Molecular Networks GmbH, Erlangen; http://www.mol-net.de.

(19) Hemmer, M. C.; Gasteiger, J. ARC, version 1.1; University of Erlangen-Nuremberg, 1998.

(20) Zupan, J.; Gasteiger J. *Neural Networks in Chemistry and Drug Design;* Wiley-VCH: Weinheim, Germany, 1999; Vol. 2..

(21) Dayhoff, J. *Neural Network Architectures*. Van Nostrand Reinhold: New York, 1990.

(22) Li, X.; Wagener, M.; Teckentrup, A.; Gasteiger, J. KMAP, Version 4.0; University of Erlangen-Nuremberg, 1999. An extended version, SONNIA, is available from Molecular Networks GmbH, Erlangen; http://www.mol-net.de.

(23) 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.

CI034223V