Assessing the Predictive Power of Unsupervised Visualization Techniques to Improve the Identification of GPCR-Focused Compound Libraries

Modest von Korff* and Kurt Hilpert

Actelion Pharmaceuticals Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland

Received February 1, 2006

Principal component analysis and self-organizing maps (SOMs) were compared to cluster and visualize the chemical space of a large and diverse data set. The data set comprised about 3000 G-protein-coupled receptor (GPCR) ligands for about 130 receptors and 3000 non-GPCR ligands from the World Drug Index. The molecules were described with a topological pharmacophore point histogram descriptor and a chemical fingerprint descriptor. To assess the predictive power of the clustering, a leave-multiple-out cross validation with *k* nearest neighbor classification was performed. The results of the classification tests and the visualization showed a clear superiority of the SOM method. SOM correctly divided the data set into two main clusters, one for the GPCR and the other for the non-GPCR ligands. Our results suggest that a continuous GPCR-ligand space exists.

INTRODUCTION

Slides such as those shown in Figure 1 are often presented in the desire to visualize the chemical space of small organic molecules in two or three dimensions.[1] The implicit hypothesis of such a picture is to visualize a correlation between the target type and the representation of the molecules in low dimensional space. Such clustering is appropriate in the trivial case where molecules are grouped according to their target types without regard for their chemical properties. However, what medicinal chemists need is a clustering considering both the target type and the molecular features. The clustering of molecules according to the target type is one of the main subjects in chemoinformatics. Many tools have been developed and evaluated for this purpose.[2−6] The visualization of the clustered objects is tightly connected to the clustering itself. We compared the predictive power of two of the most frequently used clustering and visualization methods to visualize molecular features space: the principal component analysis (PCA)[7−14] and the Kohonen maps (self-organizing maps, SOMs).[15−21] A visualization of the chemical space in relation to the target type provides the medicinal chemist with a valuable tool to rapidly overview the "biochem space". We decided to use unsupervised clustering techniques, because unsupervised learning has the advantage of learning larger and more complex models. In supervised learning, one is trying to find a connection between two sets of observations. This connection introduces a bias to the created model. If there is no real relationship between the observations, that is, between the descriptor and the target type, the model fails. In our study, unsupervised means that the clustering algorithm has no knowledge about the receptor class of the ligands. So, the information has to be taken only from the ligand structure and not from the target receptor. This lack of information should guarantee the independence of the clustering method.

* Corresponding author phone: +41 61 5656323; e-mail: modest.korff@actelion.com.

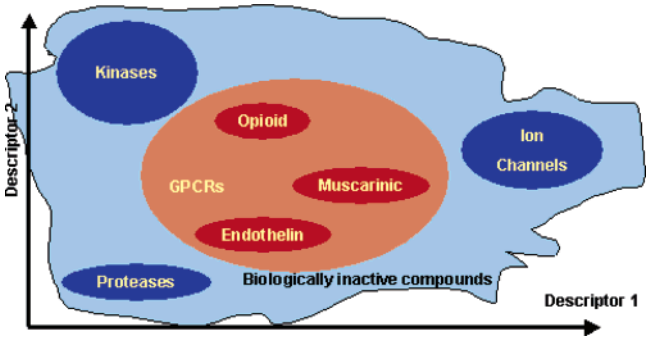

Figure 1. Imaginary projection of the "chemical space" related to "biological activity space".

An important starting point in modern drug discovery is high-throughput screening (HTS). Large compound collections are often acquired for subsequent use in HTS. However, the selection of appropriate compound collections for the screening has a major impact on subsequent drug discovery and development. Because the chemical space is huge and the HTS resources are limited, such molecules have to be selected in a way that promises the highest probability of finding interesting hits. For this aim, the selection has to be target-focused and diverse enough to find new scaffolds. Clearly, for such a selection process, the support by visualization techniques is quite helpful.

The chemical space we want to describe in two dimensions is, in fact, multidimensional. First, it comprises the topological information, given by the molecular graph. Then, we have to consider the conformer space, which contains the 3D structural information. Together with the selected metrics, these multiple dimensions span the similarity space. An additional dimensionality is given by the behavior of the molecules in the target type space. This dimension describes the interaction of the ligand molecules with the proteins. According to the enhanced induced fit theory, the ligand and the protein change their conformation during the approach of the ligand to the protein.[22] At the end of the recognition

10.1021/ci060037o CCC: $33.50
Published on Web 04/22/2006

process, the molecule fits to the binding site of the protein. All of these multidimensional and dynamic relationships have to be considered for an exact visualization of the biochem space. Many of these parameters are unknown or difficult to model; this determines the limit of performance of each visualization method. Furthermore, it is not possible to generate one single general model. The reduction of the molecular recognition process to a ligand-based approach enforces the use of local models. The ligand-based local model relies on molecular similarity. The assumption that similar ligand molecules will have similar binding affinities to the protein is widely accepted.[4,23−25] The challenge remains in finding a molecular descriptor that describes the molecular similarity as nearly as possible from the viewpoint of the target receptor.

Ligand-based modeling approaches often take into account only topological information. The reason for this is the increase of complexity required to model additional dimensions. Yet, recent approaches in ligand-based design show no real superiority of more complex models over the 2D approaches.[3,26,10]

For the visualization, the multiple dimensions of the descriptor have to be projected into two or three dimensions. Such projection should preserve the original patterns in the lower dimension and, even more, enable the human eye to detect useful patterns in the projected space.

METHODS

Data. The data set under consideration was a proprietary G-protein-coupled receptor (GPCR) ligand data set with more than 3000 molecules targeting 130 different receptors.[27] The data have been acquired from the literature and by in-house experiments. For comparison, a diverse data set of equal size, without molecules known as active on GPCR, was assembled with compounds from the World Drug Index.[28] This was done by a random selection of molecules that display their biological activity at a variety of other targets, such as enzymes (kinases, phosphatases, proteases, etc.), transporters, channels, and other receptors (nucleotide, integrin, or hormone receptors, etc.) or mechanisms, for instance, transcription, translation, protein modification, and so forth.

Two structure databases with HTS libraries from different suppliers served as databases for analytic purposes. The databases are abbreviated in the following as DB_A and DB_B. For DB_A, the supplier claims a GPCR-targeted library, whereas DB_B is a not particularly specified library.

The supplier molecules were filtered with our substructure-based proprietary in-house filtering tools to remove inappropriate molecules. The filter criteria applied are druglikeness,[29,30] toxic structure patterns,[31] and reactive chemical groups. After filtering, about 11 000 molecules for DB_A and 2300 molecules for DB_B remained.

For comparison reasons, we compiled a publicly available benzodiazepine data set with 245 structures.[32]

Descriptors. ActelionFp. To search in large molecular databases for identical and similar molecules, Actelion has developed a proprietary dictionary-based fingerprint.[33]

Pharmacophore Fingerprint (PFp2D). The PFp2D descriptor is closely related to the atom pair descriptor[34] and the binding property pair's descriptor.[35] For the detection of the pharmacophore points, we use a combination of the defined substructures and logic expressions. Preliminary experiments resulted in the selection of the following pharmacophore types within a molecule: hydrogen-bond donor (d), hydrogen-bond acceptor (a), hydrophobic (h), and aromatic (r). The pharmacophore points are put into a relationship by the histograms of the topological distance counts. To generate the histogram, the topological distance between each pair of atoms belonging to a certain pharmacophore type is counted. The maximum topological distance was set to 12 bond lengths. The counts are added to the histogram for the pharmacophore point pair combination. We do this for each two-point pharmacophore point combination, and all resulting histograms from one molecule are written into a descriptor vector. In contrast to the ActelionFp, the PFp2D descriptor captures the relative arrangement of the functional groups in the molecule.

Dimension Reduction Techniques. Both techniques used are well-known in the scientific community. Because the central aspect of this work is a comparison of the significance and the performance of these methods, a short description is nonetheless briefly given here.

Principal Component Analysis. PCA is a linear projection technique, whereby data vectors are projected into a lower dimensional space.[36] A descriptor matrix ${}_m\mathbf{A}^n$ is transformed via the covariance matrix **S** into p eigenvectors and their corresponding eigenvalues. The variance from the original matrix **A** is summarized in the eigenvectors in the way that the vectors are orthogonal to each other. The eigenvectors are arranged in decreasing order of explained variance. The first latent variable consequently contains the largest amount of variance and, therefore, the largest portion of information from the original variables. The principle components are the eigenvectors of **S**. With the principal components, the original descriptor variable space can be transformed into a new latent variable space. These latent variables, so-called "scores", are used as a new descriptor set. The scores can be computed for an arbitrary number of principal components. If all of the significant principal components are used to calculate the scores, the scores are identical to the original descriptor variables. If variables in the original matrix **A** are linearly dependent, their variance can be summarized in one principle component. Because of the fact that multivariate chemical descriptors often contain redundant information, the number of significant principal components is less than the number of original variables. Without a loss of information, principal components can be used to collect the information from the original variables in fewer scores. Conversely, the selected principal components determine the information content in the latent variable space.

If it is possible to collect the significant information in the first two or three latent variables, the result can easily be visualized in a two- or three-dimensional coordinate system. This is called principle coordinates analysis.

Self-Organizing Maps. SOMs or Kohonen maps belong to the class of unsupervised nonlinear multidimensional mapping tools.[15] In contrast to the linear multidimensional mapping methods such as PCA, SOMs are able to preserve the underlying nonlinear relationships in a data set. Another important point for our decision was that the generation of SOMs is not supervised. That means the training algorithm knows nothing about the class membership of a compound. So, the algorithm does not waste any degrees of freedom by

correcting the result with information from the class memberships.

The SOM consists of an array of vectors. The number of rows and the number of columns define the dimensions of the SOM. The field entries are the weight vectors. They are the templates for the classification of the descriptors. Hence, the weights and the descriptor vectors must be of the same length. To generate a relationship between the weight vectors and the data set under consideration, the SOM has to be trained. For this purpose, a training set of data is selected. For the training set, the descriptors are calculated. In the training set, the maximum and minimum values for each descriptor are determined. Then, the SOM is initialized with equal distributed random numbers. The maxima and minima are equal to the maxima and minima values of the descriptor vectors. The training descriptor vectors are now shuffled. The first one is taken, and the most similar weight vector on the SOM is determined. Then, this central weight is adapted toward the descriptor vector of the selected training object with an adaptation function. All weights, which are localized in the first sphere around the central weight, are now adapted with the adaptation function. According to the adaptation function, the adaptation in the sphere is less strong than that for the central weight. This is done for all spheres until a user-defined radius r is reached. The adaptation and r are lowered with each training sample that is given to the map until the training phase ends. For the calculation of the adaptation, we use a toroidal SOM topology. The array with the weights is projected onto a torus. This projection has the advantage that the calculations of the adaptation are performed on an endless plane. In this way, we avoid border effects. For the visualization, the torus is projected onto a rectangular plane. The left side of the plane proceeds then on the right side of the plane, and the upper side of the plane proceeds on the lower side of the plane.

After the SOM is trained, the success can be measured either by visualization or by an objective function. For the visualization, the training data are mapped onto the SOM. For each training object, the most similar weight is determined. Then, this position is visualized by a colored point. The color stands for the class membership of the object. The majority of objects determines the color if more than one object is mapped to the same weight. For our studies, we use a quadratic SOM with a side length of 100. The resulting number of nodes is higher than the number of training objects. This means that the trained SOM contains a lot of unoccupied weights. These weights contain valuable information. If the descriptor of a molecule is solely localized on a weight and the neighbor weights are unoccupied, there are no ligands with similar descriptors around the ligand under consideration.

One disadvantage of SOMs is that they are not reproducible. The random shuffling generates for each SOM run a new pattern. The reason lies in the existence of an almost infinite number of possible solutions for the multidimensional mapping problem. And there exists, depending on the size of the SOM and the structure of the training data set, a large number of solutions (patterns) with equal quality. But if the adaptation rate and radius for the SOM training are not set too high, the final patterns are similar in terms of explanatory power.

Similarity Measurements. For the similarity comparisons, we used the inverse Tanimoto coefficient.[37] The inverse Tanimoto coefficient is calculated by subtraction from 1. A distance of zero means that two descriptors are identical; a distance of 1 is the maximum difference. The Tanimoto coefficient has proved to be a powerful distance metric for descriptor comparisons.[38−40] For nonbinary vector comparisons, the XOR operator is replaced by the dot product of the two vectors.

Validation and Training Data. The PCA and the SOM algorithm do not use any response information, and therefore, the result of the training data set classification is already meaningful, even though the model is biased by the training data used for its generation. To assess the true modeling power of the models, it is necessary to use an independent validation data set. For this reason, we used "leave multiple out cross-validation".[41] The GPCR data set was split by random into two equal-sized data sets, $GPCR_{valid}$ and $GPCR_{train}$. The same procedure was performed with the non-GPCR data set, resulting in $NonGPCR_{valid}$ and $NonGPCR_{train}$. $GPCR_{train}$ and $NonGPCR_{train}$ were then combined to the $GPCR-NonGPCR_{train}$ data set. In the same way, we created the $GPCR-NonGPCR_{valid}$ data set. From the training data, the models were generated. First, the descriptors for validation and training data are calculated. The descriptor vectors from $GPCR-NonGPCR_{train}$ are stored as row objects in a matrix $\mathbf{A}_{train}$; the descriptor vectors from $GPCR-NonGPCR_{valid}$ are stored in a matrix $\mathbf{A}_{valid}$. This procedure was repeated 11 times.

Classification. To assess the quality of the clustering, we implemented a *k*-nearest-neighbor-objective function (*k*NN).[42] For each object in a class, the class membership is predicted. The classification quality is expressed by the ratio between the number of correctly predicted class memberships and the number of objects in the class under consideration. This is done for all classes. To assess the performance of the PCA and SOM algorithms, we used the variables from the reduced (512 to 2, 120 to 2) dimensional space as descriptors. The transformation algorithms PCA and SOM are applied to $\mathbf{A}_{train}$. The resulting models are the eigenvectors in the PCA and the map with the weight vectors in the SOM algorithm. In the PCA, the models are used to project the descriptors of $\mathbf{A}_{valid}$ into the latent variable space; in the SOM, the position on the map for each descriptor vector in $\mathbf{A}_{valid}$ is determined by searching the most similar weight vector. To assess the quality of the model, we took the first two latent variables from the PCA because the position on the SOM is given by two variables. These two variables serve as new descriptors for the matrices $\mathbf{A}_{train,proj}$ and $\mathbf{A}_{valid,proj}$. With the *k*NN classification algorithm for each object $\mathbf{a}_{valid,proj,i}$ in $\mathbf{A}_{valid,proj}$, the k most similar objects are selected from $\mathbf{A}_{train,proj}$. The class membership of $\mathbf{a}_{valid,proj,i}$ is then assumed to be the class membership of the majority of the selected objects. The quality value is the fraction of correctly predicted objects. The generation of the principle components and the generation of the SOM were repeated with all 11 training data sets.

Implementation. All implementations were done in Java. For the pharmacophore fingerprint, the GenerateMD class from ChemAxon was used.[43] The ActelionFp is an in-house development,[33] as well as the DataWarrior, our tool for data exploration and visualization developed at Actelion.[44] The SOM, the PCA, and the statistic functions were implemented by ourselves.

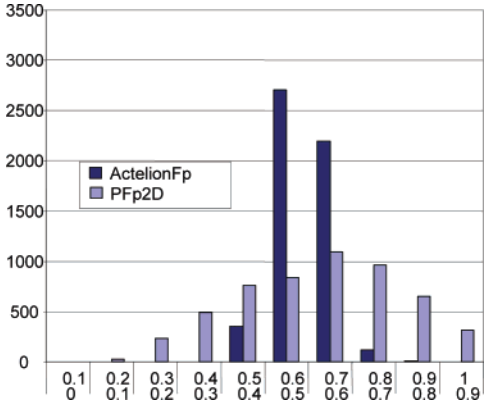

Figure 2. Histogram for distance distributions to the mean vector in GPCR space. *x* axis: Bins for the inverse Tanimoto distance to the mean vector. *y* axis: Frequency of occurrence.

RESULTS

Similarity/Dissimilarity Analysis Descriptors. For a first analysis of the behavior of the descriptors in the distance space, we calculated the distances of all objects to the average descriptor. Calculating the geometrical mean from the complete GPCR−NonGPCR data set generated the average descriptor; the distribution of the distances is shown in Figure 2. The inverse Tanimoto value served as a distance metric. The PFp2D descriptor shows the broadest distribution, starting at 0.2 and ending up at 1.0 units. The ActelionFp shows a less broad distribution, starting at 0.5 and ending at 0.9. More than 90% of the distance values of the ActelionFp are located in the bins 0.6 and 0.7, whereas the PFp2D descriptor shows a much broader distribution.

For a more detailed analysis, we randomly sampled 700 000 pairs of molecules from the combined GPCR and NonGPCR library. The descriptors for these molecules were calculated, and from these descriptors, the inverse Tanimoto distance was calculated. The distances were binned into a histogram shown in Figure 3.

For comparing molecules, the first quartile of the Tanimoto is of interest. In the zoomed graph (Figure 3b), the slope for the pharmacophore descriptor is much steeper than that for the Actelion fingerprint.

The pharmacophore descriptor assesses the molecules to be more similar than the Actelion fingerprint. In the next experiment, we analyzed the behavior in the similarity space. We chose from the literature a database with 245 benzodiazepines.[32] For all benzodiazepines, the descriptors were calculated. The complete distance matrix was calculated with the inverse Tanimoto as the metric. The distances were binned like before into the histogram (Figure 4). In contrast to the histogram of the dissimilarity space, the descriptors show an almost uniform distribution. For both data sets, the distance distributions calculated with the PFp2D descriptor are shifted to lower values than those for the ActelionFp.

Table 1. Percentage Explained Variance for the Descriptors

PC	ActelionFP	PFp2D
1	6.6357	46.4956
2	4.3973	20.2929
3	3.6827	5.5266
4	3.3475	4.1812
5	2.1235	2.3227
6	1.8727	1.7545
7	1.6187	1.691
8	1.4821	1.3584
9	1.4603	1.3417
10	1.413	1.2445

The comparison of the distance histograms for the benzodiazepines and the GPCR−NonGPCR data set indicates a high chemical diversity for the latter.

In summary, we can conclude that the dissimilarity behavior of the descriptors will result in a different behavior in classification and clustering.

PCA Plots. In our first experiments, we tried to separate GPCR and non-GPCR ligands with PCA. The cumulative explained variance for the principal components is shown in Figure 5. The principal components were derived from the complete GPCR−NonGPCR data set. The cumulated explained variance for the pharmacophore histogram reaches 90% after the first 15 principle components, whereas the 90% explained variance for the ActelionFp is reached after about 260 principle components. In Table 1, it is shown that the first two principle components for the pharmacophore and the ActelionFp descriptors explain 11% and 67%, respectively, of the overall variance.

The plots for the first two principle components are shown in Figure 6. In both plots, the upper area is dominated by the non-GPCR ligands, whereas the GPCR molecules dominate the lower area. However, there is also a broad overlap of the ligands belonging to the different class families. The hopper form of the plot for the pharmacophore histogram is generated by the distribution of the descriptors in the descriptor space. If in a descriptor only a few fields are occupied and these fields are not part of a principal

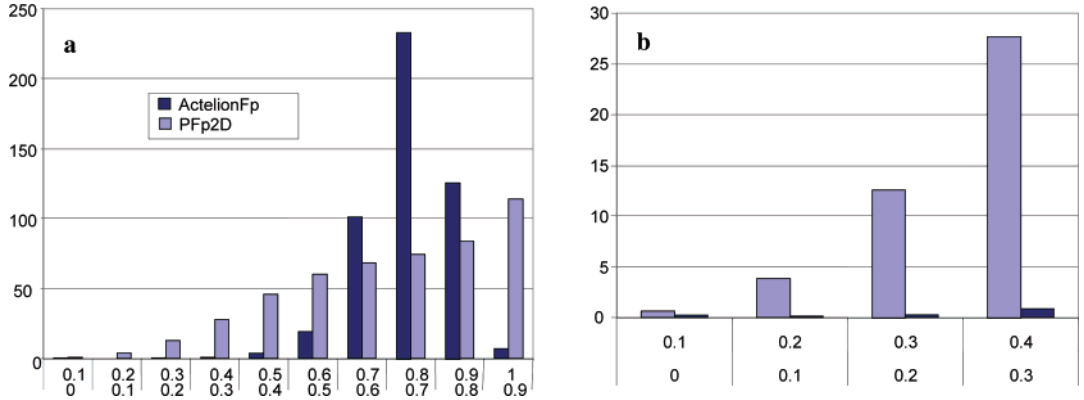

Figure 3. Histogram for distance distributions from random sampled descriptor pairs. *x* axis: Bins for the inverse Tanimoto distance. *y* axis: Frequency of occurrence × 1000. (a) Complete histogram; (b) zoomed-in image of the first four bins.

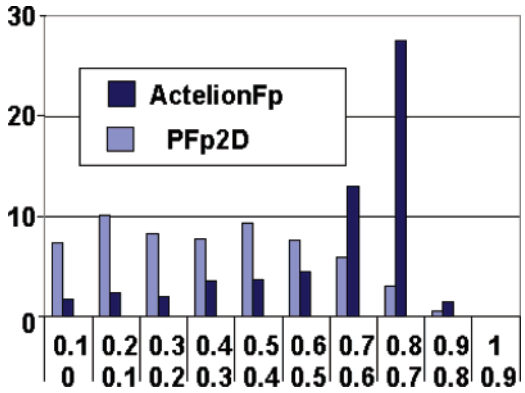

Figure 4. Histogram for the complete distance matrix of the benzodiazepine data set. *x* axis: Bins for the inverse Tanimoto distance. *y* axis: Frequency of occurrence × 1000. (a) Complete histogram; (b) zoomed-in image of the first four bins.

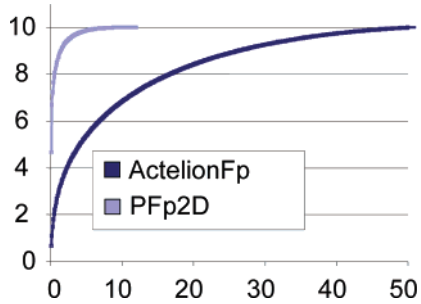

Figure 5. Cumulative explained variance for ActelionFp and PFp2D. *x* axis: Number principle components. *y* axis: Percentage explained variance.

component, its value is near zero. So, the molecules near zero contain rarely occupied descriptors. Consequently, the molecules can have different descriptors and be near together with respect to the original descriptor space. For many descriptors, the first principal component is correlated with the molecular weight. In Figure 7, we plotted the first principle component versus the molecular weight of the complete GPCR−NonGPCR database. It can be seen that there is no correlation for the ActelionFp, whereas the pharmacophore histogram shows a correlation. Solely non-GPCR molecules occupy the upper right diagonal of the plot. A closer look shows that these compounds belong mainly to the class of antibiotics. The PFp2D descriptor is more influenced by the number of atoms in the molecule than the ActelionFP; therefore, the number of atoms determines the number of entries in the histogram. Preliminary experiments showed that a normalization of the PFp2D descriptors with the molecular weights did not significantly change the clustering performance.

Other plots of the principle components showed no better separation of the GPCR−NonGPCR data.

Table 2. Results of the Classification Experiments: Original Descriptors and First Two PC and SOM Coordinates[a]

		fraction match validation	fraction match train
ActelionFP	non-GPCR	0.895	0.896
	GPCR	0.872	0.874
PFp2D	non-GPCR	0.876	0.877
	GPCR	0.849	0.854
ActelionFP PCA	non-GPCR	0.676	0.680
	GPCR	0.667	0.695
PFp2D PCA	non-GPCR	0.677	0.679
	GPCR	0.698	0.691
ActelionFP SOM	non-GPCR	0.867	0.858
	GPCR	0.867	0.861
PFp2D SOM	non-GPCR	0.849	0.830
	GPCR	0.839	0.832

[a] Figures of merit are the correct matches for the validation and the training data in terms of the fraction of correctly predicted class membership.

Classification. The results for the classification experiments are given in Table 2 and Figure 8. The table contains the average values for 11 runs. The standard deviation was in all cases below 2% and is not shown. The original descriptors of both, the ActelionFP and the PFp2D, show a high classification capability. The ActelionFP distinguishes slightly better than the PFp2D descriptor between GPCR and non-GPCR molecules.

If the first two principal components of the descriptors are taken for the classification, the percentage of correctly predicted training and validation objects is quite low. The extracted variance in the first two principal components results in a correct prediction for the validation data of less the 70% for both descriptors.

The classification with the two SOM coordinates works much better than the classification with the principal components. The percentage of correctly predicted molecules is around 85%. Like in the case of the original descriptors, the prediction for the transformed ActelionFp is slightly better than that for the transformed PFp2D descriptor.

SOM. After the classification experiments, the complete GPCR−NonGPCR data set was taken to generate two SOMs. The ActelionFp was used to generate the SOM in Figure 9a and the PFp2D descriptor to generate the SOM in Figure

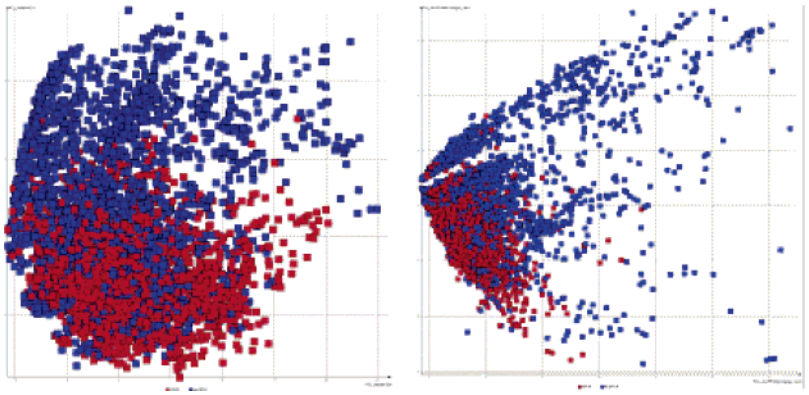

Figure 6. Plot of the first two principle components for the ActelionFp and PFp2D descriptors. GPCR ligands are red, and non-GPCR ligands are blue. *x* axis: First PC. *y* axis: Second PC.

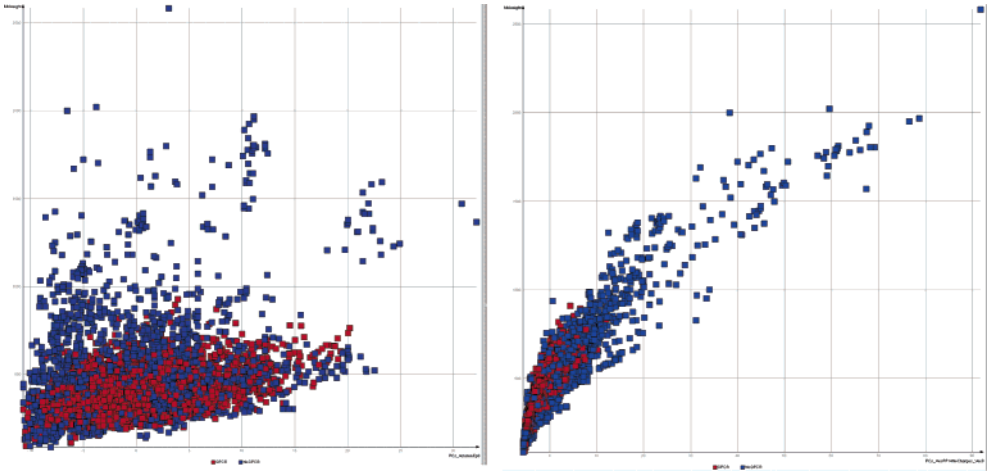

Figure 7. Plot of the first principle component versus the molecular weight for the ActelionFp and PFp2D descriptors. GPCR ligands are red, and non-GPCR ligands are blue. *x* axis: First PC. *y* axis: Second PC.

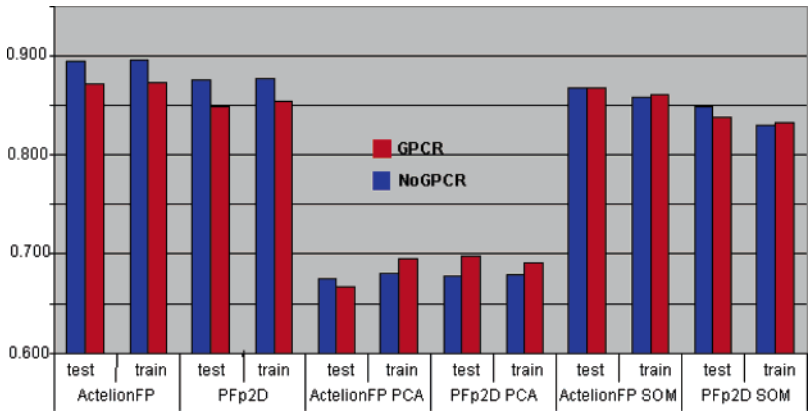

Figure 8. Classification experiments for the ActelionFp and PFp2D descriptors. *x* axis: Descriptor. *y* axis: Fraction correct matches.

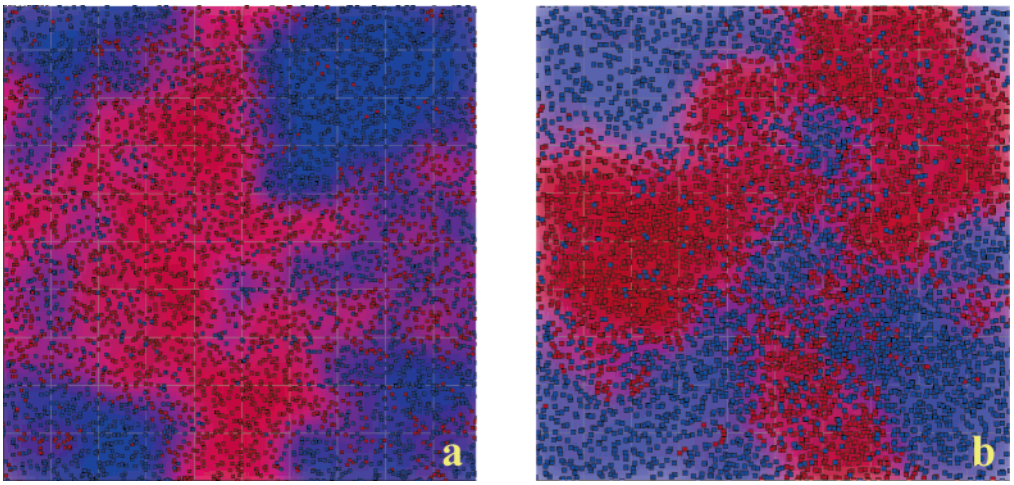

Figure 9. Kohonen maps for the (a) ActelionFp descriptor and (b) PFp2D descriptor. Objects were GPCR ligands (red) and non-GPCR ligands (blue).

9b. In both SOMs, two large clusters of GPCR and non-GPCR ligands determine the figure. Some GPCR ligands are scattered in the non-GPCR area. The two classes can be clearly differentiated from each other. It has to be considered that the SOMs in the figures are a projection of the torus into the plane. This means that the lower border of the SOM is connected to the upper border and the left border is connected to the right border. All SOMs were reproducible in the sense that we observed two large clusters in each run.

After generating the SOMs from the complete GPCR−NonGPCR data set, we mapped two supplier libraries, DB_A and DB_B, onto the SOM. The result for supplier A is shown in Figure 10 for both descriptors. In the SOM generated with the PFp2D descriptor, the DB_A is spread over the whole GPCR ligand area, whereas the SOM generated with the ActelionFp shows discrete clusters within the GPCR area. The compound library DB_B from supplier B (Figure 11) shows much less overlap with the GPCR-ligand-dominated

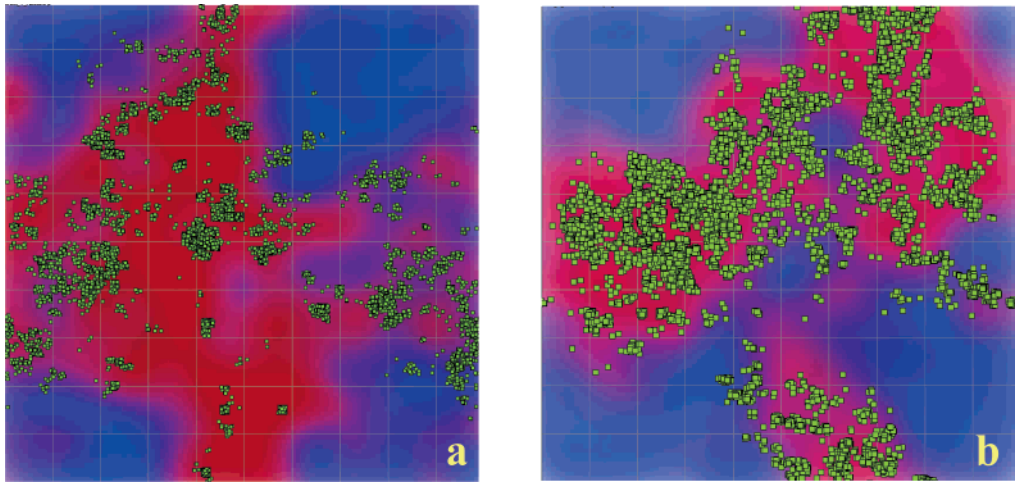

Figure 10. Compounds from supplier A mapped onto a SOM for the (a) ActelionFp descriptor and (b) PFp2D descriptor. Objects were GPCR ligands (red), non-GPCR ligands (blue), and supplier A (green).

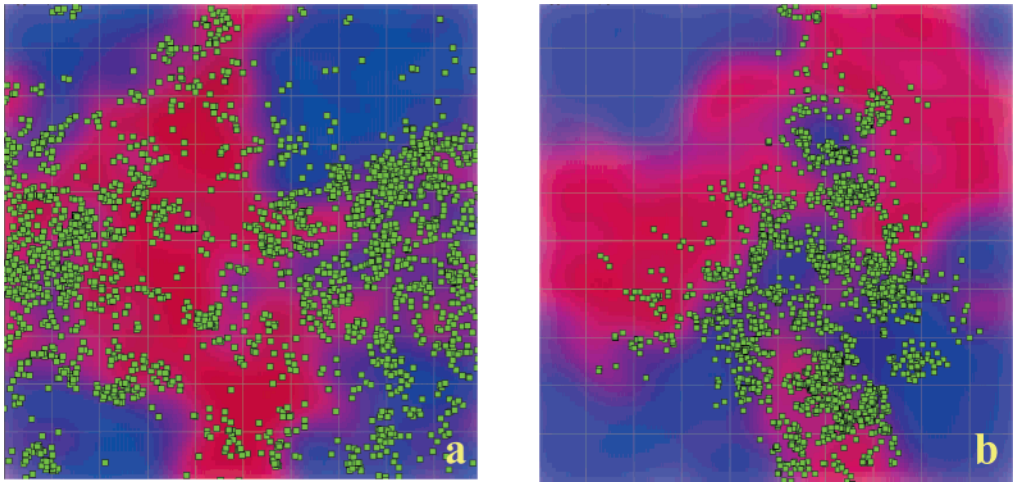

Figure 11. Compounds from supplier B mapped onto a SOM for the (a) ActelionFp descriptor and (b) PFp2D descriptor. Objects were GPCR ligands (red), non-GPCR ligands (blue), and supplier B (green).

area than the compounds of supplier A. The compounds in Figure 11a,b are mostly located at the borders of the clusters. In Figure 11b, it can be seen that the supplier compounds from library DB_B are more spread than those from library DB_A in Figure 10b.

CONCLUSIONS

From the diversity−similarity analysis, it can be concluded that both of our training and validation data sets are highly diverse. The differences in the histograms for the pharmacophore descriptor and the ActelionFp show that the chemical space is described in a different way. The broader distribution of the PFp2D descriptor compared to the distribution of the ActelionFp descriptor means that, for the latter, the diversity of the compounds is lower. The offset of the ActelionFp descriptor to higher dissimilarity values indicates that it assesses the compounds to be more dissimilar than the PFp2D descriptor does. The good clustering results for both descriptors show that both points of view are correct.

From the good correlation between the original descriptors and the target type, we conclude that our descriptors are appropriate to model the desired relationships. The slightly better result for the ActelionFp compared to the PFp2D descriptor indicates that the ligand-based GPCR space is determined by the molecular structure and not by pharmacophore patterns.

Furthermore, the classification experiments show an enormous difference between the PCA and the SOM. Concluding from our experiments, we state that, for the chosen data sets, the projection into two-dimensional space by SOMs is superior to PCA. The amount of information in the first two principal components is not sufficient to generate a powerful model. The poor modeling capability of the PCA indicates a highly nonlinear structure of the descriptor space. This is in line with the general assumption of medicinal chemists that only local models are able to explain complex chemical space. Mapping the supplier data sets onto the SOMs shows a clear difference between suppliers A and B. Future biological screenings will show whether the predicted preference for GPCR targets for supplier A is true. The results will be reported elsewhere. In conclusion, we have shown that chemical space can be clustered in an unsupervised fashion with respect to the target proteins. For this purpose, the self-organizing map is superior to the principal component analysis. Another interesting finding is that a GPCR-ligand space exits. This is attested by the classification results and the large GPCR-ligand cluster in the SOMs. Most of the GPCR ligands are continuously linked in the GPCR space. As we have shown for the supplier libraries, this

finding can be used for targeting compound libraries toward receptor classes of interest.

ACKNOWLEDGMENT

The authors wish to acknowledge Thomas Weller and Michael Scherz for proofing the manuscript and Thomas Sander for support and many valuable discussions.

REFERENCES AND NOTES

(1) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855−861.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of Non-Hierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(3) 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.
(4) 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*, 1219−29.
(5) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(6) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.
(7) Giddings, J. Modeling the Behavior of Rats in an Elevated Plus-Maze. Bachelor of Science with Honours in Mathematics and Statistics, Acadia University, Wolfville, NS, Canada, 2002.
(8) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear Mapping Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.
(9) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.
(10) Cruciani, G.; Pastor, M.; Mannhold, R. Suitability of Molecular Descriptors for Database Mining. A Comparative Analysis. *J. Med. Chem* **2002**, *45*, 2685−2694.
(11) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(12) Grigorov, M. G.; Schlichtherle-Cerny, H.; Affolter, M.; Kochhar, S. Design of Virtual Libraries of Umami-Tasting Molecules. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1248−1258.
(13) Haggarty, S. J.; Clemons, P. A.; Wong, J. C.; Schreiber, S. L. Mapping Chemical Space Using Molecular Descriptors and Chemical Genetics: Deacetylase Inhibitors. *Comb. Chem. High Throughput Screening* **2004**, *7*, 669−676.
(14) Nordling, E.; Homan, E. Generalization of a Targeted Library Design Protocol: Application to 5-HT7 Receptor Ligands. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2207−15.
(15) Kohonen, T. *Self-Organizing Maps*; Springer: New York, 2001; p 30.
(16) 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, NL, 1998; Vol. 2, pp 273−299.
(17) Kirew, D. B.; Chretien, J. R.; Bernard, P.; Ros, F. Application of Kohonen Neural Networks in Classification of Biologically Active Compounds. *SAR QSAR Environ. Res*. **1998**, *8*, 93−107.
(18) Allen, B.; Grant, G.; Richards, W. Similarity Calculations Using Two-Dimensional Molecular Representations. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 330−7.
(19) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Properties, and Drug-Likeness. *J. Med. Chem* **2002**, *45*, 3345−3355.
(20) Balakin, K. V.; Lang, S. A.; Skorenko, A. V.; Tkachenko, S. E.; Ivashchenko, A. A.; Savchuk, N. P. Structure-Based Versus Property-Based Approaches in the Design of G-Protein-Coupled Receptor-Targeted Libraries. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1553−1562.
(21) Panek, J.; Jezierska, A.; Vracko, M. Kohonen Network Study of Aromatic Compounds based on Electronic and Nonelectronic Structure Descriptors. *J. Chem. Inf. Model.* **2005**, *45*, 264−72.
(22) Bosshard, H. Molecular Recognition by Induced Fit: How Fit Is the Concept? *News Physiol. Sci.* **2001**, *16*, 171−3.
(23) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(24) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(25) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(26) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(27) von Korff, M.; Steger, M. GPCR-Tailored Pharmacophore Pattern Recognition of Small Molecular Ligands. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1137−1147.
(28) Derwent, T. World Drug Index 2002/03. www.derwent.com (accessed MMM YYYY).
(29) Lipinski, C. A.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(30) Oprea, T. Lead Structure Searching: Are We Looking at the Appropriate Property? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.
(31) Charifson, P.; Walters, W. Filtering Databases and Chemical Libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 311−23.
(32) Harrison, P. W.; Barlin, G. B.; Davies, L. P.; Ireland, S. J.; Mátyus, P.; Wong, M. G. Syntheses, Pharmacological Evaluation and Molecular Modelling of Substituted 6-Alkoxyimidazo[1,2-b] Pyridazines as New Ligands for the Benzodiazepine Receptor. *Eur. J. Med. Chem*. **1996**, *31*, 651−662.
(33) Sander, T. *ActelionFp*, unpublished work, 2002. The Actelion Fingerprint is a nonhashed binary fingerprint encoding the existence or absence of 512 predefined substructure fragments that partially contain wildcard atoms and atom query features. These fragments were computationally created by fragmenting large collections of diverse organic molecule structures. Initially, hundreds of thousands of fragments were obtained. From these, we created additional fragment sets by introducing wildcards or query features to retain original substitution patterns. Of all of the fragments, we selected computationally 512 fragments, balancing two required criteria. First, they needed to be orthogonal concerning their occurrence in diverse organic compounds, and second, their frequencies in those compounds had to be reasonable high. Both criteria ensure that these fragments are optimally suited for their original purpose, i.e., to achieve an optimal discrimination in the prescreening of a substructure search. Thrown into the bargain, these fragments also proved to be a very valuable foundation for Tanimoto-based similarity calculations. Because of the frequent usage of wildcards, calculated similarity values are more tolerant to single atom diffences, and two molecules with similar skeletal backbones are still considered similar even if some of the atoms differ.
(34) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(35) Baumann, K. An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 26−35.
(36) Jollife, I. T. *Principal Component Analysis*; Springer: New York, 2002.
(37) Tanimoto, T. Metrics and Nearest-Neighbor Classification. In *Pattern Classification*; Duda, R. O., Hart, P. E., Stork, D. G., Eds.; Wiley: New York, 2000; pp 187−192.
(38) Chen, X.; Charles, R. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−14.
(39) Whittle, M.; Gillet, V. J.; Willett, P. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.
(40) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.
(41) Baumann, K.; Albert, H.; von Korff, M. A Systematic Evaluation of the Benefits and Hazards of Variable Selection in Latent Variable Regression. Part I. Search Algorithm, Theory and Simulations. *J. Chemom.* **2002**, *16*, 339−350.
(42) Duda, R. O. kNearest Neighbor Estimation. In *Pattern Classification*; Duda, R. O., Hart, P. E., Stork, D. G., Eds.; Wiley: New York, 2000; p 187.
(43) *GenerateMD*; ChemAxon: Budapest, Hungary, 2001. http://www.chemaxon.com/
(44) Sander, T. DataWarrior, a Tool for Visualizing Data. Personal communication to Fischli, W.; Allschwil, Switzerland, 2000.

CI060037O