GPCR-Tailored Pharmacophore Pattern Recognition of Small Molecular Ligands

Modest von Korff† and Matthias Steger*,†

Axovan Ltd., Gewerbestrasse 16, 4123 Allschwil, Switzerland

Received November 25, 2003

The goal of our work was to differentiate between patterns, which are responsible for the activity of small molecular ligands binding to G-protein coupled receptors (GPCRs) and molecules, which are pharmacologically active on other target classes. Second the aim was to go one step further and analyze the chemical space occupied by GPCR active ligands itself, to distinguish between the actives of different subclasses or even cluster ligands for single receptors. To achieve these objectives, we have built a database of small, organic molecules, which bind to GPCRs. Once this crucial foundation for pattern recognition has been laid, we needed to find a descriptor, which is able to detect the compulsory features responsible for activity within a molecule. In this matter we found that the well accepted pharmacophore descriptor served us well. Finally we needed to find a method to display the clustering or separation of the specific ligands. We found that self-organizing maps (SOMs) perform excellently in this task. We herein present the analysis of the chemical space of active compounds, depending on their biological target, the GPCRs. We will also discuss the techniques used to create the chemical spaces. The findings can be applied and have an impact at various stages of the drug discovery process.

INTRODUCTION

A large percentage (>30%) of the ∼500 currently marketed drugs are modulators of G-protein coupled receptors (GPCRs) function with most major therapeutic areas being served.[1] Also, the majority of the top selling pharmaceutical products target GPCRs, making the GPCR superfamily the most successful of any target class in terms of therapeutic benefit and commercial sales. In addition, the deciphering of the human genome sequence[2] has revealed that approximately 750 genes encode GPCRs, whereby around 400 of these could be considered to be potential drug targets, of which ∼30 are the targets of currently marketed drugs. The natural ligand has been identified for a further 210 receptors, which leaves around 160 so-called 'orphan' receptors with no known ligand or function. With the rapid deorphanizing of these receptors, clearly, no single class of proteins ranks higher than GPCRs in terms of drug discovery potential.[3]

GPCRs all consist of a central core of characteristically seven transmembrane helices and are therefore also called seven trans-membrane receptors. The GPCRs are a large family of receptors and are activated by a broad variety of natural ligands such as light, Ca^{2+}, odorants, small molecules, including amino acid residues, nucleotides, and peptides as well as proteins.[4] However, known GPCRs can be classified into subfamilies, according to their pharmacological nature and sequence similarities,[3,5] resulting in different ligand binding sites, which can be within the transmembrane helices, at the extra cellular loop or even at the extra cellular N-terminal domain. Since the GPCRs are located in the cell membrane, there is very limited access to three-dimensional structures of this target family. Therefore we have decided to use the vast information of small-molecular ligands of GPCRs, together with phylogenetic relations for pattern recognition and activity predictions.

Clustering and classification methods of biologically active molecules according to their receptor affinity[6] have proven to be an inspirational source for medicinal chemists. Therefore these techniques are widely applied in the drug discovery process,[7] e.g. for the selection of compounds for high-throughput screening (HTS),[8] the design of target family focused combinatorial libraries,[9] for 'scaffold hopping' to yield novel active chemotypes[10] or to improve adverse physicochemical or toxicity issues in the lead optimization phase.[11] Extensive efforts have been undertaken to develop appropriate descriptors and appropriate clustering algorithms.[12] The classification task is based on pattern recognition, which is a natural phenomenon that occurs on all different kind of levels, such as camouflage, communication, but also for attraction.[13] The challenge is to find an algorithm that is able to detect the pattern(s), which is (are) accountable for a certain response. To enhance the speed, reduce costs, and add creativity for the discovery of novel drugs it is therefore crucial to find the molecular descriptors within a set of active molecules, which are responsible for the affinity and selectivity toward a specific biological target or a target family.

In this paper we present the analysis of the chemical space of biologically active, small organic molecules and in particular those, which are active on GPCRs. Some of these ligands have been analyzed by a variety of different computational techniques.[14] The aim of the present study was the analysis and the differentiation of the chemical space occupied by molecules, with an affinity to GPCRs and those, which are active on other drug targets such as enzymes, channels, or nucleotide receptors. The possibility of identify-

* Corresponding author phone: + +41 61 487 7654; matthias.steger@actelion.com.

† Current address: Actelion Ltd., Gewerbestrasse 16, CH-Allschwil, Switzerland.

10.1021/ci0303013 CCC: $27.50
Published on Web 04/01/2004

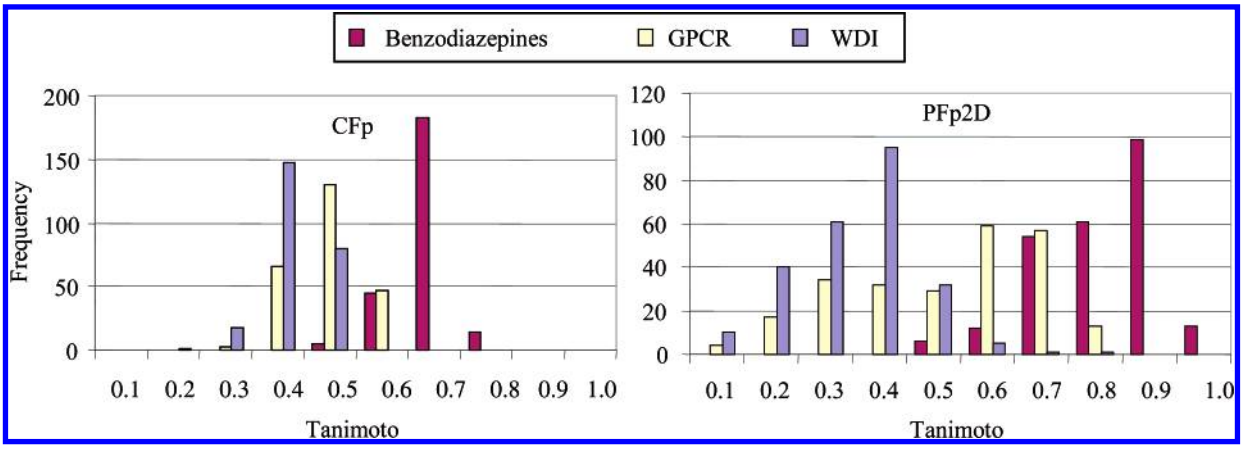

Figure 1. Histograms for the molecular diversity analysis of molecules from the WDI, the GPCR-ligands database, and the benzodiazepine data set. Each data set contains 245 molecules.

ing and defining regions in the complex chemical space with a preference for GPCR-ligands would provide us a valuable guide in the selection of molecules for our compound collection for HTS. These 'GPCR-biased' compounds could therefore increase the 'hit-rate', to give us multiple starting-points in our GPCR drug discovery programs. Once the separation in the chemical space between GPCR-ligands and 'NonGPCR'-ligands was achieved, our aim was to analyze further the chemical space occupied by the GPCR-ligands. In this study the goal was to find descriptors and chemical space analysis tools, which were able to differentiate between small organic ligands of the above-mentioned subfamilies of GPCRs and even between synthetic ligands of specific receptors. The successful partitioning of ligands within the chemical space should give us a valuable tool to improve the drug discovery process, e.g. by finding active molecules on receptors with no-known ligands by the analysis of ligands of closely related receptors, designing tailored libraries or scaffold hopping in a lead optimization program for a specific receptor to improve patent situation or physicochemical properties. The latter perceptions are based on the structure−activity relationship homology concept,[15] which describes the principle that similar ligands bind to similar targets. Furthermore we aimed for the validation of the various techniques and tools that we have employed to create the chemical spaces presented herein.

METHODS

The Data. Chemical databases have become a powerful research tool for discovering new lead compounds.[16] Fundamental to all pattern recognition tools is a solid and dynamic database. Database systems, such as MDDR[17] or WDI,[18] provide structural information about biologically active molecules but have limited information about their affinity, the target name, and do not provide any further information among the targets. Therefore the application of these systems for the elucidation of structure−activity relationships is limited.[19] We have consequently decided to build our own GPCR-ligands database of known, small, organic molecules, which are active on GPCRs. The molecules in this database are in various stages of the drug discovery process, such as drugs that are on the market or are at the stage of clinical candidate selection and effect their benefit via the modulation of GPCRs. The minimum requirement for a compound to be added is a chemical lead optimization program with a good knowledge about the structure−activity-relationship (SAR) around a certain chemical scaffold. The molecules have been collected from various sources, such as medicinal chemistry journals, patents, and very recent information from the Internet.[20] Another source of GPCR active molecules are our own drug discovery programs and screening data on more then 25 different GPCRs. However, for molecules coming from this internal source, the same criteria are applied as for published compounds. There are different criteria for a molecule to be entered into Axovan's GPCR-ligands database: it needs to be selectively active on a GPCR and the affinity should be at least in the low nanomolar (nM) range. If we have several compounds, which fulfill the affinity requirements, but have the same scaffold, we restrict the number of entries to a maximum of five molecules. If new molecules arise from more recent publications containing scaffolds, which are already present in our database, we replace the corresponding molecules by the newer ones if they show a better property profile. Therefore we call it a dynamic database. This careful chemical selection attempts to avoid a bias in our database and therefore our prediction tools toward certain redundant chemotypes. At the same time we also balance our database regarding the biological space so as not to bias the database and prediction tools toward a certain receptor or a receptor family (see Figures 1 and 2). The information, which is stored in each record in the GPCR-ligand database is as follows: the chemical structure of the active molecule and important structural information, such as MW, physicochemical properties etc. but also on which receptor the molecule is active and to which receptor-class or -family it belongs to. Additionally we add information about the mode of action, whether this is agonistic or antagonistic. With all that information we aim to link the biological space with the chemical space, which is crucial for the clustering or separation of small, active molecules. The database contains at present 2653 molecules, which are active on 128 different GPCRs.

For the analysis of the chemical space, we also needed a database that contains biologically active molecules, which bind to other targets than GPCRs. Therefore, the *NonGPCR-ligands database* was assembled with compounds from the WDI,[18] by a random selection of molecules, which display

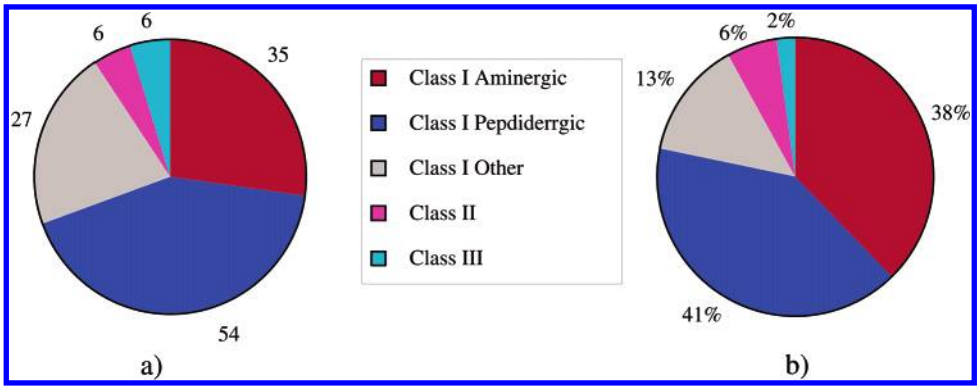

Figure 2. Histograms for the biological diversity analysis of the GPCR-ligands database. (a) Distribution of the targeted 128 different receptors, based on GPCR family classification. (b) Distribution of the 2653 GPCR-ligands regarding their target-receptor-class.

their biological activity at a variety of other targets, such as enzymes (kinases, phosphatases, proteases, etc.), transporters, channels, other cell-membrane receptors (nucleotide-, integrin- or hormone receptors etc.), or mechanisms, for instance transcription, translation, protein-modification etc. We have chosen a data set of 2726 molecules.

Molecular Descriptors. To be able to compare chemical structures with each other, molecular descriptors have to be calculated.[21] Descriptors can be subdivided into three classes: one-, two-, and three-dimensional descriptors. One-dimensional or bulk descriptors represent the molecules as a whole, e.g. polar surface area, cLogP, the number of rotatable bonds, number of rigid bonds, number of aromatic rings, shape, and branching descriptors. These descriptors are appropriate to assess the drug likeness of a molecular database.[22] Even though the above-mentioned descriptors are intrinsically related to the molecular structure, but do not represent single specific (sub)structural components, they are not ideal to capture the essence of molecular functions required for biological activity. In our work, we are following the theory that the pharmacophore points, which are represented by specific functional group types, are mainly responsible for the selectivity and the positioning of the ligand molecule at the binding site of the receptor. The concept of pharmacophore[23] is a key consideration in medicinal chemistry, which is underlined by various successful applications[10,24] and is part of the common background of each medicinal chemist. On one hand the receptor−ligand interactions can be explained by pharmacophore point interactions between defined polar groups/spheres in the molecules, and on the other hand the binding enthalpy can be explained by interactions of aromatic/hydrophobic parts of the ligand with the receptor molecule. Of course three-dimensional representations of molecules can model these interactions best, but because there is actually just one X-ray structure of a GPCR known,[25] only for a minority of GPCR-ligands reliable conformational information is available.[26]

PFp2D. Bearing this in mind, we decided to reduce the complexity of our task, which was to map all possible pharmacophore points present in each structure in the databases and therefore concentrated our studies on a two-dimensional descriptor. To represent the pharmacophores in a 2D descriptor we have developed a pharmacophore histogram descriptor vector, PFp2D. It is closely related to the atom pair descriptor[27] and to the binding property pairs descriptor.[28] The PFp2D is based on the general understanding of pharmacophore types within a molecule, which we have defined as follows: hydrogen bond donor (d); hydrogen bond acceptor (a); positive charge (+); negative charge (−); hydrophobic (h), and aromatic (r). In the descriptor the pharmacophore points are set 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 is set by the user. 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. The pharmacophore points have to be detected in the molecules and transferred into the descriptor. An example is given in Figure 3. For the classification of the atoms into pharmacophore types, we did not attempt a pK_a prediction but overcame the problem that a certain atom can represent different pharmacophore types, depending on its environment, by defining substructures.[23,24(f),29] For the detection of the pharmacophore points we then use a combination of the defined substructures and logic expressions. Nitrogen atoms are for example hydrogen bond acceptors, unless they are alkyl-amines, as they will be protonated and act as hydrogen bond donors, but become again acceptors, if the nitrogen is attached to an aromatic ring or is located in an aromatic ring. The definition for the hydrophobic pharmacophore points comprises in addition to the carbon atoms also oxygen atoms, which are attached to an aromatic ring or one of the oxygen atoms in ester- or sulfone-groups.

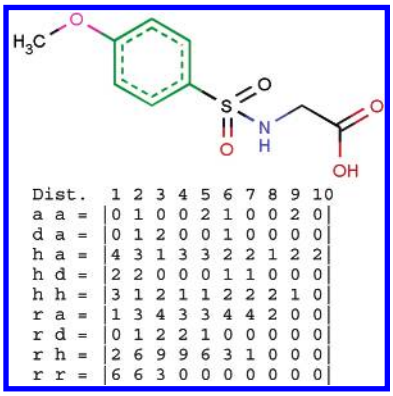

Figure 3. Example of a PFp2D descriptor histogram. Dist.: Topological distance, a: hydrogen bond acceptor, d: hydrogen bond donor, h: hydrophobic, r: aromatic.

CFp. For comparison reasons we used a chemical fingerprint descriptor from ChemAxon.[30] The CFp encodes the topological information between the atoms of a molecular graph as a binary vector. The descriptor length and therefore the resolution are user-defined. In the first step of the descriptor generation the procedure equals the generation of the atom pair descriptor and the SESP descriptor.[27] The atom types in the CFp descriptor are defined by the elements. Between all atoms the possible walks on the molecular graph are calculated and encoded as a binary vector. Hence each possible walk is represented by a bit pattern. This bit pattern is added to the descriptor with a logical OR operation. With this technique overlaps between different bit patterns are possible. This results in a hashed chemical fingerprint. The principle difference of the CFp descriptor to the PFp2D descriptor is the higher generalization power of the PFp2D descriptor.

Principal Component Analysis. The Principal Component Analysis (PCA) is a linear projection technique, whereby data vectors are projected into a lower dimensional space.[31] The PCA algorithm extracts the variance from the original vector. 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 significant principal components are used to calculate the scores, the scores are identical with the original descriptor variables. In most applications the number of significant principal components is less than the number of original variables. Without 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. Hence principal component selection can be used to exclude unwanted variance from the new descriptors. In our algorithm we used the PCA to reduce the number of variables to increase the performance of the self-organizing map algorithm. In the self-organizing map algorithm the distance calculations between the weights and the descriptor vectors are the most time-consuming task. The performance depends linearly on the length of the descriptor vectors.

Self-Organizing Maps. In the book of Duda and Hart[32] it is mentioned that, for the descriptive approach, the classification of objects is not sufficient to solve a complex pattern recognition task. Deciding for the successful pattern recognition is the context of the objects. To achieve the original context of transformed descriptor data, pure clustering is not sufficient. Another important point is the visualization, based on the knowledge that one of the most powerful tools for pattern recognition is the human eye. To use this tool an appropriate visualization algorithm is necessary. Techniques, which allow a visualization of classified data, are hierarchical clustering[33] and nonlinear mapping.[34] In nonlinear mapping the high dimensional descriptor vectors are transformed into 2D or 3D vectors. These vectors can be used as coordinates to visualize the corresponding objects. We decided to use self-organizing maps as a well-known nonlinear mapping technique, which has been developed by Tihaven Kohonen.[35] For this technique it is known that it retains the original distances by mapping high dimensional vectors into a 2D space. 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 wastes no degrees of freedom by correcting the result with information from the class memberships.[36] We used for all herein generated SOMs 100×100 weights arranged on a toroidal surface.

Objective Function. To assess the quality of the clustering we implemented a *k* nearest neighbor objective function.[32,37] For each object o in a class, the class membership is predicted. The classification quality is expressed by the ratio between the number of correct predicted class memberships and the number of objects in the class under consideration. This is done for all classes. To predict the class membership the next *k* objects to the object under consideration o are determined. The majority of the *k* next neighbors predict the class membership of o. To assess the performance of the SOM, we calculate the distances between the objects on the torus of the SOM, to find the next neighbors. In comparison we calculate the distances between the descriptors respective of their scores, which we have used to generate the SOM. The comparison between the two objective functions allows the assessment of the quality of the nonlinear mapping. If the difference is small the mapping was successful with respect to the original distances of the descriptor vectors. Due to statistical reasons we have chosen to set $k = 1$.

RESULTS AND DISCUSSION

The GPCR ligands database should be able to capture the chemical diversity of known GPCR-ligands as well as the biological diversity of this receptor family.

Molecular Diversity. The assessment of the molecular diversity of our GPCR-ligands database was done by comparing the latter data set with the NonGPCR-ligands database from the WDI and with a publicly available data set of 245 benzodiazepines.[38] We chose the benzodiazepine data set to have a uniform data set, which is nevertheless of pharmacological interest and located in the space of bioactive molecules. This enabled us to compare a relatively uniform data set with the highly diverse WDI-data set and a data set of ligands for a certain class of receptors. We have taken a subset of 245 molecules of each of the three databases, generated for each of them the PFp2D descriptor and transformed it into a vector. Additionally we calculated for each molecule a chemical fingerprint descriptor (CFp).[39] The length of the CFp was set to 512 bits and with a walk length of four.

For the CFp as well as for the PFp2D, a mean descriptor vector was calculated and the Tanimoto distance matrix was calculated between the mean descriptor and all other descriptors. The distances for each descriptor were collected in a histogram. The results are shown in Figure 1. Tanimoto 1.0 means total similarity, i.e., they are the same molecules, whereas Tanimoto 0.0 is equal to total dissimilarity. The results of the chemical diversity analysis do reflect what we aimed to achieve regarding the chemical diversity of the entries in our GPCR-ligands database. It shows that the collected GPCR-ligands are much more diverse than the set of benzodiazepines but less diverse then the molecules from the WDI. However, the diversity of the GPCR-ligands is much closer to that from the WDI, hence proving that we

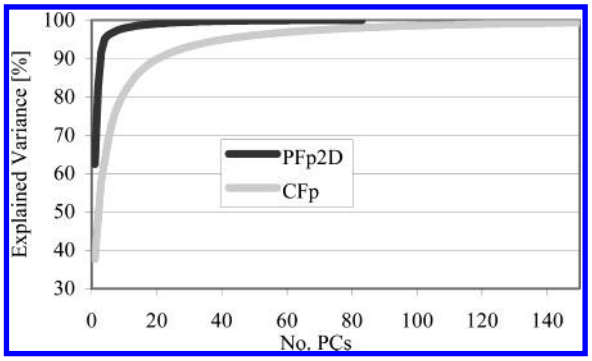

Figure 4. Explained variance for the PFp2D descriptor and the CFp descriptor.

do cover a broad diversity regarding the chemical- and pharmacophoric space of molecules, which display their activity at targets within the GPCR superfamily. Additionally, it can be seen that the distance matrix of the pharmacophore descriptor spans a wider distribution. The values in the histograms of the CFp descriptor range from a Tanimoto coefficient 0.2−0.8, whereas the values in the histograms of the PFp2D descriptor range from absolute dissimilarity to identical. This reflects the high sensitivity of the PFp2D compared to the CFp.

Biological Diversity. The content of the GPCR-ligands database was analyzed not only according to the chemical diversity of the entries but also toward the biological diversity of receptors on which the ligands display their activity. The superfamily of GPCRs has been classified into the three main receptor subfamilies I, II, and III.[3,4] Family I has been further divided into the three following subfamilies: (1) aminergic-receptors, such as muscarinic-, adrenergic-, dopamine-, histamine-, or serotonin-receptors. (2) peptidergic receptors, like angiotensin-, bombesin-, bradykinin-, chemokine-, endothelin-, melanocortin, or neuropeptide Y-receptors. (3) Others are represented by prostanoid-, nucleotide-like-, cannabinoid-, melatonin-, or leukotriene-receptors. Receptors belonging to the family II are the corticotropin-releasing factor-, glucagon-, ghrelin-, or the motilin-receptor. Last, family III contains the metabotropic glutamate receptor. Figure 2a shows that the distribution of the 128 different, targeted receptors are reflecting the natural occurrence in the genome sequence of receptors belonging to family I (89%), II (7%), and III (4%).[3] A very similar allocation pattern can be seen in Figure 2b in which the number of ligands per subfamily is shown. One of our main goals was to not bias our database toward aminergic-receptor ligands, due to the many, historically known ligands. This task has been achieved, as the database contains at least as many ligands, which are active on peptidergic receptors.

PCA of the Descriptors. To study the variance in the molecular descriptors we calculated the chemical fingerprints (CFps) from ChemAxon[39] for the GPCR data set. The resulting descriptor vectors are written into row vectors. These descriptor vectors are piled up to a matrix on which we performed a PCA. Additionally we calculated the PFp2D descriptors and performed a PCA. The cumulated, explained variance for both descriptor types is shown in Figure 4. The PCA of the CFp descriptor yields a much higher number of significant principal components. This phenomenon was expected, because the PFp2D descriptor summarizes several substructures into one pharmacophore point. The lower number of principal components means that less latent variables are needed to explain the total variance of the descriptor. The result is a sparse model, compared to the models yielded from the CFp descriptor. The few numbers of latent variables was one of the factors choosing the PFp2D descriptor for exploring the GPCR space and not choosing the CFp descriptor.

Chemical Space Analysis of GPCR− vs NonGPCR-Ligands. Our first undertaking in the analysis of the chemical space was to investigate the ability of the PFp2D descriptor to discriminate between GPCR-ligands and NonGPCR-ligands. On this behalf, the PFp2D histograms were calculated, followed by a PCA of the descriptor for each molecule in both databases. We realized in some preliminary results that the first principal component of the PFp2D descriptor contains mainly the molecular weight information. This is probably due to the fact that the PFp2D histogram descriptors relate to the size and the molecular weight of drug-like molecules, because the number of functional groups and the number of pharmacophore points is strongly correlated with the molecular weight of drug-like molecules. However, because the variance of the molecular weight of the GPCR-ligands as well as that of the NonGPCR-ligands is very large it is not relevant for their respective classification. Hence the first principal component, which contains the highest fraction of variance, was omitted for the calculation of the latent variables for the PFp2D descriptors.

No Charges. As outlined above, we did not use the charge (pK_a) information to generate the PFp2D descriptors. With our substructure orientated approach for the pharmacophore point definition, the implementation of charges showed no improvement for the clustering and classification power of our PFp2D descriptor. But the introduction of charges increases the size of the descriptor vector from 120 to 252. So we decided to omit the charge information from our descriptor.

PCA. We started our clustering experiments by drawing the first two and three principal components against each other. No clusters were detectable, neither for the GPCR-ligands nor for the NonGPCR-ligands. Other combinations of principal components than the first three were also not successful. The conclusion is that the PCA is not able to cluster the ligands according to the corresponding receptor information.

SOM. Because the PCA method was not successful for the discrimination of GPCR-ligands from biologically active molecules binding to other targets within the chemical space, we decided to implement a nonlinear mapping method. To reduce the number of descriptor variables to less latent variables, we first performed a principal component analysis (PCA) on the calculated PFp2D descriptors of the GPCR-ligands. The PCA prevents the SOM from the modeling collinearity. The eigenvectors from the PCA were used to calculate the latent variables for the NonGPCR-ligands database. With the restriction to the descriptor vectors from the GPCR DB for the calculation of the principal components we introduce a small bias in direction of the GPCR compounds. This restriction based on the realization that the variance in the NonGPCR DB is much higher than in the GPCR DB. So the variance in the NonGPCR DB would have over determined the variance in the GPCR DB. If not mentioned otherwise from the PFp2D descriptors the prin-

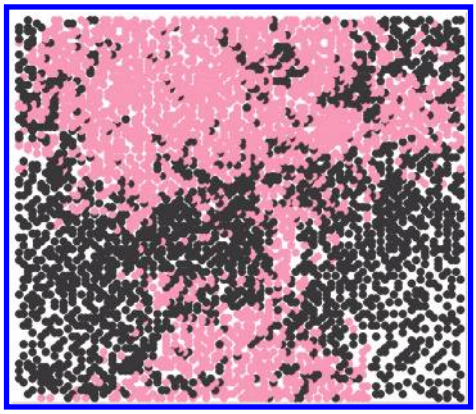

Figure 5. SOM of 2653 GPCR-ligands (pink) and 2726 NonGPCR-ligands (dark gray).

cipal components 2−30 were used for the generation of the SOM. The eigenvectors from GPCR-ligands database are used to generate the latent variables for the GPCR DB and for the NonGPCR DB. The principal components 2−30 were used to generate the latent variables as descriptors. The number of training cycles for all SOM training runs with the GPCR−as well as with the NonGPCR-ligands data set was set to 10^6. The result of the SOM, based on the PFp2D descriptor, for the analysis of the chemical space, taken by GPCR-ligands and other biologically active molecules is shown in Figure 5. A position for a weight occupied by a GPCR-ligand is colored in pink, and the ones occupied by a NonGPCR-ligand are colored in dark gray. The map shows two big clusters, whereby subspaces can be easily detected with a clear preference of GPCR-ligands and other parts, where preferably NonGPCR-ligands are located. A few NonGPCR clusters are enclosed by GPCR ligands, and a few NonGPCR ligands are scattered in the GPCR area. This may account for the fact that the analyzed molecules have not been tested on all possible biological targets. Therefore molecules selected from the WDI[18] may also be active on GPCRs and vice versa. Only a few GPCR ligands are enclosed in the NonGPCR area. The structure of the map can be further resolved between dense and sparse occupied areas. The successful separation of GPCR active molecules and pharmacologically active compounds, which bind to other targets, allows us to build a compound collection for HTS with a GPCR bias. This resulted in an improved hit-rate. Because we have selected compounds based on their pharmacophore pattern and not on their chemical similarity, providing a choice of several chemotypes to start a medicinal chemistry program. Accounting for rules for bioavailability,[22] when choosing molecules for our compound collection, the identified hits delivered typically good starting-points to improve the potency and selectivity of the lead structures.

GPCR-Ligands Space. With the successful separation of the chemical space occupied by GPCR-ligands from the space covered by NonGPCR-ligands in hand, our next task was the analysis of the GPCR-ligands space itself. We therefore attempted to solve the problem of the differentiation of the chemical space of GPCR-ligands from the specific families or subfamilies.

Aminergic- vs Peptidergic-Receptor−Ligands. Within the chemical space of GPCR-ligands, our first focus was on the family I, as defined above. Our goal was the separation of the small-molecular ligands of biogenic amine receptors from those binding to peptidergic receptors. Again we calculated the PFp2D of 1005 aminergic receptor−ligands and 1075 peptidergic receptor−ligands. The result of the SOM is shown in Figure 6. A position for a weight occupied by an aminergic-receptor−ligand is colored in red, and those occupied by peptidergic-receptor−ligands are colored in blue. As shown in Figure 6, there exists a clear difference between ligands active on the aminergic receptors (red) and ligands active on the peptidergic receptors (blue). We can easily detect subspaces, which are preferably covered by aminergic receptor ligands and other regions with a clear preference for peptidergic receptor ligands. We would like to emphasize that the clustered ligands are not the natural ligands of the receptors but small organic molecules, taken from our GPCR-ligands database. Demonstrating that molecules in the peptidergic family have not a peptidic substructure but are also small, organic molecules that bind to receptors, which belong to the family I of GPCRs and have a peptide as natural ligand.

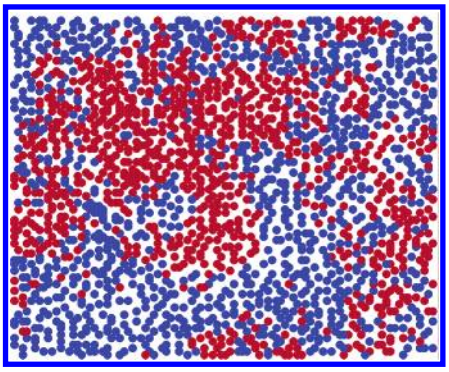

Figure 6. SOM of 1005 aminergic GPCR-ligands (red) and 1075 peptidergic GPCR ligands (blue).

GPCR Classes I, II, and III. Next, we aimed at the separation and clustering of ligands active on the three different GPCR main families. As the family I is by far the largest, we divided the ligands again in those active on aminergic-, peptidergic-, or other receptors in family I, such as prostanoid-, nucleotide-like-, or cannabinoid-receptors. The aminergic ligands (red), peptidergic ligands (blue), and the class II ligands (magenta) show clear clusters. As mentioned before, the classification relies on the structure types of the natural ligands. The class III ligands show a cluster from the upper left corner to the upper right corner, which is actually one cluster, because of the toroidal nature of the map. The cluster is not dense, and the space between the class III ligands is filled with ligands from other classes. The miscellaneous molecules are grouped in several clusters. The successful clustering and separation of small molecular ligands displaying their activity GPCRs belonging to different subfamilies can have several impacts on the discovery of novel drugs. It allows us to apply, what is known as a chemogenomics knowledge-based strategies.[19] The knowledge of structure−activity relationship on already explored targets can serve as starting points for the design of pharmacologically active structures on novel targets, which belong to the same family. With the clear clusters in the map of Figure 7, we are well positioned to explore recently deorphanized receptors, belonging to these subfamilies.

Receptor Class Discrimination. Our ultimate task of chemical space analysis, after the successful discrimination

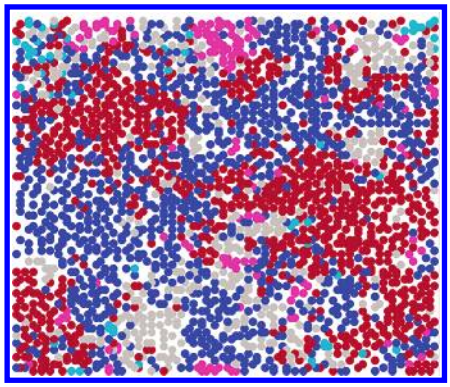

Figure 7. SOM of GPCR ligands. Coloring: class I aminergic red, peptidergic blue, and other light gray; class II magenta and class III cyan.

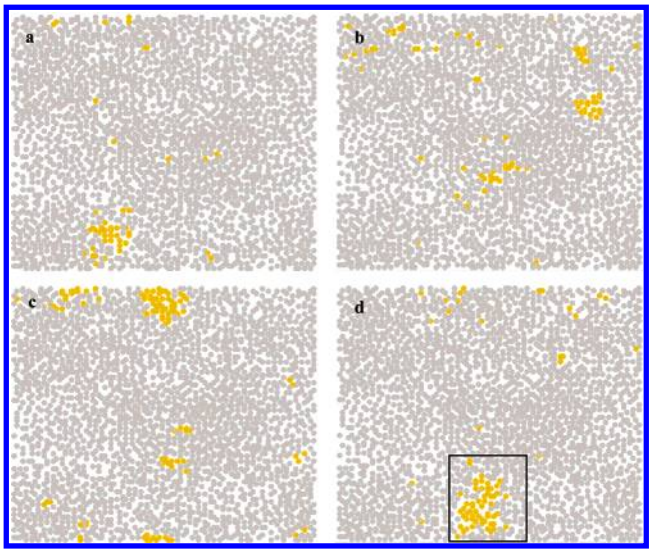

Figure 8. SOM of GPCR-ligands colored: (a) adenosine A2A, (b) cannabinoid, (c) CRF, and (d) endothelin.

of GPCR−from NonGPCR-ligands and the successful clustering of ligands from different GPCR subfamilies, was to go even deeper into the chemical space and group the ligands from specific receptors.

To demonstrate therefore the clustering power of the SOM with the PFp2D descriptor we colored the ligands, taken from our GPCR-ligands database, of four different receptors: adenosine A2a, a nucleotide-like receptor belonging to family I, cannabinoid (family I, other), endothelin (family I, peptidergic), and CRF (family II). The result is shown in Figure 8. The 47 molecules, active on the A2A receptor in Figure 8a, show a main cluster in the left bottom corner of the map, whereas the 83 actives on the cannabinoid (CB) receptor group mainly in the right top corner, but additionally another cluster can be found in the middle and some CB-ligands are spread over the SOM. The latter may be due to the fact that CB-ligands are actually coming from two receptor subtypes, CB1 and CB2. Again very nicely grouped together are the 113 CRF-ligands, in the middle at the top, continuing at the bottom of the toroidal map (Figure 8c). As with the CB-receptors, CRF has two receptor subtypes existing. However, so far only very few molecules are published, which bind to the CRF2 receptor, and therefore the map shows mainly CRF1 receptor−ligands. Figure 8d shows the positioning of 85 molecules, which display their activity at the endothelin (ET) receptor. Although, also the ET receptors are occurring in two subtypes, ETA and ETB for which we have about the same number of actives stored, only one large cluster in the middle at the bottom of the map can be seen, and very few compounds are not located within this cluster. It can therefore be concluded that the CB-receptor-subtypes require more diverse ligands than the ET-receptors, which seem to have more similar binding pockets.

The classification of single, specific receptor−ligands gives us the possibility to further support and add creativity to our drug discovery programs. We are now able to choose compounds for acquisition from commercial sources for specific receptors or related receptor subtypes, based on the chemogenomics knowledge-based strategies.[19] Furthermore it offers the opportunity to perform a 'scaffold-hopping', based on a lead series with a known structure−activity relationship. This may be desirable because of e.g. adverse patent situation, bioavailability, selectivity, or toxicity issues of the current lead-scaffold. To demonstrate the feasibility for 'scaffold-hopping' with the herein described technique of applying a combination of the pharmacophore descriptor PFp2D together with the PCA and SOM methods, we have taken a random selection of five ET receptor−ligands. The selected ligands **a**−**e**[40−44] are displayed in Figure 9, which also shows the magnified cluster of ET receptor−ligands from Figure 8d. It is obvious that the chemical structures of these ET active molecules are diverse and they all have a different scaffold. However, these molecules have the same pharmacophore, which is the reason why they are clustered in the SOM, but demonstrate diverse chemotypes, which is the goal to be achieved in 'scaffold hopping'.

Objective Functions and Comparison of Descriptors. In Table 1 it is shown that the classification power for the GPCR−NonGPCR data set is at least 85%. The classification of the original descriptors, with around 89% for the GPCR- as well as for the NonGPCR data set is slightly better than the classification with the SOM coordinates as descriptors. From this slight difference it can be concluded that the SOM conserves the original distances very well. This result is in coherence with the literature.[45] The probability to classify correctly the objects by chance is equal to the percentile of the objects belonging to one class, in the described case approximately 50%. A classification of 90% correctly predicted objects in a highly diverse data set with an unsupervised method emphasizes the power of the PFp2D descriptor. This is also supported by the classification results of the family I ligands with the KNN objective function (K=1), which are 0.837 for the peptidergic ligands and 0.835 for the aminergic ligands.

The obvious clustering shown in the Figures 5−8 and the values of the KNN SOMTopo objective function are already strong hints for the robustness of the applied clustering method. To check the clusters for stability with a numerical value, the KNN objective function was applied with different k values. The result displayed in Table 2 shows that with increasing k the classification power of the objective function decreases only slightly. We can therefore conclude that the clustering is robust.

To test for the loss of important information through the PCA we compared the classification power of the PFp2D descriptor and the latent variable descriptor generated by the scores from the principal components 2−30 of the PFp2D

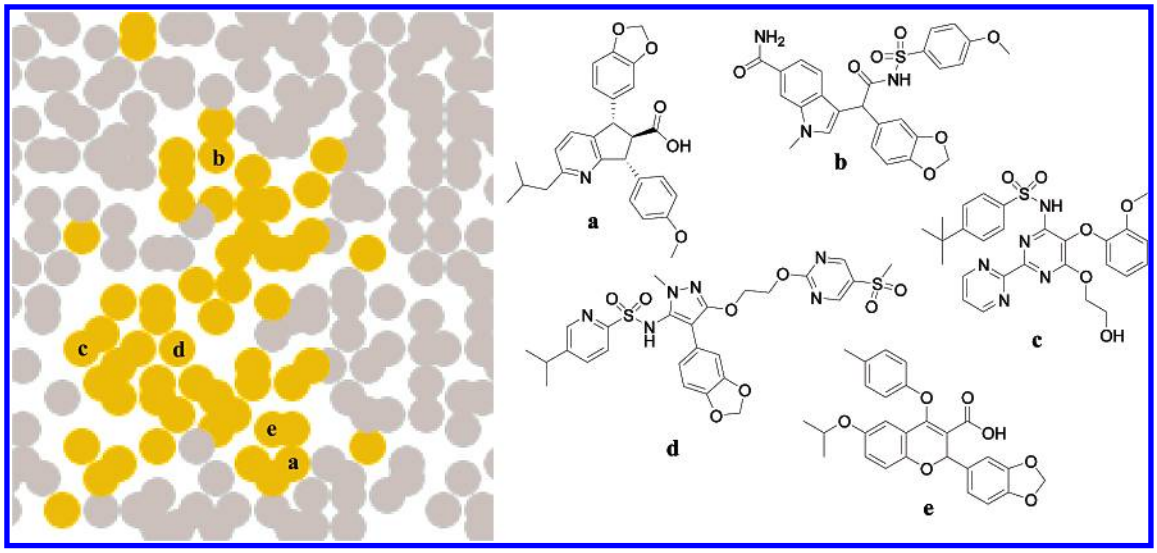

Figure 9. Part of the GPCR SOM from Figure 7d. The endothelin ligands are colored.

Table 1. Classification Result for PFp2D Descriptors of the GPCR vs the NonGPCR Data Set

objective function KNN $K = 1$	class	distribution of objects in classes[a] PFp2D descriptor	PFp2D descriptor PCs 2-30
SOMTopo	GPCR	0.851	0.849
SOMTopo	NoGPCR	0.861	0.863
descriptors	GPCR	0.897	0.886
descriptors	NoGPCR	0.885	0.890

[a] Distribution of objects in classes given by number (percentile): GPCR 2653, (0.493), NoGPCR 2727 (0.507). SOMTopo: objective KNN function with Euclidean distance on the topology of the SOM. Descriptors: objective KNN function with Euclidean distance of the descriptors (latent variables or original variables).

Table 2. Objective for GPCR−NonGPCR Classification[a]

objective function KNN $K = 1$	class	PFp2D PCs 2-30 $K = 1$	$K = 3$	$K = 5$	$K = 7$
SOMTopo	GPCR	0.845	0.863	0.865	0.865
SOMTopo	NoGPCR	0.867	0.852	0.841	0.846
descriptors	GPCR	0.885	0.885	0.884	0.881
descriptors	NoGPCR	0.885	0.865	0.862	0.860

[a] SOMTopo: objective KNN function with Euclidean distance on the topology of the SOM. Descriptors: objective KNN function with Euclidean distance of the descriptors (latent variables or original variables).

matrix as described in the PCA paragraph. The results are given in Table 2: there is no significant difference between the classification power of the two descriptors for the GPCR−NonGPCR data set. The transformation of the descriptors into the principal component space preserves the information, which is necessary for the successful clustering. We observed that the PCA can successfully be used to reduce the dimensionality of the input vectors in the SOM. This results in a linear increase of performance and additionally gives the possibility to get rid of unwanted information in the descriptors, like the molecular weight.

Another interesting question was the clustering power of the SOM if more than two groups are analyzed. Therefore the KNN objectives were calculated for the SOM of the

Table 3. Objective SOMTopo for GPCR-Family Classification[a]

no. in class	percentile	class	PFp2D	PFp2D PCs 2-50
1005	0.379	I aminergic	0.819	0.841
1072	0.404	I peptidedergic	0.746	0.762
362	0.136	I other	0.61	0.638
157	0.059	class II	0.764	0.739
57	0.021	class III	0.351	0.456

[a] Distribution of objects in classes given by number (percentile): I aminergic 1005 (0.379), I peptidergic 1072 (0.404), I other 362 (0.136), class II 157 (0.059), and class III 57 (0.021).

GPCR-ligands, grouped according to their family membership (Figure 7). The results of the KNN objectives are shown in Table 3. For the latent variable PFp2D descriptor we decided to take the 49 PCs and not the 29 PCs we used to classify the GPCR−NonGPCR data set, because of the higher complexity of the problem to cluster more than two classes in the GPCR space. It can be assessed from Figure 4 that the first 50 PCs include 100% of the variance. The first principal component was again omitted, as it was not desirable to bias our result by describing the molecular weight by the first latent variable. The results for the PFp2D descriptor and the latent variable PFp2D descriptor differ a little bit. For the aminergic, peptidergic, other class I ligands, and the class III ligands, the latent variable PFp2D descriptor shows higher classification power than the PFp2D descriptor. But compared to the chance classification, given in the table by the percentile, the differences in the descriptors are insignificant. We can state that the classification power of the KNN SOMTopo objective function compared to the classification by chance is especially high for the classes with only few class members.

To assess the classification power of the SOM for the ligands of single receptors, we selected ligands from four examples from different receptor classes. The results of the KNN objectives calculation are shown in Table 4. With a correct classification percentile of 0.81 the CRF receptor ligands yield the best results, whereas the classification of the ligands for the A2A-, cannabinoid-, and the endothelin receptor results in a percentile between 0.62 and 0.68. Compared with the classification by chance, the classification power for each receptor is by more than 1 order of magnitude

Table 4. Classification of the Ligands for Four Different Receptors

receptor	A2A	cannabinoid	CRF	endothelin
number ligands	47	83	113	85
objective[a]	0.617	0.651	0.805	0.682
classification by chance	0.0177	0.0313	0.0426	0.0320

[a] Objective function SOMTopo with $K = 1$.

better. Concluding, the PFp2D descriptor is able to cluster successfully ligands of different, single receptors.

Summarizing the results from applying the KNN objective function, it can be stated that all objective functions show a strong coherent behavior. The classification power and therefore the probability to predict the pharmacological activity of compounds with the PFp2D and with the derived latent variable PFp2D descriptor is high for all cases we considered. The variability of the classes classified by the KNN objective function in this study comprises the very inhomogeneous GPCR−NonGPCR data set, the more homogeneous data set with the aminergic- and the peptidergic GPCR-ligands and the single-receptor−ligands data set. So the PFp2D descriptor is able to cluster molecular data with a broad range of homogeneity.

CONCLUSIONS

The aim for the presented work was the understanding of the phenomena of pattern recognition, based on the fact that similar ligands bind to similar targets. Combining a small number of distinct pattern modules creates diverse patterns. The pattern modules have to be detected by the matching descriptor. A generally applicable 'default set' of molecular descriptors does not exist, and the choice of the 'right' descriptors will usually be governed by the nature of the questions to be answered.[46] The pharmacophore descriptor captures the information about the potential points of interactions between ligands and the target protein, and it is well accepted within the medicinal chemistry community. Our PFp2D pharmacophore fingerprint descriptor based on 2D representations of molecules proved to be an excellent choice for our clustering tasks. However, for the classification of objects it is not sufficient to solve a complex pattern recognition task, we also had to find a method for visualization. The unsupervised self-organizing maps (SOMs), which are based on the nonlinear mapping technique, was successfully applied to our clustering tasks.

We achieved the successful clustering and separation of GPCR-ligands and molecules, which display their pharmacological activity on other targets. Our second task was to discriminate between chemical spaces covered by GPCR-ligands, according to their family membership. Within this chemical space occupied by GPCR-ligands, we were able to separate ligands binding to aminergic receptors from those binding to peptidergic receptors. In another challenging task we attempted to identify clusters of ligands belonging to each of the three main subfamilies of GPCRs. Also the clustering of ligands displaying their activity at five different subfamilies could be solved in one self-organizing map by the application of our carefully assembled data sets in combination with the 'right' descriptors and the 'right' visualization tools. Using the same combination of cheminformatic tools, we even achieved to separate ligands from different specific receptors.

The impacts of these 'virtual' findings on the drug discovery process are manifold and can be underlined by several applications. With the knowledge of the separation of compounds, according to their biological targets, it is possible to select and/or acquire from commercial compound vendors molecules, which have a bias toward a certain target family. This may result in a higher hit rate and therefore giving a choice of starting-points for the initiation of a medicinal chemistry program. An interesting task in the postgenomic age is the finding of actives on previously unexploited targets. This can be achieved, by the chemogenomics knowledge-based strategy,[19] whereby the basis is the clustering of ligands according to their subfamily membership. In a later stage in the drug discovery process, the successful pattern recognition of pharmacologically active molecules can help finding novel chemotypes in the lead optimization process. This can be achieved by the so-called 'scaffold-hopping' and is desirable, if the current lead series do not fulfill certain expectations, such as bioavailability, toxicity, or patent-ability. One can think of many more applications of a deciphered pattern, such as the design of combinatorial libraries, which have beneficial effects on the discovery of novel drugs.

The visual impression of the evident clustering one gains when looking at the herein presented SOMs is confirmed by the calculation of the objective functions. The essential foundation for the successful clustering studies however was the selection of the pharmacophore PFp2D descriptor and the underlying GPCR-ligands database. With this data set we have achieved to capture the chemical diversity of molecules binding to GPCRs and also to reflect the biological diversity of the specific receptors and subfamilies of this diverse target class.

ACKNOWLEDGMENT

The authors wish to express their thanks to Bernard Przybylski for his invaluable assistance in the construction of the whole system and ChemAxon for the implementation of the CFp descriptor and of the PFp2D descriptor. We also would like to thank Caroline Petitjean, Eveline Engel, and Barbara Kammer for the content management of the databases and Richard Huckle for proofreading the manuscript. Hugo Kubinyi, Knut Baumann, and Thomas Sander are gratefully acknowledged for fruitful discussions.

REFERENCES AND NOTES

(1) Wise, A.; Gearing, K.; Rees, S. Target validation of G-protein coupled receptors. *Drug Discovery Today* **2002**, *7(4)*, 235−246.

(2) (a) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Balwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860−921. (b) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A. The sequence of the human genome. *Science* **2001**, *291*, 1304−1351.

(3) Chalmers, D. T.; Behan, D. P. The use of constitutively active GPCRs in drug discovery and functional genomics. *Nature Rev. Drug Discovery* **2002**, *1*, 599−608.

(4) Bockaert, J. G protein-coupled receptors. *Encyclopedia Life Sci.* **2001**, 1−9.

(5) Kawasawa, Y.; McKenzie, M. L.; Hill, D. P.; Bono, H.; Yanagisawa, M. G protein-coupled receptor genes in the Fantom2 database. *Genome Res.* **2003**, *13*, 1466−1477.

(6) (a) Downs, G. M.; Barnard, J. M. Clustering Methods and their uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.

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

(7) (a) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity?. *J. Med. Chem*. **2002**, *45*, 4350−4358. (b) Oprea, T. I.; Zamora, I.; Ungell, A.-L. Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.* **2002**, *4*, 258−266. (c) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5(2)*, 49−58. (d) Kubinyi, H. Molekulare Aehnlichkeit. Pharmazie in unserer Zeit **1998**, *27(3)*, 92−106 and 158−172. (e) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem*. **1998**, *41*, 3325−3329.

(8) Manallack, B. G.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C. Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1256−1262.

(9) (a) Schneider, G.; Nettekoven, M. Ligand-based combinatorial design of selective purinergic receptor (A_{2A}) antagonists using self-organizing maps. *J. Comb. Chem.* **2003**, *5*, 233−237. (b) Pascual, R.; Mateu, M.; Gasteiger, J.; Borrell, J. I.; Teixido, J. Design and analysis of combinatorial library of HEPT analogues: comparison of selection methodologies and inspection of the actually covered chemical space. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 199−207. (c) Brunne, R. M.; Hessler, G.; Muegge, I. In *Handbook of Combinatorial Chemistry*; Nicolaou, K. C., Hanko, R., Hartwig, W., Eds.; Wiley-VCH: Weinheim, 2002; Vol. 1, Chapter 27, pp 761−783. (d) Poulain, R.; Horvath, D.; Bonnet, B.; Eckhoff, C.; Chapelain, B.; Bodinier, M.-C.; Déprez, B. From hit to lead. Combining two complementary methods for focused library design. Application to m opiate ligands. *J. Med. Chem.* **2001**, *44*, 3378−3390. (e) Andrews, K. M.; Cramer, R. D. Towards general methods of targeted library design: Topomer shape similarity searching with diverse structures as queries. *J. Med. Chem*. **2000**, *43*, 1723−1740.

(10) (a) Flohr, S.; Kurz, M.; Kostenis, E.; Brkovich, A.; Fournier, A.; Klabunde, T. Identification of nonpeptidic urotensin II receptor angagonists by virtual screening based on a pharmacophore model derived from structure−activity relationship and nuclear magnetic resonance studies on urotensin II. *J. Med. Chem*. **2002**, *45*, 1799−1805. (b) Palomer, A.; Cabré, F.; Pascual, J.; Campos, J.; Trujillo, M. A., Entrena, M. A. G.; Garcia, L.; Mauleon, D.; Espinosa, A. Identification of novel cyclooxygenase-2 selective inhibitors using pharmacophore models. *J. Med. Chem*. **2002**, *45*, 1402−1411. (c) Fisher, L. S.; Güner, O. F.; Seeking novel leads through structure-based pharmacophore design. *J. Braz. Chem. Soc.* **2002**, *13 (6)*, 777−787. (d) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. 'Scaffold-hopping' by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed*. **1999**, *38 (19)*, 2894−2896.

(11) (a) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem*. **2003**, *46*, 1250−1256. (b) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 647−679. (c) Härter, M. W.; Keldenich, J.; Schmitt, W. In *Handbook of Combinatorial Chemistry*; Nicolaou, K. C., Hanko, R., Hartwig, W., Eds.; Wiley-VCH: Weinheim, 2002; Vol. 1, Chapter 26, pp 743−759. (d) Cronin, M. T. D. Computational methods for the prediction of drug toxicity. *Curr. Opin. Drug Discovery Dev.* **2000**, *3(3)*, 292−297.

(12) (a) Gobbi, A.; Lee, M.-L. DISE: Directed sphere exclusion. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 317−323. (b) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 182−188. (c) Godden, J. W.; Xue, L.; Bajorath, J. Classification of biologically active compounds by median partitioning. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1263−1269. (d) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 391−405.

(13) Anderson, J. C.; Baddeley, R. J.; Osorio, D.; Shashar, N.; Tyler, C. W.; Ramachandran, V. S.; Crook, A. C.; Hanlon, R. T. Modular organization of adaptive colouration in flounder and cuttlefish revealed by independent component analysis. *Network: Comput. Neural Syst.* **2003**, *14*, 321−333.

(14) (a) Balakin, V. K.; Stanley, A. L.; 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. (b) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* **2002**, *3*, 928−944. (c) Balakin, V. K.; Tkachenko, S. E.; Stanley, A. L.; Okun, I.; Ivashchenko, A. A.; Savchuk, N. P. Property-based design of GPCR-targeted library. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1332−1342. (d) Jacoby, E. A novel chemogenomics knowledge-based ligand design strategy − application to G protein-coupled receptors. *Quant. Struct.-Act. Relat.* **2001**, *20*, 115−123.

(15) Frye, S. V. Structure−activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. *Chem. Biol*. **2001**, *6*, R3−7.

(16) Miller, M. A. Chemical database techniques in drug discovery. *Nature Rev. Drug Discovery* **2002**, *1*, 220−227.

(17) MDL Drug Data Report, MDL ISIS/HOST software, MDL Information Systems, Inc.

(18) Derwent World Drug Index, MDL ISIS/HOST software, Derwent Information Ltd.

(19) Jacoby, E.; Schuffenhauer, A.; Floersheim, P. Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect* **2003**, *16(2)*, 93−102.

(20) Prous Science, DailyDrugNews.com, www.prous.com/home_daily/index.html

(21) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

(22) (a) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubilty and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25. (b) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem*. **2003**, *46*, 1250−1256. (c) 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. (d) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem*. **2002**, *45*, 2615−2623. (e) Walters, W. P.; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.

(23) Böhm, H.-J.; Klebe, G. What can we learn from molecular recognition in protein ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed*. **1996**, *35*, 2588−2614.

(24) (a) Bernard, D.; Coop, A.; MacKerell, A. D. 2D Conformationally sampled pharmacophore: A ligand-based pharmacophore to differentiate δ opioid agonists from antagonists. *J. Am. Chem. Soc*. **2003**, *125*, 3101−3107. (b) Singh, J.; Van Vlijmen, H.; Liao, A.; Lee, W.-C.; Cornebise, M.; Harris, M.; Shu, I.; Gill, A.; Cuervo, J. H.; Abraham, W. M.; Adams, S. P. Identification of potent and novel $\alpha 4\beta 1$ antagonists using in silico screening. *J. Med. Chem*. **2002**, *45*, 2988−2993. (c) Ooms, F.; Wouters, J.; Oscari, O.; Happaerts, T.; Bouchard, G.; Carrupt, P.-A.; Testa, B.; Lambert, D. M. Exploration of the pharmacophore of 3-Alkyl-5-arylimidazolidinediones as new CB_1 cannabinoid receptor ligands and potential antagonists: synthesis, lipophilicity, affinity and molecular modeling. *J. Med. Chem*. **2002**, *45*, 1748−1756. (d) Debnath, A. K. Pharmacophore mapping of a series of 2,4,-Diamino-5-deazapteridine inhibitors of *Mycobacterium avium* complex dihydrofolate reductase. *J. Med. Chem*. **2002**, *45*, 41−53. (e) Poulain, R.; Horvath, D.; Bonnet, B.; Eckhoff, C.; Chapelain, B.; Bodinier, M.-C.; Déprez, B. From hit to lead. Analyzing structure-profile relationships. *J. Med. Chem.* **2001**, *44*, 3391−3401. (f) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. Poulain, R.; Horvath, D.; Bonnet, B.; Eckhoff, C.; Chapelain, B.; Bodinier, M.-C.; Déprez, B. From hit to lead. Analyzing structure-profile relationships. *J. Med. Chem*. **2001**, *44*, 1841−1846. (g) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 117−125.

(25) Teller, D. C.; Okada, T.; Behnke, C. A.; Palczewski, K.; Stenkamp, R. E. Advances in determination of a high-resolution three-dimensional structure of Rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry* **2001**, *40(26)*, 7761−7772.

(26) Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? *Prot.: Struct., Funct. Genet.* **2003**, *50*, 5−25.

(27) (a) 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. (b) Baumann, K. An alignment-indipendent versatile structure descriptors for QSAR and QSPR based on the distribution of molecular features. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 26−35.

(28) (a) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Morsley, R. T.; Sheridan, R. P. Chemical similarity using physichochemical property descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 118−127. (b) Sheridan, P. R.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 915−924.

(29) Zuccotto, F. Pharmacophore features distribution in different classes of compounds. *J. Med. Chem*. **2003**, *46*, 1542−1552.
(30) Website: http://www.jchem.com/doc/user/GenerFP.html
(31) Jollife, I. T. *Principal Component Analysis*; Springer: New York, 1986; Chapter 1, pp 1−5.
(32) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley: New York, 2000.
(33) Gordon, A. *Classification*; Chapman and Hall: London, 1981; Chapter 3, pp 33−53.
(34) Sammon, J. W. A nonlinear mapping for data analysis. *IEEE Trans. Comput.* **1969**, *C(18)*, 401−409.
(35) Kohonen, T. *Self-Organizing Maps*; Springer: New York, 2001.
(36) Ye, J. On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* **1998**, *93*, 401−409.
(37) Dasarathy, B. V. *Nearest Neighbor (NN) Norms: NN Pattern Classication Techniques;* IEEE Computer Society Press: Los Alamitos, CA, 1990.
(38) Harrison, P. W.; Barlin, G. B.; Davies, L. P.; Ireland, S. J.; Matyus, P.; Wong, M. G. Syntheses, pharmacological evaluation and molecular modeling of substituted 6-alkoxyimidazo[1,2-*b*]pyridazines as new ligands for the benzodiazepine receptor. *Eur. J. Med. Chem*. **1996**, *31*, 651−662.
(39) Website: http://www.chemaxon.com/jchem/doc/user/fingerprint.html
(40) Niiyama, K.; Takahashi, H.; Nagase, T.; Kojima, H.; Amano, Y.; Katsuki, K.; Yamakawa, T.; Ozaki, S.; Ihara, M.; Yano, M.; Fukuroda, T.; Nishikibe, M.; Ishikawa, K. Structure−activity relationship of 2-substituted 5,7-diarylcyclopenteno[1,2-*b*]pyridine-6-carboxylic acids as a novel class of endothelin receptor antagonists. *Bioorg. Med. Chem. Lett*. **2002**, *12*, 3041−3045.
(41) Rawson, D. J.; Dack, K. N.; Dickinson, R. P.; James, K. The design and synthesis of a novel series of indole derived selective ETA antagonists. *Bioorg. Med. Chem. Lett*. **2002**, *12*, 125−128.
(42) Roux, S.; Breu, V.; Ertel, S. I.; Clozel, M. Endothelin antagonism with bosentan: a review of potential applications. *J. Mol. Med*. **1999**, *77*, 364−376.
(43) Banks, B. J.; Chubb, N. A. L.; Eshelby, J. J.; Schulz, D. J. *Pyrazoles and their use as endothelin antagonists*; EP1072597; Pfizer Ltd. (GB), Pfizer (US), 2001.
(44) Ishizuka, N.; Ken-ichi, M.; Katsunori, S.; Masafumi, F.; Shin-ichi, M.; Yamamori, T. Structure−activity relationships of a novel class of Entdothelin-A receptor antagonists and discovery of potent and selective receptor antagonists, 2-(Benzo[1,3]dioxol-5-yl)-6-isopropyl-oxy-4-(4-methoxyphenyl)-2*H*-chromene-3-carboxylic acid (S-1255). 1. Study on structure−activity relationships and basic structure crucial for ETA antagonism. *J. Med. Chem*. **2002**, *45*, 2041−2055.
(45) (a) Bauer, H.-U.; Pawelzik, K. R. Qunatifying the neighbourhood preservation of selforganizing feature maps. *Computation* **1997**, *9*, 1291−1303. (b) Flexer, A. On the use of self-organizing maps for clustering and visualization. *Intell.-Data-Analysis* **2001**, *5*, 373−384.
(46) Sauer, W. H. B.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 987−1003.

CI0303013