Binding Site Detection via Mutual Information

Jörg D. Wichard, *Member, IEEE,* Ronald Kühne, Antonius ter Laak

***Abstract*— About 40% of all marketed drugs have the so-called G-protein coupled receptors (GPCRs) as their target protein. There exist more than 800 different GPCRs in humans, of which at least 300 GPCRs are believed to be druggable. Yet, for only two GPCRs there are three-dimensional (3D) protein crystal structures available and consequently little is known about the molecular interactions between pharmacologically active substances and the receptors in this important drug target protein family. A way to overcome the lack of 3D structural information is to deduce GPCR structure-function relationships directly from the enormous amount of small molecule biological activity data that exist for GPCRs. In this work, we suggest a new approach for the detection of interdependence of features in the GPCR sequences and properties of the related ligands based on the mutual information between ligand and sequence space.**

***Index Terms*— Drug Discovery, GPCR, Mutual Information, Chemogenomics**

I. Introduction

G-protein-coupled receptors (GPCRs) are a protein family of transmembrane receptors that modulate sensory perception, chemotaxis, neurotransmission, cell communication and several other vital physiological events. GPCRs comprise one of the largest superfamily in the genome and the estimated numbers of GPCRs vary widely. In a recent analysis Fredriksson et al. [1] identified more than 800 human GPCR sequences and provided a phylogenetic analysis.

The involvement in many biological processes has the consequence that GPCRs play a key role in many pathological conditions, which has led to GPCRs being the target of up to 40 % of today's marketed drugs [2], [3]. The expected future development is also quite promising because it is evident that drugs have still only been developed to affect a relatively small number of GPCRs. In a recent study, Russ and Lampel [4] estimated that there are up to 300 druggable GPCRs in the humane genome.

For the drug discovery process it is very important to understand the interaction between the ligand (in general a small molecule) and the target receptor. Therefore it is useful to have a 3D model of the receptor (the tertiary structure of the protein). The common experimental method of structure determination is X-ray crystallography, but like other membrane proteins, GPCRs are very difficult to crystallize. For several years there was only one mamalien GPCR crystal structure published at atomic resolution. It was the inactive conformation of Bovine-Rhodopsin, the optical receptor protein solved in 2000 by Palczewski et al. [5] with improvements in resolution reported later [6]–[9]. The structure is shown in Figure 1. Recently the first human GPCR crystal structure was published [10]–[12], the β_2 adrenoreceptor (β_2AR) which shows similar structural features as the Bovine-Rhodopsin structure.

J. Wichard and R. Kühne are with the FMP Berlin, Molecular Modelling Group, Robert-Roessle-Str. 10, 13125 Berlin, Germany, (email: {wichard, kuehne}@fmp-berlin.de), Antonius ter Laak is with the Bayer Schering Pharma AG, Computational Chemistry, 13342 Berlin, (email: antoniuster.laak@bayerhealthcare.com)

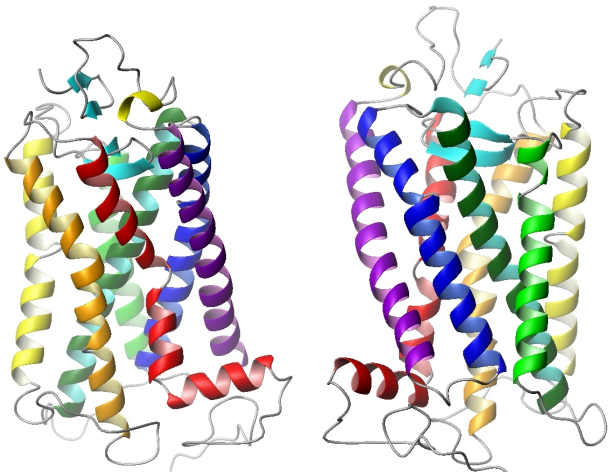

Fig. 1. The crystal structure of Bovine-Rhodopsin as described by Palczewski et al. [5], shown from two sides, the extracellular region on top and the intracellular region on the bottom. The 7 transmembrane domains are marked with different colors: TM1 in purple, TM2 in blue, TM3 in dark green, TM4 in light green, TM5 in yellow TM6 in orange and TM7 in red. The helices are forming a bundle that is perforating the lipid layer of the cell membrane and contains a binding pocket, where the chromophore 11-*cis*-retinylidene is located (not shown in this figure).

In order to overcome the obstacles that arise from the absence of more GPCR structures one could try to generate these structures computationally. The common approach to do so is homology modeling, where the structure of the target protein is modeled using the crystal structure of Bovine-Rhodopsin as a template [13]. Thus structure-based drug discovery which is an important part of the modern drug discovery process, so far plays only a minor role in the lead finding and optimization process of small molecule ligands targeting GPCRs, because homology models based on the few available crystal structures may have limited applicability, see Bywater [14] for a detailed discussion. Another approach to generate 3D structures is the *de novo* structure prediction, wherein the model is generated from first principles [15], [16]. An overview and a comparison of these two approaches could be found in [17].

In this paper, we like to introduce an approach to analyze the relationship between the GPCR sequences and the GPCR ligands without making any explicit use of a 3D model.

The main idea is to estimate the mutual information between the joint distribution of sequence and ligand properties.

Therefor we had to annotate and prepare the GPCR-ligand database that we used for this purpose. The first step was the construction of a non-redundant multiple sequence alignment. This is straightforward for GPCRs thanks to sufficient sequence homology within the seven transmembrane helical domains. The second step is the calculation of molecular descriptors in order to describe the physico-chemical features of the GPCR-ligands. The last step applies mutual information to detect the interdependence of sequence features in determined alignment positions and distinct chemical features in the ligand space. We further used a Monte-Carlo test to assess the validity of our estimates.

II. GPCRs

All GPCRs share a common motif of seven membrane spanning domains joined together by three extracellular and three intracellular loops with an extracellular N terminus and an intracellular C terminus as shown in Figure 1.
The seven trans-membrane helices (TMs) have several highly conserved residues which allow to detect and align them properly. In Table I we listed the alignment positions and the most conserved residues.
The second extracellular loop (ECL2) and the TM3 contain a highly conserved cysteine residue which build a disulphide bond to stabilize the receptor structure. The intracellular loops interact with the G proteins which are mainly involved in the process of signal transduction.
Several classification systems have been used to group the GPCRs into families. Following the work of Fredriksson et al. [1], the human GPCRs can be grouped into five main families named Glutamate, Rhodopsin, Adhesion, Frizzled/Taste2 and Secretin, forming the GRAFS classification system, wherein the Rhodopsin family is the largest one.
Of particular interest to small-molecule drug discovery is the trans-membrane region of the receptor. Most of the small molecule GPCR ligands are supposed to bind in an hypothetical binding pocket within the transmembrane domain [18]. Activation of a GPCR upon ligand binding induces conformational changes in the position of the TMs that act like a switch and result in specific interactions with the G-proteins and the related secondary messangers.

III. Data

A. The Database

We used the GPCR Inhibitor Database from GVKBIO [19] of about 58,000 GPCR ligands referring to more than 200 different GPCR sequences (mostly human and rodents) that were gathered from various sources such as articles published in different journals and patent data. There are several open source databases available now, that might have a smaller number of ligands as the commercial ones, see for example Horn et al. [20] or Okono et al. [21]. We had to do some preprocessing in order to annotate the database and to achieve unique names for the protein sequences. Apparently it is not common, to use unique names for the sequences under investigation, because different authors published their ligand binding data using different names. We decided to use the Swiss-Prot ID for our purpose, removed all ambiguous entries from the database and renamed the remaining ones with unique IDs. After this cleaning process we had about 30,000 GPCR ligands (as 2D-structures in sd-files) referring to 192 GPCRs from the Rhodopsin family.

Domain	Alignment Position	Conserved Amino Acid
TM1	1.31-1.58	N: 1.50
TM2	2.38-2.67	D: 2.50
TM3	3.25-3.55	R: 3.50
TM4	4.39-4.62	W: 4.50
TM5	5.35-5.65	P: 5.50
TM6	6.31-6.59	P: 6.50
TM7	7.32-7.56	P: 7.50

TABLE I
THE POSITIONS OF THE TRANSMEMBRANE DOMAINS IN THE BALLESTEROS WEINSTEIN NUMBERING SCHEME AND THE CONSERVED AMINO ACIDS.

.

B. Sequence Alignment

The Rhodopsin family GPCRs share several highly conserved residues in all seven helices so the multiple sequence alignment is straightforward and was performed with a Profile Hidden Markov Model (HMM) as included in the HMMER software [22]. In addition we adjusted the outcome of the HMM by hand in order to align the second extracellular loop (ECL2) around the conserved Cysteine residue between TM4 and TM5. We compared our HMM alignment of the seven TMs with the GPCR alignments published by Horn et al. [20] and found congruence in almost all cases. The last transmembrane domain (TM7) has an extention that is also known as Helix 8. The extracted TMs and the conserved residues are given in Table I wherein we used the numbering scheme of Ballesteros and Weinstein [23] to describe the alignment position.

C. Molecular Descriptors

According to Todeschini and Consonni [24], a molecular descriptor is the *'final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment'*.
We calculated several molecular descriptors in order to characterize the physical and chemical properties of the compounds in our ligand data base. All molecular descriptors were calculated with the Pipeline Pilot software [25]. The molecular descriptors were simple element counts, molecular property counts, Ghose-Crippen AlogP counts [26] and electrotopological state counts [27], [28]. We listed the most important descriptors in Table II ranked by their mean mutual information that was taken over all alignment positions.

IV. Mutual Information

The mutual information measures the mutual dependence of two random variables. In our case, we are dealing with

discrete random variables, following the mathematical definitions of Shannon's original work [29]. A detailed introduction including rigorous mathematical proofs of the theorems below could be found in [30].

If we consider a discrete random variable X with n possible values taken from the alphabet $A_x = \{x_1, x_2, \ldots, x_n\}$ with the associated probability distribution $P(X) = \{p(x_1), p(x_2), \ldots, p(x_n)\}$ and $\sum_{i=1}^{n} p(x_i) = 1$, then the entropy $H(X)$ is defined as

$$H(X) := -\sum_{i=1}^{n} p(x_i) \log(p(x_i)), \tag{1}$$

with the following properties:

- $H(X)$ is maximized if the probabilities are equally distributed over all letters off the alphabet $p(x_i) = \frac{1}{n}$.
- If there is only one possible outcome for X, then $H(X) = 0$. This reflects the common interpretation of *entropy as as a measure of uncertainty about a random variable*.
- In the case of impossible events $X = x_i$ with $p(x_i) = 0$ we define $p(x_i) \log(p(x_i)) \equiv 0$ because in the limit case we have $\lim_{\rho \to 0} \rho \log(\rho) = 0$.
- We use the natural logarithm in the following work. The basis of the logarithm doesn't affect any of the above properties but could be used to scale the entropy.

The entropy can be extended to the case of two random variables X and Y. With the alphabet $A_y = \{y_1, y_2, \ldots, y_m\}$ and the joined probability distribution $P(X,Y) = \{p(x_1, y_1), p(x_1, y_2), \ldots, p(x_n, y_m)\}$ we define the joined entropy $H(X,Y)$ as

$$H(X,Y) := -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log(p(x_i, y_j)), \tag{2}$$

wherein $p(x_i, y_j)$ denotes the probability of the joined occurrence of x_i and y_j. If the two random variables X and Y are statistically independent the joined probabilities factorize and the joined entropy turns into

$$H(X,Y) = H(X) + H(Y). \tag{3}$$

The conditional entropies are given by

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) \\ H(Y|X) &= H(X,Y) - H(X). \end{aligned} \tag{4}$$

If the entropy $H(X)$ is regarded as a measure of uncertainty about the random variable X, then $H(X|Y)$ can be seen as the average amount of uncertainty remaining about X after Y is known. The mutual information $I(X,Y)$ between two random variables X and Y is defined as

$$\begin{aligned} I(X,Y) &:= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{5}$$

The mutual information as defined in Eq. 5 is a basic concept of information theory and provides a general measure of interdependence between random variables.

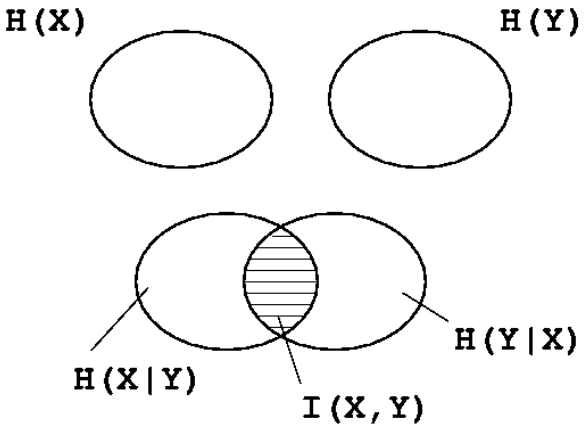

Fig. 2. The Venn-Diagram of the mutual information. If the entropy $H(X)$ is regarded as a measure of uncertainty about the random variable X then the mutual information $I(X,Y)$ measures how much the uncertainty of X is reduced if Y has been observed.

From Eq. 5 and Eq. 3 we see, that the mutual information is zero if X and Y are statistically independent. In this sense mutual information quantifies the distance between the joint distribution of X and Y and what the joint distribution would be if X and Y were independent random variables. Since it makes no assumption about the type of relation between X and Y, mutual information is sometimes considered to be an extension of the linear correlation coefficient [31]. But in contrast to the correlation coefficient, the mutual information is always nonnegative. Here are some basic properties of the mutual information:

$$\begin{aligned} I(X,Y) &= I(Y,X) \\ H(X,Y) &\geq I(X,Y) \geq 0 \\ I(X,X) &= H(X). \end{aligned} \tag{6}$$

In Figure 2 we depicted the relation of entropy, joined entropy and mutual information in the case of two random variables. In this context, mutual information quantifies the reduction of uncertainty of a random variable X (as given by $H(X)$) if Y is known and *vice versa*. Methods of feature evaluation were developed and discussed in different areas by many researchers, see for example Guyon et al. [32] for a detailed discussion. Mutual information was shown in the literature to be very robust and precise to evaluate a feature set [33].

A. Estimating Mutual Information

In order to calculate the mutual information of two random variables, one should know the associated probability distributions. In general these probabilities are not known and have to be estimated. A common way to do this is the histogram approach which is straightforward. This approach calculates the relative frequencies in the histogram bins in order to estimate the probability distribution of the random variable. In general the choice of the number and the width of the bins is crucial and has some effects to the outcome of

the estimate, see Steuer et al. [34] for a detailed discussion. In our case the situation is more comfortable because the 20 amino acids which are the building blocks of the proteins are the canonical alphabet that we use. The same is true for almost all molecular descriptors that are discrete, like element counts or molecular property counts. In the case of continuous descriptors[1] we follow a suggestion of Li [31] who proposed the number of bins N_{bin} for non-Gaussian distributions should be

$$N_{bin} = \log_2(K) + 1 + \log_2(1 + \widehat{\kappa}\sqrt{K/6}),$$

wherein $\widehat{\kappa}$ is the estimated kurtosis of the distribution and K is the number of samples. This is a generalization of *Sturges' rule* which is an established rule of thumb for binning Gaussian distributions [35].

Let us consider K simultaneous measurements of the two random variables X and Y with the alphabets $A_x = \{x_1, x_2, \ldots, x_n\}$ and $A_y = \{y_1, y_2, \ldots, y_m\}$ and let k_{ij} be the total number of measurements with $X = x_i$ and $Y = y_j$. Then the probabilities $p(x_i, y_j)$ are approximated by the corresponding relative frequencies of the pairwise occurrence

$$p(x_i, y_j) = \frac{k_{ij}}{K}$$

and accordingly for the single probabilities

$$\begin{aligned} p(x_i) &= \frac{k_i}{K} \\ p(y_j) &= \frac{k_j}{K}. \end{aligned}$$

The mutual information $I(X, Y)$ of the two random variables X and Y turns to

$$I(X, Y) = \log(K) + \frac{1}{K}\sum_{ij} k_{ij} \log\left(\frac{k_{ij}}{k_i k_j}\right). \qquad (7)$$

B. Significance Tests

The degree of interdependence between two random variables could be measured in several ways, but the central question is how much interdependence is necessary to exclude more trivial explanations. All quantifiers of interdependence (like mutual information or cross-correlation) show fluctuations when estimated, but the distributions are not available analytically. It is therefore necessary to use Monte Carlo techniques to assess the significance of results. We calculated the mutual information on randomly drawn subsets of the whole data set but only 50% of the size and repeated this 100 times. The variances of these estimates were very small and in all cases below 2% of the mean values, which means that we have very stable estimates.

V. Binding Site Detection

The mutual information of the different TMs and for the 20 most important molecular descriptors is shown in Figure 4. It offers very interesting insights into the interrelation between ligands and amino acid positions.

[1]We have only a few molecular descriptors that have continuous values: Molecular weight, AlogP, estimated solubility and molecular surface area.

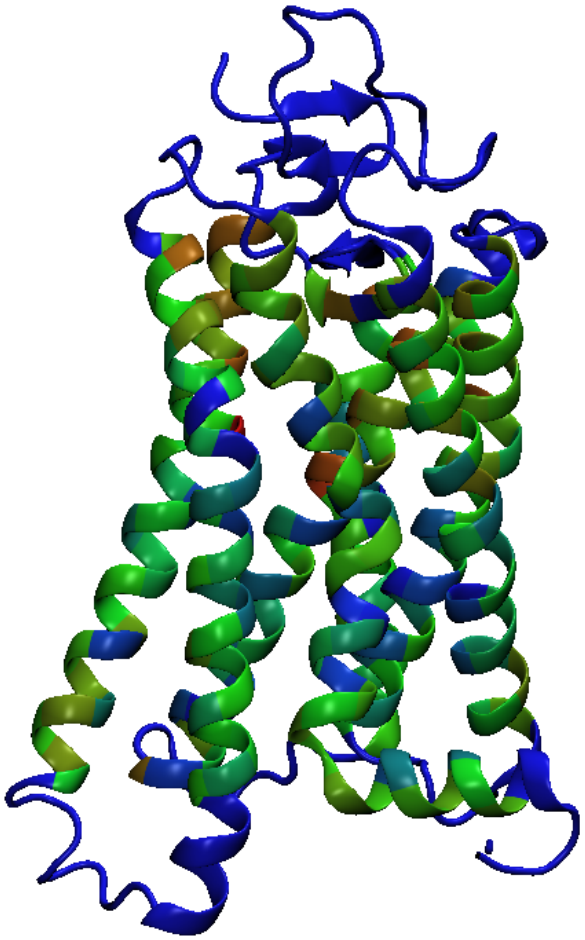

Fig. 3. The Rhodopsin structure wherein the color of the residues represents the mean mutual information. The positions with the highest mutual information are facing the inner part of the receptor and are located in the part that points to the extracellular site, where the hypothetical small molecule binding pocket is supposed to be.

Rank	Descriptor	Mean MI
1	Number of Aromatic Bonds	0.132
2	C-Count	0.132
3	Number of Hydrogens	0.125
4	Number of Chain Assemblies	0.119
5	Number of Atoms	0.119
6	Electrical State Count ssCH2	0.117
7	Number of Bonds	0.116
8	N-Count	0.116
9	Electrical State Count aaN	0.116
10	Molecular Weight	0.115
11	Number of Chains	0.115
12	Electrical State Count sssCH	0.112
13	Molecular Solubility	0.112
14	Electrical State Count dO	0.111
15	Molecular Surface Area	0.111
16	O-Count	0.111
17	Electrical State Count dssC	0.101
18	Electrical State Count aaCH	0.101
19	Number of H-Acceptors	0.098
20	Number of Ring Bonds	0.094

TABLE II

The twenty most important molecular descriptors ranked by their mean mutual information

In TM1 are the positions 1.32, 1.35, 1.39 and 1.43 very prominent, reflecting the helical structure of the domain. This becomes even more evident if we use the mean mutual information to color the Rhodopsin structure as shown in Figure 3. The positions with the highest mutual information are facing the inner part of the receptor and are located in the part that points to the extracellular site. This is both in accordance with the postulated binding pocket in this region [18] and thus reflects structural information that we generated without any explicit structural assumption.

The most important outcome of our study is the detailed connection of sequence and ligand features that allows to rank the chemical features of the ligands for each position of the binding site. This is a useful tool for the construction of focused screening libraries [36] in the early stages of the drug discovery process.

VI. Applications to Library Design

A common method in the drug discovery process is High Throughput Screening (HTS), where a huge collection of compounds from a compound library is tested against the biological target of interest and to identify new active compounds as candidates for the further drug development process. In order to reduce costs and to speed up the screening process it is preferable to use only a smaller subset of a huge library. In the recent years, several methods were described how to design, select or synthesize gene family-focused or -biased libraries (see [37]–[39] for an overview). The mutual information approach allows to detect the interdependence of sequence features and distinct chemical features in the ligand space. Moreover it could be used to rank the molecular descriptors of the ligands and the alignment position of the GPCRs with respect to their importance and to learn a mapping between the sequences and their ligands. The main paradigm of this approach is that molecules sharing a high similarity to existing ligands have an enhanced probability to share the biological profile. The crucial point is the definition of *similarity*. Therefore we have to establish a distance measure in the space of the sequences and in the space of their ligands by building vectors from the most important alignment positions and the most important molecular descriptors.

First we have to select the N positions in the alignment with the highest mean mutual information. There is no general rule how many positions should be selected, i.e. how to set the cutoff in the mean mutual information. Starting from structural considerations, Surgand et al. [18] identified 30 critical positions supposed to line the generic binding cavity inside the TM region and constructed phylogenetic trees for several GPCR sub-families. Based on an analysis of the functional conservation indices for the 7 TM helices of the Rhodopsin family GPCRs, Kratochwil et al. [40] identified 28 critical positions in the 7 TMs. We decided to select 46 positions in 7 TMs based on our mutual information analysis. In the first step, a descriptor for each of the 20 amino acids is created, for example the hydropathicity [41] or simply the mutability with respect to a certain amino acid, that could

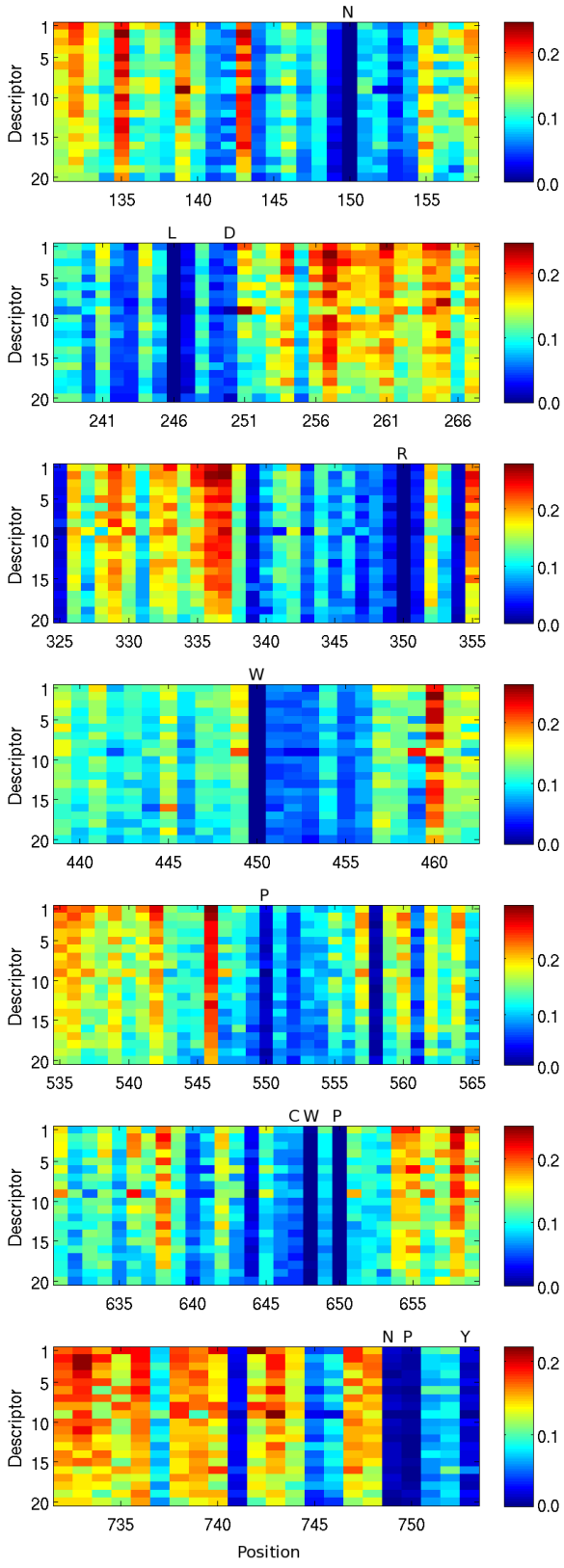

Fig. 4. The mutual information of the seven transmembrane regions. The positions are named according to the Ballesteros Weinstein numbering scheme. The conserved alignment positions are marked on the top of each image. The twenty most important molecular descriptors are ranked as defined in Table II.

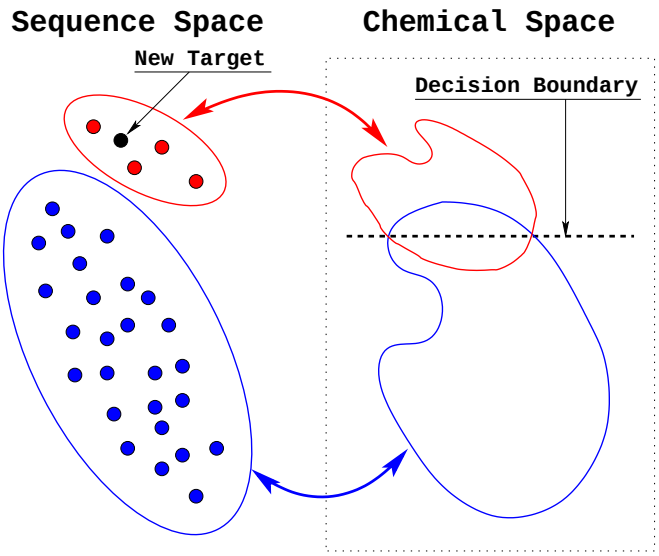

Fig. 5. See Section VI for a description

be taken from a scoring matrix (BLOSUM, PAM)), which encodes empirically derived substitution probabilities. These values are used to build the vector $\vec{x}^k$ for the k-th GPCR sequence

$$\vec{x}^k = (x_1, \ldots, x_N), \tag{8}$$

wherein x_i denotes the descriptor of the amino acid at the i-th selected alignment position.
The molecular descriptors for the ligands are calculated as described in Section III-C. The 20 most important descriptors from Table II are combined to build the vector

$$\vec{y}^l = (y_1, \ldots, y_{20}) \tag{9}$$

for the l-th compound.
With these definitions we created two real-valued vector spaces wherein we can use the euclidean metric as the desired similarity measure. Together with our annotated database described in Section III-A this offers a way to build focused compound libraries for new GPCR-targets or smaller groups of target receptors. We show the main principle in Figure 5. If we consider the task to find ligands for specific GPCR target, than we align the sequence of the new target according to alignment of our GPCR database and calculate the vector $\vec{x}^{new}$ of the sequence features from Eq. 8 and look for the closest neighbors in sequence space. If we shift to the chemical space, we can build to different sets of compounds. One set consists of compounds that interact with the closest neighbors in sequence space and one set of compounds that interact with the receptors apart. Now we can use the standard machine leaning tools for classification in order to create a decision boundary that wors as a filter for compound selection and look for a subset of a huge combinatorial library that is supposed to be focused to the specific new GPCR target.

VII. Conclusion

We introduced mutual information to study the mutual dependence between a wide array of small molecular-weight ligands on a wide array of GPCR targets. Although we focus on GPCRs, it is expected that such chemogenomics approaches which are connecting the chemical and the bioinformatics world could open new perspectives in drug discovery in respect of other target families.

Acknowledgment

The authors would like to thank the members of the Molecular Modelling Group at FMP Berlin and the Computational Chemistry Department of Bayer-Schering Pharma for fruitful discussions. The pictures of the Bovine-Rhodopsin structure in Figure 1 were prepared with the program MOLMOL [42].

References

[1] R. Fredriksson, M. Lagerström, L. Lundin, H. Schiöth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints, Mol. Pharmacol. 63 (2003) 1256–1272.
[2] A. Hopkins, C. Groom, The druggable genome, Nature Reviews Drug Discovery 1 (9) (2002) 727–730.
[3] D. Filmore, It's a GPCR world, Modern Drug Discovery 7 (11) (2004) 24–28.
[4] A. Russ, S. Lampel, The druggable genome: An update, Drug discovery today 10 (23/24) (2005) 1607–1610.
[5] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. Fox, I. L. Trong, D. Teller, T. Okada, R. Stenkamp, M. Yamamoto, M. Miyano, Crystal structure of rhodopsin: A G-protein-coupled receptor, Science 289 (5480) (2000) 739–745.
[6] D. Teller, T. Okada, C. Behnke, K. Palczewski, R. Stenkamp, Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors, Biochemistry 40 (2001) 7761–7772.
[7] T. Okada, Y. Fujiyoshi, M. Silow, J. Navarro, E. Landau, Y. Shichida, Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography, Proc. Natl. Acad. Sci. USA 99 (9) (2002) 5982–5987.
[8] J. Li, P. Edwards, M. Burghammer, C. Villa, G. Schertler, Structure of bovine rhodopsin in a trigonal crystal form, J. Mol. Biol. 353 (5) (2004) 1409–1438.
[9] T. Okada, M. Sugihara, A. Bondar, M. Elstner, P. Entel, V. Buss, The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure, J. Mol. Biol. 342 (2) (2004) 571–583.
[10] V. Cherezov, D. Rosenbaum, M. Hanson, S. Rasmussen, F. Thian, T. S. Kobilka, H. Choi, P. Kuhn, W. Weis, B. Kobilka, R. Stevens, High-Resolution Crystal Structure of an Engineered Human β 2 Adrenergic G-Protein Coupled Receptor, Science 318 (5854) (2007) 1258 – 1265.
[11] S. Rasmussen, H. Choi, D. Rosenbaum, T. Kobilka, F. Thian, P. Edwards, M. Burghammer, V. Ratnala, R. Sanishvili, R. Fischetti, G. Schertler, W. Weis, B. Kobilka, Crystal structure of the human β-2 adrenergic G-protein-coupled receptor, Nature (450) (2007) 383–387.
[12] D. Rosenbaum, V. Cherezov, M. Hanson, S. Rasmussen, F. Thian, T. Kobilka, H. Choi, Y. X., W. Weis, R. Stevens, B. Kobilka, GPCR Engineering Yields High-Resolution Structural Insights into β 2 Adrenergic Receptor Function, Science 318 (5854) (2007) 1266 – 1273.
[13] J. Ballesteros, K. Palczewski, G protein-coupled receptor drug discovery: Implications from the crystal structure of rhodopsin, Curr. Opin. Drug Discov. Devel. 4 (5) (2001) 561–574.
[14] R. Bywater, Location and nature of the residues important for ligand recognition in G-protein coupled receptors, Journal of molecular recognition 18 (2005) 60–72.
[15] O. Becker, Y. Marantz, S. Shacham, B. Inbal, A. Heifetz, S. Kalid, O. Bar-Haim, D. Warshaviak, M. Fichman, S. Noiman, G-protein-coupled receptors: In silico drug discovery in 3D, Proc. Natl. Acad. Sci. USA 101 (3).

[16] N. Vaidehi, W. Floriano, R. Trabanino, S. Hall, W. Freddolino, E. Choi, G. Zamanakos, W. Goddard, Prediction of structure and function of G-protein-coupled receptors, Proc. Natl. Acad. Sci. USA 99 (20) (2002) 12622–12627.

[17] S. Schlyer, R. Horuk, I want a new drug: G-protein-coupled receptors in drug development, Drug Discovery Today 11 (11–12) (2006) 481–493.

[18] J. Surgand, J. Rodrigo, E. Kellenberger, D. Rognan, A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors, Proteins: Structure, Function and Bioinformatics 62 (2) (2005) 509–538.

[19] GVKBIO: GPCR Inhibitor Database, http://www.gvkbio.com.

[20] F. Horn, J. Weare, M. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne, G. Vriend, GPCRDB: An information system for G-protein-coupled receptors, Nucleic Acids Res. 26 (1) (1998) 275–279.

[21] Y. Okuno, J. Yang, K. Taneishi, H. Yabuuchi, G. Tsujimoto, GLIDA: GPCR-ligand database for chemical genomic drug discovery, Nucl. Acids Res. 34 (2006) 673–677.

[22] S. Eddy, Profile hidden markov models, Bioinformatics 14 (9) (1998) 755–763.

[23] J. Ballesteros, H. Weinstein, Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors, Methods in Neurosciences 25 (1995) 366–428.

[24] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, New York, 2000.

[25] Pipeline Pilot, http://www.scitegic.com/.

[26] A. Ghose, V. Viswanadhan, J. Wendoloski, Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods, Journal of Physical Chemistry A 102 (21) (1998) 3762–3772.

[27] L. Lowell H. Hall, B. Mohney, L. Kier, The electrotopological state: structure information at the atomic level for molecular graphs, J. Chem. Inf. and Comput. Sci. 31 (1) (1991) 76–82.

[28] L. Hall, L. Kier, The e-state as the basis for molecular structure space definition and structure similarity, J. Chem. Inf. and Comput. Sci. 40 (3) (2000) 784–791.

[29] C. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423 and 623–656.

[30] R. Gray, Entropy and Information Theory, Springer, 1990.

[31] W. Li, Mutual information functions versus correlation functions, Journal of Statistical Physics 60 (1990) 823–837.

[32] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature Extraction. Foundations and Applications, Studies in Fuzziness and Soft Computing, Springer Verlag, 2006.

[33] D. Xu, J. Principe, Feature evaluation using quadratic mutual information, in: Neural Networks: Proceedings of IJCNN 2001, Vol. 1, 2001, pp. 459–463.

[34] R. Steuer, J. Kurths, C. O. Daub, J. Weise, J. Selbig, The mutual information: Detecting and evaluating dependencies between variables, Bioinformatics 18 (2) (2002) 231–240.

[35] H. A. Sturges, The choice of a class-interval, J. Am. Statist. Assoc. 21 (1926) 65–66.

[36] K. Balakin, S. Tkachenko, S. Lang, I. Okun, A. Ivashchenko, N. Savchuk, Property-based design of GPCR-targeted library, Journal of Chemical Information and Computer Sciences 42 (6) (2002) 1332–1342.

[37] R. Crossley, The design of screening libraries targeted at G-protein coupled receptors, Current Topics in Medicinal Chemistry 4 (6) (2004) 581–588.

[38] J. L. Miller, Recent developments in focused library design: Targeting gene-families, Current Topics in Medicinal Chemistry 6 (1) (2006) 19–29.

[39] K. Balakin, A. Kozintsev, A. Kiselyov, N. Savchuk, Rational design approaches to chemical libraries for hit identification, Curr. Drug. Discov. Technol. 3 (1) (2006) 49–65.

[40] N. Kratochwil, P. Malherbe, L. Lindemann, M. Ebeling, M. Hoener, A. Mhlemann, R. . P. Porter, M. Stahl, P. Gerber, An automated system for the analysis of G-protein-coupled receptor transmembrane binding pockets: Alignment, receptor-based pharmacophores, and their application, J. Chem. Inf. Model. 45 (5) (2005) 1324 –1336.

[41] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (5) (1982) 105–132.

[42] R. Koradi, M. Billeter, K. Wütherich, MOLMOL: A program for display and analysis of macromolecular structures, J. Mol. Graphics 14 (1996) 51–55.