Virtual screening of GPCRs: An in silico chemogenomics approach

Laurent Jacob, Brice Hoffmann, Véronique Stoven, Jean-Philippe Vert

BMC Bioinformatics

Research article

Virtual screening of GPCRs: An *in silico* chemogenomics approach

Laurent Jacob*†[1,2,3], Brice Hoffmann†[1,2,3], Véronique Stoven[1,2,3] and Jean-Philippe Vert[1,2,3]

Address: [1]Mines ParisTech, Centre for Computational Biology, 35 rue Saint-Honoré, F-77305, Fontainebleau, France, [2]Institut Curie, Paris, F-75248, France and [3]INSERM, U900, Paris, F-75248, France

Email: Laurent Jacob* - laurent.jacob@mines-paristech.fr; Brice Hoffmann - brice.hoffmann@mines-paristech.fr; Véronique Stoven - veronique.stoven@mines-paristech.fr; Jean-Philippe Vert - jean-philippe.vert@mines-paristech.fr

* Corresponding author †Equal contributors

Published: 6 September 2008

BMC Bioinformatics 2008, **9**:363 doi:10.1186/1471-2105-9-363

Received: 3 April 2008
Accepted: 6 September 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/363

Abstract

Background: The G-protein coupled receptor (GPCR) superfamily is currently the largest class of therapeutic targets. *In silico* prediction of interactions between GPCRs and small molecules in the transmembrane ligand-binding site is therefore a crucial step in the drug discovery process, which remains a daunting task due to the difficulty to characterize the 3D structure of most GPCRs, and to the limited amount of known ligands for some members of the superfamily. Chemogenomics, which attempts to characterize interactions between all members of a target class and all small molecules simultaneously, has recently been proposed as an interesting alternative to traditional docking or ligand-based virtual screening strategies.

Results: We show that interaction prediction in the chemogenomics framework outperforms state-of-the-art individual ligand-based methods in accuracy both for receptor with known ligands and without known ligands. This is done with no knowledge of the receptor 3D structure. In particular we are able to predict ligands of orphan GPCRs with an estimated accuracy of 78.1%.

Conclusion: We propose new methods for *in silico* chemogenomics and validate them on the virtual screening of GPCRs. The methods represent an extension of a recently proposed machine learning strategy, based on support vector machines (SVM), which provides a flexible framework to incorporate various information sources on the biological space of targets and on the chemical space of small molecules. We investigate the use of 2D and 3D descriptors for small molecules, and test a variety of descriptors for GPCRs. We show that incorporating information about the known hierarchical classification of the target family and about key residues in their inferred binding pockets significantly improves the prediction accuracy of our model.

Background

The G-protein coupled receptor (GPCR) superfamily is comprised of an estimated 600–1,000 members and is the largest known class of molecular targets with proven therapeutic value. They are ubiquitous in our body, being involved in regulation of every major mammalian physiological system [1], and play a role in a wide range of disorders including allergies, cardiovascular dysfunction, depression, obesity, cancer, pain, diabetes, and a variety of central nervous system disorders [2-4]. They are integral membrane proteins sharing a common global topology that consists of seven transmembrane alpha helices, an

intracellular C-terminal, an extracellular N-terminal, three intracellular loops and three extracellular loops. There are four main classes of GPCRs (A, B, C and D) defined in terms of sequence similarity [5]. Their location on the cell surface makes them readily accessible to drugs, and 30 GPCRs are the targets for the majority of best-selling drugs, representing about 40% of all prescription pharmaceuticals on the market [6]. Besides, the human genome contains several hundreds unique GPCRs which have yet to be assigned a clear cellular function, suggesting that they are likely to remain an important target class for new drugs in the future [7]. Predicting interactions *in silico* between small molecules and GPCRs is not only of particular interest for the drug industry, but also a useful step for the elucidation of many biological process. First, it may help to decipher the function of so-called *orphan* GPCRs, for which no natural ligand is known. Second, once a particular GPCR is selected as a target, it may help in the selection of promising molecule candidates to be screened *in vitro* against the target for lead identification.

In silico virtual screening of GPCRs is however a daunting task, both for receptor-based approaches (also called docking) and for ligand-based approaches. The former relies on the prior knowledge of the 3D structure of the protein, in a context where only two GPCR structures are currently known (bovine rhodopsin and human β_2-adrenergic receptor). Indeed, GPCRs, like other membrane proteins, are notoriously difficult to crystallize. As a result, docking strategies for screening small molecules against GPCRs are often limited by the difficulty to model correctly the 3D structure of the target. To circumvent the lack of experimental structures, various studies have used 3D structural models of GPCRs built by homology modeling using bovine rhodopsin as a template structure. Docking a library of molecules into these modeled structures allowed the recovery of known ligands [8-11], and even identification of new ligands [12,13]. However, docking methods still suffer from docking and scoring inaccuracies, and homology models are not always reliable-enough to be employed in target-based virtual screening. Methods have been proposed to enhance the quality of the models for docking studies by global optimization and flexible docking [9], or by using different sets of receptor models [11]. Nevertheless, these methods have been applied only to class A receptors and they are expected to show limited performances for GPCRs sharing lower sequence similarity with rhodopsin, especially in the case of receptors belonging to classes B, C and D. Alternatively, ligand-based strategies, in particular quantitative structure-activity relationship (QSAR), attempt to predict new ligands from previously known ligands, often using statistical or machine learning approaches. Ligand-based approaches are interesting because they do not require the knowledge of the target 3D structure and can benefit from the discovery of new ligands. However, their accuracy is fundamentally limited by the amount of known ligands, and degrades when few ligands are known. Although these methods were successfully used to retrieve strong GPCR binders [14], they are efficient for lead optimization within a previously identified molecular scaffold, but are not appropriate to identify new families of ligands for a target. At the extreme, they cannot be pursued for the screening of orphan GPCRs. In this paper, we present a contribution to the screening of GPCRs, that is complementary to the above docking and ligand-based approaches. The method is related to ligand-based approaches, but because it allows to share information between different GPCRs, it can be used for orphan GPCRs, possibly in parallel to docking methods in order to increase the prediction quality.

Indeed, instead of focusing on each individual target independently from other proteins, a recent trend in the pharmaceutical industry, often referred to as *chemogenomics*, is to screen molecules against several targets of the same family simultaneously [15,16]. This systematic screening of interactions between the chemical space of small molecules and the biological space of protein targets can be thought of as an attempt to fill a large 2D *interaction matrix*, where rows correspond to targets, columns to small molecules, and the (i, j)-th entry of the matrix indicates whether the j-th molecule can bind the i-th target. While in general the matrix may contain some description of the strength of the interaction, such as the association constant of the complex, we will focus in this paper on a simplified description that only differentiates binding from non-binding molecules, which results in a binary matrix of target-molecule pairs. This matrix is already sparsely filled with our current knowledge of protein-ligand interactions, and chemogenomics attempts to fill the holes. While classical docking or ligand-based virtual screening strategies focus on each single row independently from the others in this matrix, *i.e.*, treat each target independently from each others, the chemogenomics approach is motivated by the observation that similar molecules can bind similar proteins, and that information about a known interaction between a ligand and a GPCR could therefore be a useful hint to predict interaction between similar molecules and similar GPCRs. This can be of particular interest when, for example, a particular target has few or no known ligands, but similar proteins have many: in that case it is tempting to use the information about the known ligands of similar proteins for a ligand-based virtual screening of the target of interest. In this context, we can formally define *in silico* chemogenomics as the problem of predicting interactions between a molecule and a ligand (*i.e.*, a hole in the matrix) from the knowledge of all other known interactions or non-interactions (*i.e.*, the known entries of the matrix).

Recent reviews [15-18] describe several strategies for *in silico* chemogenomics. A first class of approaches, called *ligand-based chemogenomics* by [18], pool together targets at the level of families (such as GPCR) or subfamilies (such as purinergic GPCR) and learn a model for ligands at the level of the family [19,20]. Such strategies could be facilitated by the design of libraries of annotated ligands [21]. Other approaches, termed *target-based chemogenomic* approaches by [18], cluster receptors based on ligand binding site similarity and again pool together known ligands for each cluster to infer shared ligands [22]. Finally, a third strategy termed *target-ligand* approach by [18] attempts to predict ligands for a given target by leveraging binding information for other targets in a single step, that is, without first attempting to define a particular set of similar receptors. This strategy was pioneered by [23]. [24] predicted ligands of orphan GPCR. They merged descriptors of ligands and targets to describe putative ligand-receptor complexes, and used SVM to discriminate real complexes from ligand-receptors pairs that do not form complexes. A similar approach termed *proteochemometrics* was used in [25,26] to correlate ligand-receptor descriptions to the corresponding binding affinities. [27] followed a similar idea to [24] with different descriptors, and showed in particular that the SVM formulation allows to generalize the use of vectors of descriptors to the use of positive definite kernels to describe the chemical and the biological space in a computationally efficient framework. [27] were not able to show, however, significant benefits with respect to the individual approach that learns a separate classifier for each GPCR (except in the case of orphan GPCRs, for which their approach performed better than the baseline random classifier). Recently, in the context of predicting interactions between peptides and different alleles of MHC-I molecules, [28] followed a similar approach and highlighted the importance of choosing adequate descriptors for small molecules and targets. They obtained state-of-the-art prediction accuracy for most MHC-I allele, in particular for those with few known binding peptides. [29] on the other hand applied this approach to predict interaction between various potential targets including GPCRs, enzymes and ion channels. Using general descriptors for targets, they obtained predictors that were more accurate than state-of-the-art individual methods both for the orphan targets and for the targets for which some ligands were already known.

In this paper we go one step further in this direction and present an *in silico* chemogenomics approach specifically tailored for the screening of GPCRs, although the method could in principle be adapted to other classes of therapeutic targets. We follow the idea of [24] and the algorithmic trick of [27], which allows us to systematically test a variety of descriptors for both the molecules and the GPCRs. We test 2D and 3D descriptors to describe molecules, and five ways to describe GPCRs, including a description of their relative positions in current hierarchical classifications of the superfamily, and information about key residues likely to be in contact with the ligand. We evaluate the performance of all combinations of these descriptions on the data of the GLIDA database [30], which contains 34686 reported interactions between human GPCRs and small molecules, and observe that the choice of the descriptors has a significant impact on the accuracy of the models. However, in all cases, we obtained significant improvements of the prediction accuracy with respect to the individual learning setting.

Data

We used the GLIDA GPCR-ligand database [30] which includes 22964 known ligands for 3738 GPCRs from human, rat and mouse. The ligand database contains highly diverse molecules, from ions and very small molecules up to peptides, and a significant number of duplicates. These redundancies were eliminated. Elimination of duplicates present in the GLIDA database was important here because it could have led to over-optimistic evaluation in the cross-validation procedure described below. The remaining molecules were further filtered in order to satisfy two constraints. First, our method relies on the evaluation of similarities between molecules using kernels, which makes sense only if the molecules are comparable in size. Second, since the long term goal is to identify drug candidates targeting GPCRs, it was important to retain drug-like compounds, i.e. molecules having the adequate physico-chemical characteristics to be potential drugs candidates satisfying ADME criteria [31]. Therefore, to only keep drug-like compounds, we filtered the GLIDA database using the filter program (OpenEye Scientific Software) with standard parameters, which removes molecules according to calculated properties such as molecular weight, hydrogen bond donor and acceptor count, number of rotatable bonds, ring size and number etc... as discussed in [32-35]. For example, only molecules of molecular weights ranging from 150 Da to 450 Da were kept (the classically accepted range for drugs), since the aim was to evaluate if statistical learning was possible on drug-like compounds. Another example was the elimination of molecules with more than 10 rotatable bonds (although most of them being already filtered out on the molecular weight criterion). Indeed, they correspond to very flexible molecules that are not suitable for the use of 3D descriptors. Overall these filters retained 2446 molecules, available under a 2D description file in the GLIDA data bank, and giving 4051 interactions with the human GPCRs. The number of molecules retained is only a small fraction of the GLIDA database, but it corresponds to all drug-like compounds of this database. For each positive interaction given by this restricted set, we generated a neg-

ative interaction involving the same receptor and one of the ligands that was in the database and that was not indicated as one of its ligands. This may have generated a few false negative points in our benchmark, and it would be interesting to use experimentally tested negative interactions. However, the mean similarity between the different ligands in the database using the Tanimoto kernel, a classical normalized similarity measure for ligands which is later used in our method, is quite low (0.13). Besides, only 6.7% of the ligands have a mean similarity of more than 0.2 to the other ligands. This suggests that even if false negative have to be expected, this method to generate negative interaction is a reasonable approximation. We loaded the sequences of all GPCRs that are able to bind any of these ligands, which resulted in 80 sequences, all corresponding to human GPCRs. The retained GPCRs were significantly diverse in sequence, most of them sharing 15% to 50% pairwise sequence similarities. Furthermore, they belong to various families, according to the GLIDA classification. They are found in several subfamilies of class A (rhodopsin-like receptors), classes B (secretin family) and C (metabotropic family). In the GLIDA database, GPCRs are classified in hierarchy (as mentioned above) which was also loaded for use in the hierarchy kernel.

Methods

In this section, we first review the methods proposed by [24,27] for *in silico* chemogenomics with SVM, before presenting the particular descriptors we propose to use for molecules and GPCRs within this framework.

In silico chemogenomics with machine learning

We consider the problem of predicting interactions between GPCRs and small molecules. For this purpose we assume that a list of target/small molecule pairs $\{(t_1, m_1),...,(t_n, m_n)\}$, known to interact or not, is given. Such information is often available as a result of systematic screening campaigns in the pharmaceutical industry, or on dedicated databases. Our goal is then to create a model to predict, for any new candidate pair (t, m), whether the small molecule m is likely to bind the GPCR t.

A general method to create the predictive model is to follow these four steps:

1. Choose n_{tar} descriptors to represent each GPCR target t in the biological space by a n_{tar}-dimensional vector $\Phi_{tar}(t) = \left(\Phi^1_{tar}(t), ..., \Phi^{n_{tar}}_{tar}(t)\right)$;

2. In parallel, choose n_{mol} descriptors to represent each molecule m in the chemical space by a n_{mol}-dimensional vector $\Phi_{mol}(m) = \left(\Phi^1_{mol}(m), ..., \Phi^{n_{mol}}_{mol}(m)\right)$;

3. Derive a vector representation of a candidate target/molecule complex $\Phi_{pair}(t, m)$ from the representations of the target $\Phi_{tar}(t)$ and of the molecule $\Phi_{mol}(m)$;

4. Use a statistical or machine learning method to train a classifier able to discriminate between binding and non-binding pairs, using the training set of binding and non-binding pairs $\{\Phi_{pair}(t_1, m_1),..., \Phi_{pair}(t_n, m_n)\}$

While the first two steps (selection of descriptors) may be specific to each particular chemogenomics problem, the last two steps define the particular strategy used for *in silico* chemogenomics. For example, [24,36] proposed to concatenate the vectors $\Phi_{tar}(t)$ and $\Phi_{mol}(m)$ to obtain a $(n_{tar} + n_{mol})$-dimensional vector representation of the ligand-target complex $\Phi_{pair}(t, m)$, and to use a SVM as a machine learning engine. [27] followed a slightly different strategy for the third step, by forming descriptors for the pair (t, m) as *product* of small molecule and target descriptors. More precisely, given a molecule m described by a vector $\Phi_{mol}(m)$ and a GPCR t described by a vector $\Phi_{tar}(t)$, the pair (t, m) is represented by the tensor product:

$$\Phi_{pair}(t, m) = \Phi_{tar}(t) \otimes \Phi_{mol}(m), \quad (1)$$

that is, a $(n_{tar} \times n_{mol})$-dimensional vector whose entries are products of the form $\Phi^i_{tar}(t) \times \Phi^j_{mol}(m)$, for $1 \le i \le n_{tar}$ and $1 \le j \le n_{mol}$. A SVM is then used as an inference engine, to estimate a linear function $f(t, m)$ in the vector space of target/molecule pairs, that takes positive values for interacting pairs and negative values for non-interacting ones.

The main motivation for using the tensor product (1) is that it provides a systematic way to encode correlations between small molecule and target features. For example, in the case of binary descriptors, the product of two features is 1 if both the molecule and the target descriptors are 1, and zero otherwise, which amounts to encode the simultaneous presence of particular features of the molecule and of the target that may be important for the formation of a complex. A potential issue with this approach, however, is that the size of the vector representation $n_{tar} \times n_{mol}$ for a pair may be prohibitively large for practical computation and manipulation. For example, using a vector of molecular descriptors of size 1024 for molecules, and representing a protein by the vector of counts of all 2-mers of amino-acids in its sequence ($d_t = 20 \times 20 = 400$) results in more than 400 k dimensions for the representation of

a pair. As pointed out by [27], this computational obstacle can however be overcome when a SVM is used to train the linear classifier, thanks to a trick often referred to as the *kernel trick*. Indeed, a SVM does not necessarily need the explicit computation of the vectors representing the complexes in the training set to train a model. What it needs, instead, is the inner products between these vectors, and a classical property of tensor products is that the inner product between two tensor products $\Phi_{pair}(t, m)$ and $\Phi_{pair}(t', m')$ is the product of the inner product between $\Phi_{tar}(t)$ and $\Phi_{tar}(t')$, on the one hand, and the inner product between $\Phi_{mol}(m)$ and $\Phi_{mol}(m')$, on the other hand. More formally, this property can be written as follows:

$$\begin{aligned}&\left(\Phi_{tar}(t)\otimes\Phi_{mol}(m)\right)^{\top}\left(\Phi_{tar}(t')\otimes\Phi_{mol}(m')\right)\\&=\Phi_{tar}(t)^{\top}\Phi_{tar}(t')\times\Phi_{mol}(m)^{\top}\Phi_{mol}(m'),\end{aligned}\tag{2}$$

where $u \prec v = u_1v_1 + ... + u_dv_d$ denotes the inner product between two d-dimensional vectors u and v. In other words, the SVM does not need to compute the $n_{tar} \times n_{mol}$ vectors to describe each pair, it only computes the respective inner products in the target and ligand spaces, before taking the product of both numbers.

This flexibility to manipulate molecule and target descriptors separately can moreover be combined with other tricks that sometimes allow to compute efficiently the inner products in the target and ligand spaces, respectively. Many such inner products, also called *kernels*, have been developed recently both in computational biology [37] and chemistry [38-40], and can be easily combined within the chemogenomics framework as follows: if two kernels for molecules and targets are given as:

$$\begin{aligned}K_{mol}(m,m')&=\Phi_{mol}(m)^{\top}\Phi_{mol}(m'),\\K_{tar}(t,t')&=\Phi_{tar}(t)^{\top}\Phi_{tar}(t'),\end{aligned}\tag{3}$$

then we obtain the inner product between tensor products, *i.e.*, the kernel between pairs, by:

$$K((t, m), (t', m')) = K_{tar}(t, t') \times K_{mol}(m, m'). \tag{4}$$

In summary, as soon as two vectors of descriptors or kernels K_{mol} and K_{tar} are chosen, we can solve the *in silico* chemogenomics problem with an SVM using the product kernel (4) between pairs. The particular descriptors or kernels used should ideally encode properties related to the ability of similar molecules to bind similar targets or ligands respectively.

In the next two subsections, we present different possible choices of descriptors – or kernels – for small molecules and GPCRs, respectively.

Descriptors for small molecules

The problem of explicitly representing and storing small molecules as finite-dimensional vectors has a long history in chemoinformatics, and a multitude of molecular descriptors have been proposed [41]. These descriptors include in particular physicochemical properties of the molecules, such as its solubility or logP, descriptors derived from the 2D structure of the molecule, such as fragment counts or structural fingerprints, or descriptors extracted from the 3D structure [42]. Each classical fingerprint vector and vector representation of molecules define an explicit "chemical space" in which each molecule is represented by a finite-dimensional vector, and these vector representations can obviously be used as such to define kernels between molecules [43]. Alternatively, some authors have recently proposed some kernels that generalize some of these sets of descriptors and correspond to inner products between large- or even infinite-dimensional vectors of descriptors. These descriptors encode, for example, the counts of an infinite number of walks on the graph describing the 2D structure of the molecules [39,40,44], or various features extracted from the 3D structures [43,45].

In this study we select two existing kernels, encoding respectively 2D and 3D structural information of the small molecules:

• *The 2D Tanimoto kernel*. Our first set of descriptors is meant to characterize the 2D structure of the molecules. For a small molecule m, we define the vector $\Phi_{mol}(m)$ as the binary vector whose bits indicate the presence or absence of all linear graph of length u or less as subgraphs of the 2D structure of l. We chose $u = 8$ in our experiment, *i.e.*, characterize the molecules by the occurrences of linear subgraphs of length 8 or less, a value previously observed to give good results in several virtual screening tasks [40]. Moreover, instead of directly taking the inner product between vectors as in (3), we use the Tanimoto kernel:

$$K_{ligand}(l,l') = \frac{\Phi_{lig}(l)^{\top}\Phi_{lig}(l')}{\Phi_{lig}(l)^2+\Phi_{lig}(l')^2-\Phi_{lig}(l)^{\top}\Phi_{lig}(l')},\tag{5}$$

which was proven to be a valid inner product by [46], giving very competitive results on a variety of QSAR or toxicity prediction experiments.

• *3D pharmacophore kernel* While 2D structures are known to be very competitive in ligand-based virtual screening

for identification of molecules presenting some given chemical, physical or biological properties [43], we reasoned that the protein-ligand recognition process takes place in the 3D space. Thus, we decided to test descriptors representing the presence of potential 3-point pharmacophores. For this, we used the 3D pharmacophore kernel proposed by [45], that generalizes 3D pharmacophore fingerprint descriptors. This approach requires the choice of a 3D conformer for each molecule, in a context where there exists a large number of methods for exploring the conformation space, and where we lack significant data for bound ligands in GPCR structures. Therefore, we chose to build a 3D version of the ligand database in which molecules are represented in the conformation proposed by the Omega program (OpenEye Scientific Software), because it performs rapid systematic conformer search, and has been showed to present good performances for retrieving bioactive conformations [47]. For each of the 2446 retained ligands, the conformer was generated using the standard Omega parameters, except for a 1 Å RMSD clustering of the conformers, instead of the 0.8 default value. Partial charges were calculated for all atoms using the molcharge program (OpenEye Scientific Software) with standard parameters. This ligand database was then used to calculate a 3D pharmacophore kernel for molecules [45].

We used the freely and publicly available *ChemCPP* (available at http://chemcpp.sourceforge.net) software to compute the 2D and 3D pharmacophore kernel.

Descriptors for GPCRs

SVM and kernel methods are also widely used in bioinformatics [37], and a variety of approaches have been proposed to design kernels between proteins, ranging from kernels based on the amino-acid sequence of a protein [48-54] to kernels based on the 3D structures of proteins [55-57] or on the pattern of occurrences of proteins in multiple sequenced genomes [58]. These kernels have been used in conjunction with SVM or other kernel methods for various tasks related to structural or functional classification of proteins. While any of these kernels can theoretically be used as a GPCR kernel in (4), we investigate in this paper a restricted list of specific kernels described below, aimed at illustrating the flexibility of our framework and test various hypothesis.

• The *Dirac* kernel between two targets t, t' is:

$$K_{Dirac}(t,t') = \begin{cases} 1 & \text{if } t = t', \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

This basic kernel simply represents different targets as orthonormal vectors. From (4) we see that orthogonality between two proteins t and t' implies orthogonality between all pairs (l, t) and (l', t') for any two small molecules c and c'. This means that a linear classifier for pairs (l, t) with this kernel decomposes as a set of independent linear classifiers for interactions between molecules and each target protein, which are trained without sharing any information of known ligands between different targets. In other words, using Dirac kernel for proteins amounts to performing classical learning independently for each target, which is our baseline approach.

• The *multitask* kernel between two targets t, t' is defined as:

$$K_{multitask}(t, t') = 1 + K_{Dirac}(t, t').$$

This kernel, originally proposed in the context of multitask learning [59], removes the orthogonality of different proteins to allow sharing of information. As explained in [59], plugging $K_{multitask}$ in (4) amounts to decomposing the linear function used to predict interactions as a sum of a linear function common to all GPCRs and of a linear function specific to each GPCR:

$$f(l,t) = w^{\top}\Phi(l,t) = w_{general}^{\top}\Phi_{lig}(l) + w_t^{\top}\Phi_{lig}(l).$$

A consequence is that only data related to the the target t are used to estimate the specific vector w_t, while all data are used to estimate the common vector $w_{general}$. In our framework this classifier is therefore the combination of a target-specific part accounting for target-specific properties of the ligands and a global part accounting for general properties of the ligands across the targets. The latter term allows to share information during the learning process, while the former ensures that specificities of the ligands for each target are not lost.

• The *hierarchy* kernel. Alternatively we could propose a new kernel aimed at encoding the similarity of proteins with respect to the ligands they bind. In the GLIDA database indeed, GPCRs are grouped into 4 classes based on sequence homology and functional similarity: the *rhodopsin* family (class A), the *secretin* family (class B), the *metabotropic* family (class C) and some smaller classes containing other GPCRs. The GLIDA database further subdivides each class of targets by type of ligands, for example amine or peptide receptors or more specific families of ligands. This also defines a natural hierarchy that can be used to compare GPCRs.

The hierarchy kernel between two GPCRs was therefore defined as the number of common ancestors in the corresponding hierarchy plus one, that is,

$$K_{hierarchy}(t, t') = \langle \Phi_h(t), \Phi_h(t') \rangle,$$

where $\Phi_h(t)$ contains as many features as there are nodes in the hierarchy, each being set to 1 if the corresponding node is part of t's hierarchy and 0 otherwise, plus one feature constantly set to one that accounts for the "plus one" term of the kernel.

• The *binding pocket* kernel. Because the protein-ligand recognition process occurs in 3D space in a pocket involving a limited number of residues, we tried to describe the GPCR space using a representation of this pocket. The difficulty resides in the fact that although the GPCR sequences are known, the residues forming this pocket are *a priori* unknown. However, mutagenesis data showed that the transmembrane binding site is situated in a similar region for all GPCRs [60], and this information was confirmed by the two available X-ray structures. In order to identify residues potentially involved in the binding pocket of GPCRs of unknown structure studied in this work, we proceeded in several steps, somewhat similarly to [61]. (a) The two known structures, PDB entries 1U19 and 2RH1[62,63], were superimposed using the STAMP algorithm [64]. Although retinal is an inverse agonist and form a covalent bond with Rhodopsin, while carazolol is an agonist and binds non-covalently, root mean square deviation between these two complexed structures is only of 1.6 Å in the transmembrane helices [65]. In the superimposed structures, the retinal and 3-(isopropylamino)propan-2-ol ligands are localized in the same region of the transmembrane space, which is in agreement with global conservation of binding pockets, as shown on Figure 1. (b) The structural alignment of bovine rhodopsin and of human β_2-adrenergic receptor was used to generate a sequence alignment of these two proteins. (c) For both structures, in order to identify residues potentially involved in stabilizing interactions with the ligand (residues of the pocket), we selected residues that presented at least one atom situated at less than 6 Å from at least one atom of the ligand. Figure 2 shows that these two

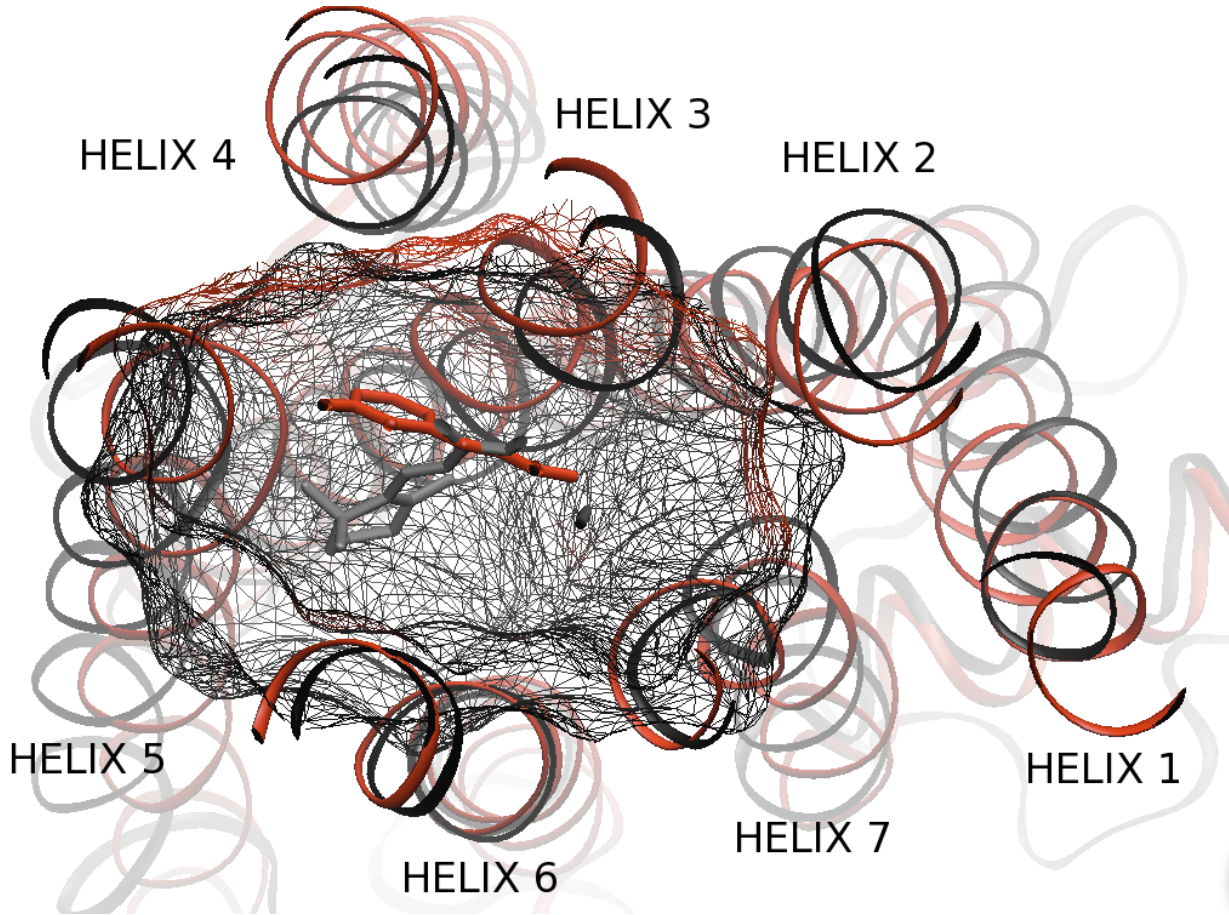

Figure 1
Binding pocket. Representation of the binding pocket of β_2-adrenergic receptor (in red) and bovine Rhodopsin (in black) viewed from the extracellular surface. On the center of the pocket, 3-(isopropylamino)propan-2-ol and cis-retinal have been represented to show the size and the position of the pocket around each ligand. Figure drawn with VMD [79].

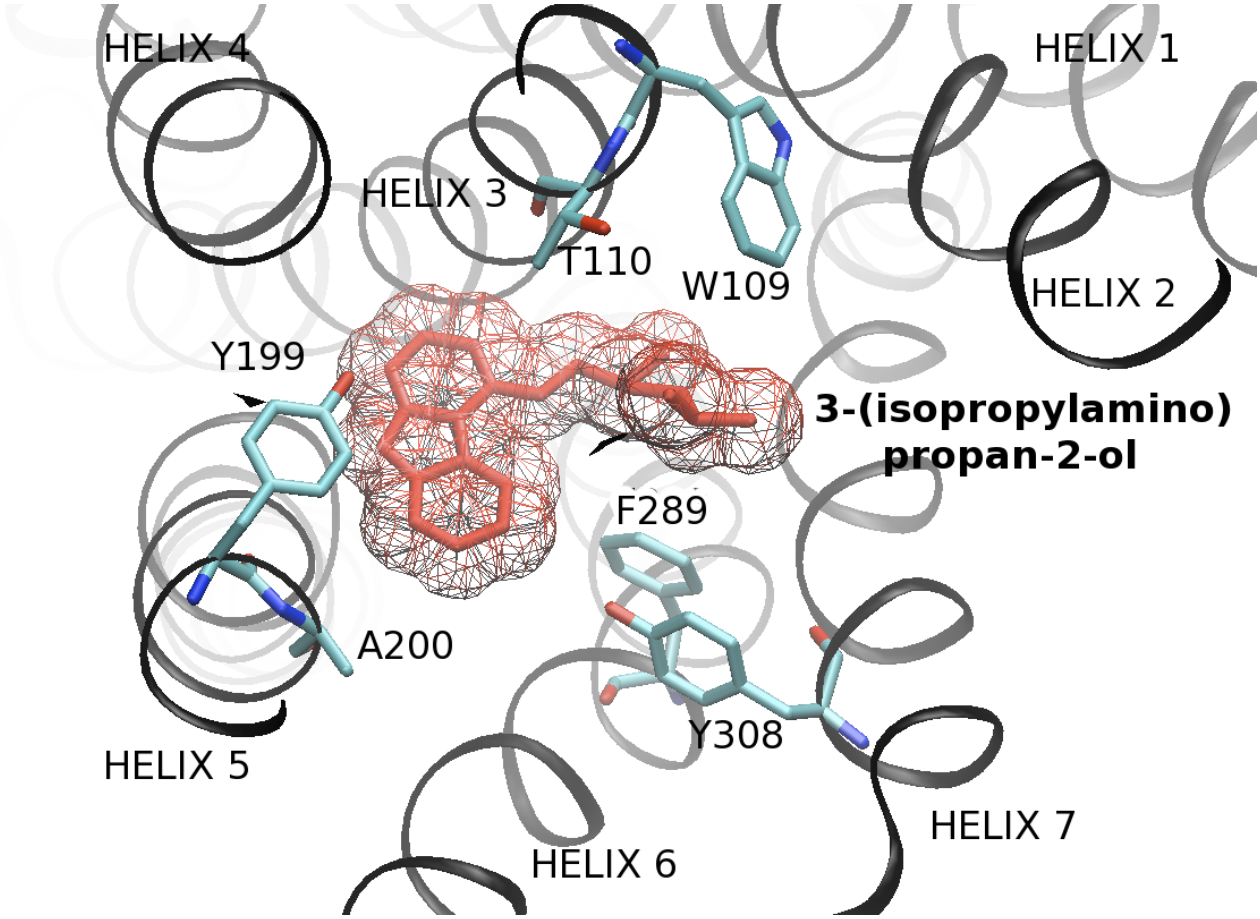

Figure 2
3-(isopropylamino)propan-2-ol and the protein environment of β_2-adrenergic receptor as viewed from the extracellular surface. 3-(isopropylamino)propan-2-ol and the protein environment of β_2-adrenergic receptor as viewed from the extracellular surface. Amino acid side chains are represented for 6 of the 31 residues (in cyan, blue and red) of the binding pocket motif. Transmembrane helix and 3-(isopropylamino)propan-2-ol are colored in black and red respectively. Figure drawn with VMD [79].

pockets clearly overlap, as expected. (d) Residues of the two pockets (as defined in (c)) were labeled in this structural sequence alignment. These residues were found to form small sequence clusters that were in correspondence in this alignment. These clusters were situated mainly in the apical region of transmembrane segments and included a few extracellular residues. Indeed, it has been previously demonstrated that extracellular loops can play a role in ligand binding together with transmembrane regions [66]. (e) All studied GPCR sequences, including bovine rhodopsin and human β_2-adrenergic receptor were aligned using CLUSTALW [67] with Blosum matrices [68]. Sequences which could not be correctly aligned (i.e. with important gaps in the transmembrane regions) were discarded in order to only keep comparable sequences. We then checked that conserved residues according to [69] of the transmembrane helices were correctly aligned, and local misalignments were corrected. In addition, the structural alignment of bovine rhodopsin and human β_2-adrenergic receptor, and known conserved positions were used to locally correct misalignments. For each protein, residues in correspondence in this alignment with a residue of the binding pocket (as defined above) of either bovine rhodopsin or human β_2-adrenergic receptor were retained. This lead to a different number of residues per protein, because of sequence variability. For example, in extracellular regions, some residues from bovine rhodopsin or human β_2-adrenergic receptor had a corresponding residue in some sequences but not in others. In order to provide a homogeneous description of the binding pocket for all GPCRs, in the list of residues initially retained for each protein, only residues situated at positions where no gaps were found in any of the GPCRs were kept. (f) Each protein was then represented by a vector

whose elements corresponded to a potentially conserved pocket. This description, although appearing as a linear vector filled with amino acid residues [see Additional file 1], implicitly codes for a 3D information on the receptor pocket, as illustrated in Figure 2. These vectors were then used to build a kernel that allows comparison of binding pockets. The classical way to represent motifs of constant length as fixed length vectors is to encode the letter at each position by a 20-dimensional binary vector indicating which amino acid is present, resulting in a 180-dimensional vector representations. In terms of kernel, the inner product between two binding pocket motifs in this representation is simply the number of letters they have in common at the same positions:

$$K_{pb}(x, x') = \sum_{i=1}^{l} \delta(x[i], x'[i]),$$

where l is the length of the binding pocket motifs (31 in our case), $x[i]$ is the i-th residue in x and $\delta(x[i], x'[i])$ is 1 if $x[i] = x'[i]$, 0 otherwise. This is the baseline pocket binding kernel. Alternatively, using a polynomial kernel of degree p over the baseline kernel is equivalent, in terms of feature space, to encoding p-order interactions between amino acids at different positions. In order to assess the relevance of such non-linear extensions we tested this polynomial pocket binding kernel,

$$K_{ppb}(x, x') = (K_{pb}(x, x') + 1)^p.$$

We only used a degree $p = 2$, although a more careful choice of this parameter could further improve the performances.

Results

We ran two different sets of experiments on this dataset in order to illustrate two important points. In a first set of experiments, for each GPCR, we 5-folded the data available, *i.e.*, the line of the interaction matrix corresponding to this GPCR. The classifier was trained with four folds and the whole data from the other GPCRs, *i.e.*, all other lines of the interaction matrix. The prediction accuracy for the GPCR under study was then tested on the remaining fold. The goal of these first experiments was to evaluate if using data from other GPCRs improved the prediction accuracy for a given GPCR. In a second set of experiments, for each GPCR we ignored ligand data available for this particular GPCR, we trained a classifier on the whole data from the other GPCRs, and tested on the data of the considered GPCR. The goal was to assess how efficient our chemogenomics approach would be to predict the ligands of orphan GPCRs. In both experiments, the C parameter of the SVM was selected by internal cross validation on the training set among 2^i, $i \in \{-8, -7, ..., 5, 6\}$. The data and source code (under GPL license) are publicly available [see Additional file 2].

For the first experiment, since learning an SVM with only one training point does not really make sense and can lead to "anti-learning" less than 0.5 performances, we set all results r involving the Dirac GPCR kernel on GPCRs with only 1 known ligand to $\max(r, 0.5)$. This is to avoid any artefactual penalization of the Dirac approach and make sure that we measure the actual improvement brought by sharing information across GPCRs.

Table 1 shows the results of the first experiments with all the ligand and GPCR kernel combinations. For all the ligand kernels, one observes an improvement between the individual approach (Dirac GPCR kernel, 86.2%) and the baseline multitask approach (multitask GPCR kernel, 88.8%). The latter kernel is merely modeling the fact that each GPCR is uniformly similar to all other GPCRs, and twice more similar to itself. It does not use any prior information on the GPCRs, and yet, using it improves the global performance with respect to individual learning. Using more informative GPCR kernels further improves the prediction accuracy. In particular, the hierarchy kernel add more than 4.5% of precision with respect to naive multitask approach. All the other informative GPCR kernels also improve the performance. The polynomial binding pocket kernel is almost as efficient as the hierarchy kernel, which is an interesting result. Indeed, one could fear that using the hierarchy kernel, for the construction of which some knowledge of the ligands may have been used, could have introduced bias in the results. Such bias is certainly absent in the binding pocket kernel. The fact that the same performance can be reached with kernels based on the mere sequence of GPCRs' pockets is therefore an important result. Figure 3 shows three of the GPCR kernels. The baseline multitask is shown as a comparison. Interestingly, many of the subgroups defined in the hierarchy can be found in the binding pocket kernel, that is, they are retrieved from the simple information of the binding pocket sequence.

The 3D kernel for the ligands, on the other hand, did not perform as well as the 2D kernel. This can be either explained by the fact the the pharmacophore kernel is not

Table 1: Prediction accuracy for the first experiment with various ligand and target kernels

$K_{tar} \backslash K_{lig}$	2D Tanimoto	3D pharmacophore
Dirac	86.2 ± 1.9	84.4 ± 2.0
multitask	88.8 ± 1.9	85.0 ± 2.3
hierarchy	93.1 ± 1.3	88.5 ± 2.0
binding pocket	90.3 ± 1.9	87.1 ± 2.3
poly binding pocket	92.1 ± 1.5	87.4 ± 2.2

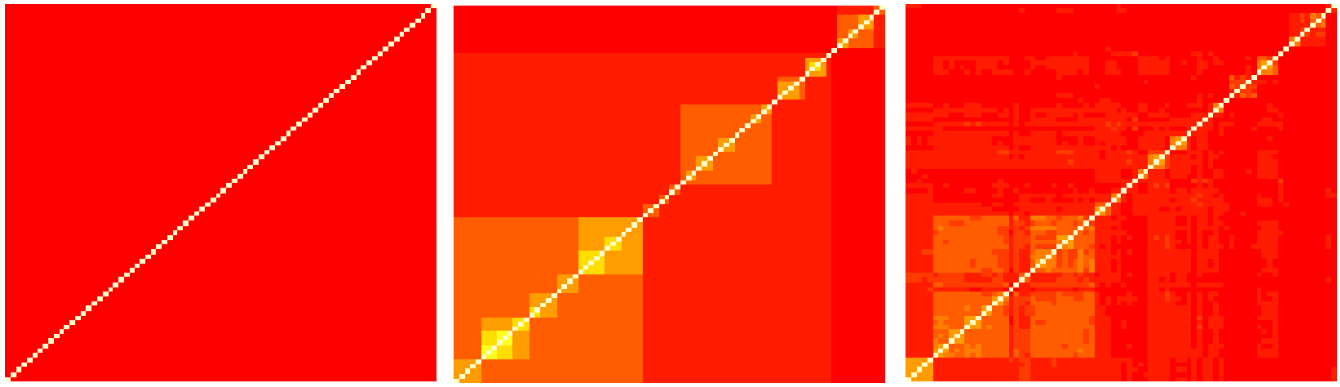

Figure 3
GPCR kernel Gram matrices. GPCR kernel Gram matrices (K_{tar}) for the GLIDA GPCR data with multitask, hierarchy and binding pocket kernels.

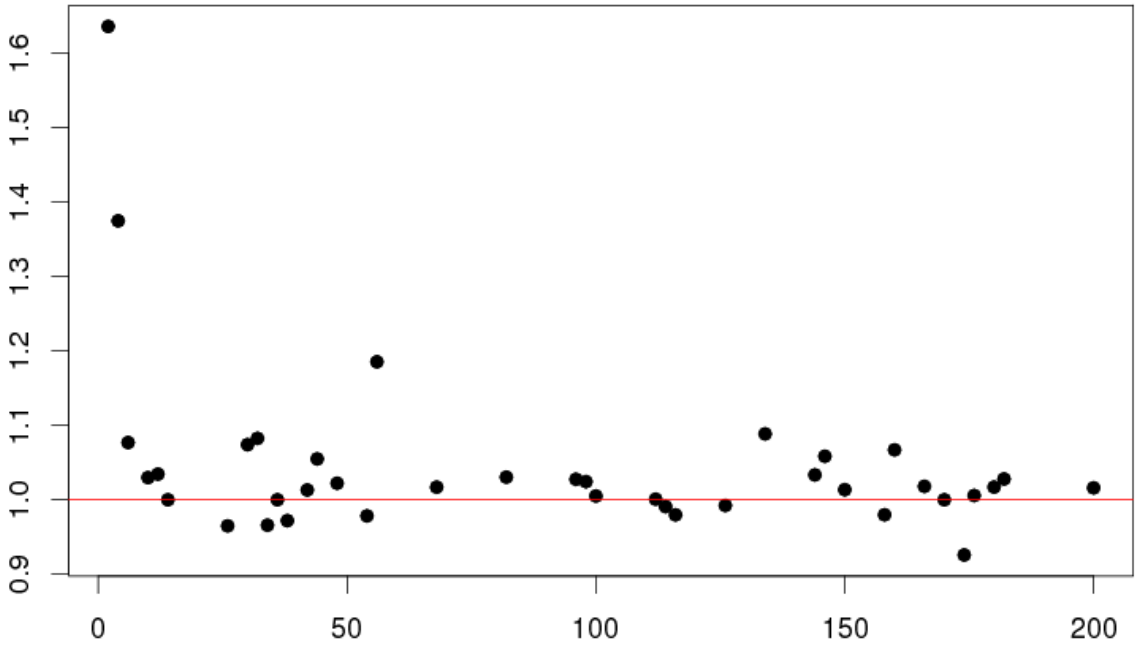

Figure 4
Improvement of the chemogenomics approach. Improvement (as a performance ratio) of the hierarchy GPCR kernel against the Dirac GPCR kernel as a function of the number of training samples available. Restricted to [2 – 200] samples for the sake of readability.

suited to this problem, or by the fact that choosing the conformer of the ligand is not a trivial task. This point is discussed below.

Figure 4 illustrates how the improvement brought by the chemogenomics approach varies with the number of available training points. As one could have expected, the strongest improvement is observed for the GPCRs with few (less than 20) training points (*i.e.*, less than 10 known ligands since for each known ligand an artificial non-ligand was generated). When more training points become available, the improvement is less important, and sharing the information across the GPCRs can even degrade the performances. This is an important point, first because, as showed on Figure 5, many GPCRs have few known ligands (in particular, 11 of them have only two training points), and second because it shows that when enough training points are available, individual learning will probably perform as well as or better than our chemogenomics approach.

Our second experiment intends to assess how our chemogenomics approach can perform when predicting ligands for orphan GPCRs, *i.e.*, with no training data available for the GPCR of interest. Table 2 shows that in this setting, individual learning performs random prediction. Naive multitask approach provides modest improvement of the performance, but informative kernels such as hierarchical and binding pocket kernels achieve 77.4% and 78.1% of precision respectively, that is, almost 30% better than the random approach one would get when no data is available. Here again, the fact that the binding pocket kernel that only uses the sequence of the receptor pocket performs as well as the hierarchy-based kernel is encouraging. It suggests that given a receptor for which nothing is known except its sequence, it is possible to make reasonable ligand predictions.

Discussion

Our results demonstrate that chemogenomic approaches outperform individual approach, in particular in cases where very limited or no ligand information is available, as shown in Table 2 and Figure 4. In the case of well stud-

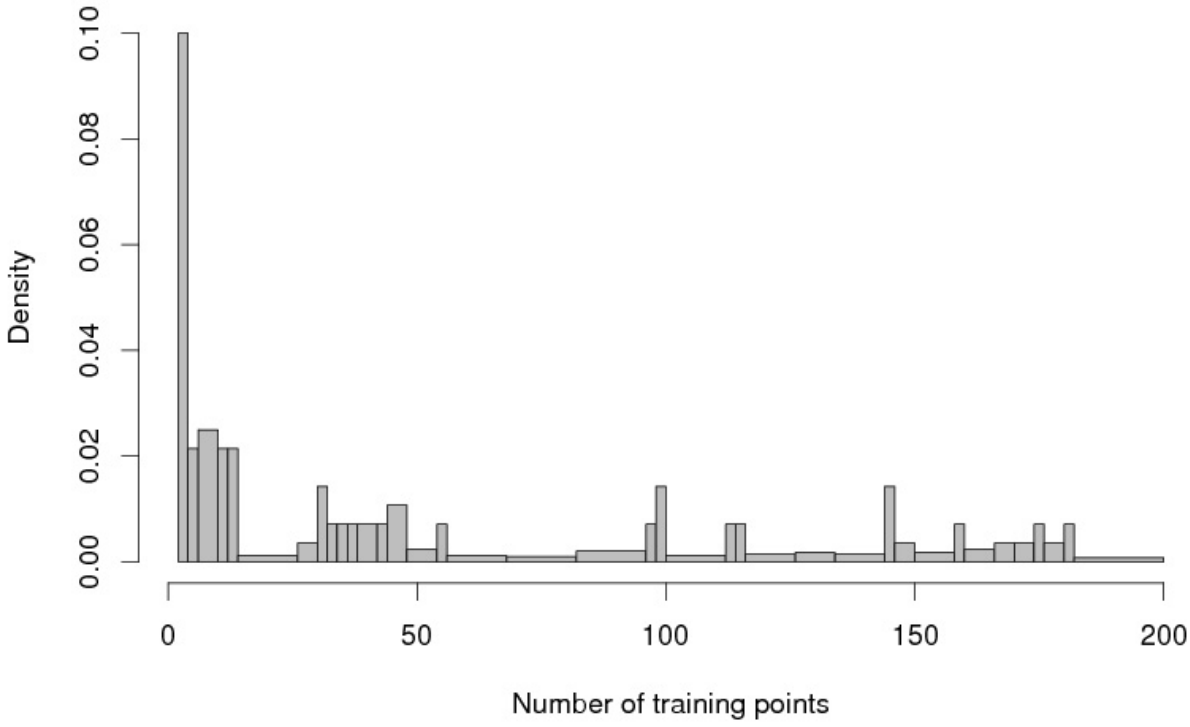

Figure 5
Distribution of the number of training points for a GPCR. Distribution of the number of training points for a GPCR. Restricted to [2 – 200] samples for the sake of readability.

Table 2: Prediction accuracy for the second experiment with various ligand and target kernels

K_{tar}\K_{lig}	2D Tanimoto	3D pharmacophore
Dirac	50.0 ± 0.0	50.0 ± 0.0
multitask	56.8 ± 2.5	58.2 ± 2.2
hierarchy	77.4 ± 2.4	76.2 ± 2.2
binding pocket	78.1 ± 2.3	76.6 ± 2.2
poly binding pocket	76.4 ± 2.4	74.9 ± 2.3

ied GPCRs, more classical ligand-based methods (QSAR) may be better suited to predict new strong binders from a large number of known ligands, as shown in Figure 4. Consistent with this observation, Tables 3 and 4 show that in the two types of experiments, the improvement is observed for all subfamilies of GPCRs retained in this study. This is an interesting result since most of published virtual screening studies on GPCRs were applied to class A GPCRs.

Since our chemogenomic approach is a ligand-based approach, it would probably be interesting to use it in combination with docking. Indeed, although prior known ligands can help tuning docking procedures to the receptor under study, it can in principle be used with little or no ligand information. When more experimental 3D structures become available for GPCRs in the future, this will help building reliable models for a wider range of GPCRs that would be suitable for docking studies. Joint use of ligand-based chemogenomic and docking would certainly improve predictions.

We chose to use a binary descriptor for the receptor-ligand interaction, while QSAR or docking methods usually try to rank molecules according to their predicted affinity for the receptor. However, affinity prediction is still a subject of research at the level of a single receptor, at least when using methods whose calculation times are compatible with the screening of large molecular databanks. In this context, we feel that in chemogenomic approaches, where information is shared between different proteins, such quantitative prediction is even more challenging. This led us to retain the binary binding and non-binding descriptors, although it would formally have been straightforward to use a regression algorithm instead of a classification one to make quantitative predictions.

It is not always easy to compare the performances of a new method to other existing methods, and particularly in the case of GPCRs. Indeed, at least to our knowledge, there is up to now no public complete data from previous screening studies available as a benchmark to compare different screening methods on the same data. This urged us to give public access to the ligand and receptor databases used in this study, to the detailed experimental protocol of the study, and to the predictions made by our chemogenomic approach for each GPCR [see Additional files 3, 4] (summarized by GPCR family in Table 3 and Table 4). This provides a benchmark which we hope will contribute to a fair evaluation of different methods and trigger new developments. This benchmark could be used to compare predictions made by other methods. Our approach boils down to the application of well-known machine learning methods in the constructed chemogenomics space. We used a systematic way to build such a space by combining a given representation of the ligands with a given representation of the GPCRs into a binding-prediction-oriented GPCR-ligand couple representation. This allows to use any ligand or GPCR descriptor or kernel existing in the chemoinformatics or bioinformatics literature, or new ones containing other prior information as we tried to propose in this paper. Our experiments showed that the choice of the descriptors was crucial for the prediction, and more sophisticated features for either the ligands or the GPCRs could probably further improve the performances. Among these features, improvements in the 3D ligand descriptors could probably be obtained. Indeed, 3D pharmacophore kernels did not always reach the performance of 2D kernels for the ligands. This is apparently in contradiction with the idea that protein-ligand interaction is a process occurring in the 3D space, and with previous work in our group [45]. Different explanations can be proposed. First, it is possible that the bioactive conformation was not correctly predicted for all molecules used in this study. For the two ligands for which it was known, *i.e.*, retinal and 3-(isopropylamino)propan-2-ol from PDB

Table 3: Prediction accuracy by GPCR family for the first experiment

Family\K_{tar}	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors (18)	73.7	80.0	85.8	83.8	83.7
Rhodopsin amine receptors (35)	91.1	92.1	94.0	93.9	94.1
Rhodospin other receptors (17)	83.6	88.0	95.7	95.9	95.9
Metabotropic glutamate family (9)	73.1	93.5	98.9	83.3	93.3
Secretin family (1)	50.0	100.0	100.0	50.0	100.0

Mean prediction accuracy for each GPCR family for the first experiment with the 2D Tanimoto ligand kernel and various target kernels. The numbers in bracket are the numbers of receptors considered in the experiment for each family. BP is the binding pocket kernel and PBP the poly binding pocket kernel.

Table 4: Prediction accuracy by GPCR family for the second experiment

Family\K_{tar}	Dirac	multitask	hierarchy	BP	PBP
Rhodopsin peptide receptors (18)	50.0	50.6	66.7	74.0	65.3
Rhodopsin amine receptors (35)	50.0	56.0	73.7	74.0	73.1
Rhodospin other receptors (17)	50.0	50.2	86.5	87.6	85.5
Metabotropic glutamate family (9)	50.0	79.7	93.9	87.2	91.3
Secretin family (1)	50.0	100.0	100.0	50.0	100.0

Mean prediction accuracy for each GPCR family for the second experiment with the 2D Tanimoto ligand kernel and various target kernels. The numbers in bracket are the numbers of receptors considered in the experiment for each family. BP is the binding pocket kernel and PBP the poly binding pocket kernel.

entries 1U19 and 2RH1 respectively, we found that the predicted conformation, using the same method as for all other molecules, was very close to the experimental conformation, with RMSD values of less than 1 Å. However, in absence of any other information on bound ligand conformations, it is not possible to rule out the possibility that for other molecules, the prediction was not correct. Although more complete conformational space exploration for all ligands was clearly out of the scope of this paper and would be a study by itself, work in this direction could improve the method. In particular, since 2D ligand-based methods are not easily suitable to make predictions outside of the molecular scaffolds for which information is known, ligand-based methods using 3D description are of particular interest, because they are expected to allow better predictions on molecules presenting diverse molecular patterns. Synergy between our method and docking would provide a means for the choice of a conformer. The principle could be to build homology models for the GPCRs, dock the molecular database in the modeled binding pockets, and derive a 3D database using, for each molecule, the conformer associated to the best docking solution. However, conformer generation and selection is a major drawback of using 3D descriptors, especially in the case of large ligands with many free torsion angles.

Various evidence suggest that, within a common global architecture, a generic binding pocket mainly involving transmembrane regions hosts agonists, antagonists and allosteric modulators. In order to identify this pocket automatically, other studies report the use of sequence alignment and the prediction of transmembrane helices. [60] detected hypervariable positions in transmembrane helices for identification of residues forming the binding pocket, although some positions were more conserved. Indeed, conserved residues are probably important for structural stabilization of the pocket, while variable positions are involved in ligand binding, in order to accommodate the wide spectrum of molecules that are GPCR substrates. Analyzing the positions of variable positions, these authors proposed potential binding pockets for GPCRs, and found that the corresponding residues were frequently in the GRAP mutant database for GPCRs [70]. Interestingly, they pointed that residues at hypervariable positions were found within a distance of 6 Å from retinal in the rhodopsin X-Ray structure, which is also a classical distance cutoff above which it is admitted that protein-ligand interactions become negligible. Therefore, this inspired the simple and automatic method used in the present work for extracting GPCRs potential binding pockets, and our results are in good agreement with this study. It is also important to note that GPCRs are known to exist in dynamic equilibrium between inactive- and several active-state conformations [71], and different ligands sometimes trigger distinct conformational changes and stabilize different receptor conformations [72]. Taking into account receptor plasticity constitutes in itself a research domain in docking. Its use is of particular interest for screening GPCR homology models since residue positions are not exactly known. Therefore flexible docking procedures have been proposed and applied on GPCR proteins [9,73]. Moreover, a modeling method has been proposed to get insights on transmembrane bundle plasticity [74]. In our case, receptor flexibility might influence the definition of the binding pocket, since it initially relies on the identification of residues in the two reference structures (1U19 and 2RH1) that present at least one atom situated at less than 6 Å of the ligand. Therefore, we made the implicit hypothesis that receptor conformational changes upon ligand binding does not drastically affect this list of residues. When more structures become available in this family of proteins, a better appreciation of such conformational rearrangements will be possible, which could be taken into account in the binding pocket definition and could help to improve the method. [70] found that hierarchical tree representations of GPCR subfamilies calculated with full-length GPCR sequences or with only binding pocket residues were similar, and that locally, the latter was in better agreement with functional data although their binding pocket included only 35 residues. This result is also in good agreement with our finding that the hierarchy kernel based on full length sequence (from GLIDA) and the kernel based on the binding pocket provided very similar performances. As mentioned in the Results section, it is however important to note that the

kernels based on the binding pocket were built without any ligand information that could lead to some bias and artificially better performance.

Conclusion

We showed how sharing information across the GPCRs by considering a chemogenomics space of the GPCR-ligand interaction pairs could improve the prediction performances, with respect to the single receptor approach. In addition, we showed that using such a representation, it was possible to make reasonable predictions even when all known ligands were ignored for a given GPCR, that is, to predict ligands for orphan GPCRs. Our results demonstrate that chemogenomic approaches is particularly suited to cases where very limited or no ligand information is available, as shown in Table 2.

This chemogenomics approach is related to ligand-based approaches. However, sharing information among different GPCRs allows, in this case, to perform prediction on orphan GPCRs, which is also possible using target-based methods. Nevertheless, the latter are limited by the number of known receptor structures and the difficulty to apply such methods on homology models.

Interesting developments of this method could include application to other important drug target families, like enzymes or ion channels [75], for which most of the descriptors used for the GPCRs in this paper could directly be transposed, and other, more specific ones could be designed. From a methodological point of view, many recent developments in multitask learning [76-78] could be applied to generalize this chemogenomics approach using, for example, other regularization methods.

Authors' contributions

LJ, with the help and under the supervision of JPV, developed and implemented all the classification methods presented in the paper, and ran the experiments. BH and VS designed the benchmark and developed the binding pocket kernel and 3D ligand kernel. All authors contributed to writing the text, read and approved the final manuscript.

Additional material

Additional file 1

Aligned receptor pocket residues. Residues of 5-hydroxytryptamine 5A receptor, Adenosine A2b receptor, Gamma-aminobutyric acid type B receptor and Relaxin 3 receptor 2 (shown as examples) aligned with β_2-adrenergic receptor binding site amino acids. The binding pocket motif of β_2-adrenergic receptor has been used as reference to determine residues involved in the formation of the binding site of the 79 other GPCRs. Bold columns correspond to the residues shown on Figure 2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-363-S1.pdf]

Additional file 2

Source and data. Source code (under GPL license) and benchmark used in the experiments in a compressed archive checker.tgz.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-363-S2.tgz]

Additional file 3

Prediction accuracy by GPCR for the first experiment. Mean prediction accuracy for each GPCR for the first experiment with the 2D Tanimoto ligand kernel and various target kernels. The GPCR identifiers are the GLIDA references. The numbers in bracket are the numbers ligands considered in the experiment for each GPCR. BP is the binding pocket kernel and PBP the poly binding pocket kernel.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-363-S3.pdf]

Additional file 4

Prediction accuracy by GPCR for the second experiment. Mean prediction accuracy for each GPCR for the second experiment with the 2D Tanimoto ligand kernel and various target kernels. The GPCR identifiers are the GLIDA references. The numbers in bracket are the numbers ligands considered in the experiment for each GPCR. BP is the binding pocket kernel and PBP the poly binding pocket kernel.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-363-S4.pdf]

Acknowledgements

We thank Mines ParisTech and Carnot Mines for financial support to the project and funding of LJ and BH.

References

1. Bockaert J, Pin JP: **Molecular tinkering of G protein-coupled receptors: an evolutionary success.** *EMBO J* 1999, **18(7):**1723-1729.
2. Deshpande DA, Penn RB: **Targeting G protein-coupled receptor signaling in asthma.** *Cell Signal* 2006, **18(12):**2105-2120.
3. Hill SJ: **G-protein-coupled receptors: past, present and future.** *Br J Pharmacol* 2006, **147(Suppl 1):**S27-S37.
4. Catapano LA, Manji HK: **G protein-coupled receptors in major psychiatric disorders.** *Biochim Biophys Acta* 2007, **1768(4):**976-993.
5. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G: **GPCRDB information system for G protein-coupled receptors.** *Nucl Acids Res* 2003, **31:**294-297.

6. Fredholm BB, Hökfelt T, Milligan G: **G-protein-coupled receptors: an update.** *Acta Physiol* 2007, **190:**3-7.
7. Lin SHS, Civelli O: **Orphan G protein-coupled receptors: targets for new therapeutic interventions.** *Ann Med* 2004, **36(3):**204-214.
8. Evers A, Klabunde T: **Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor.** *J Med Chem* 2005, **48(4):**1088-1097.
9. Cavasotto CN, Orry AJW, Abagyan RA: **Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors.** *Proteins* 2003, **51(3):**423-433.
10. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, Avisar N, Inbal B, Heifetz A, Fichman M, Topf M, Naor Z, Noiman S, Becker OM: **PREDICT modeling and in-silico screening for G-protein coupled receptors.** *Proteins* 2004, **57:**51-86.
11. 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?** *Proteins* 2003, **50:**5-25.
12. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, Bar-Haim S, Warshaviak D, Fichman M, Noiman S: **G protein-coupled receptors: in silico drug discovery in 3D.** *Proc Natl Acad Sci USA* 2004, **101(31):**11304-11309.
13. Cavasotto CN, Orry AJW, Murgolo NJ, Czarniecki MF, Kocsi SA, Hawes BE, O'Neill KA, Hine H, Burton MS, Voigt JH, Abagyan RA, Bayne ML, Monsma FJ: **Discovery of novel chemotypes to a G-protein-coupled receptor through ligand-steered homology modeling and structure-based virtual screening.** *J Med Chem* 2008, **51(3):**581-588.
14. Rolland C, Gozalbes R, Nicolaï A, Paugam MF, Coussy L, Barbosa F, Horvath D, Revah F: **G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation.** *J Med Chem* 2005, **48(21):**6563-6574.
15. Kubinyi H, Müller G, Mannhold R, Folkers G, (Eds): *Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective* Methods and Principles in Medicinal Chemistry, New York: Wiley-VCH; 2004.
16. Jaroch SE, Weinmann H, (Eds): *Chemical Genomics: Small Molecule Probes to Study Cellular Function* Ernst Schering Research Foundation Workshop, Berlin: Springer; 2006.
17. Klabunde T: **Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.** *Br J Pharmacol* 2007, **152:**5-7.
18. Rognan D: **Chemogenomic approaches to rational drug design.** *Br J Pharmacol* 2007, **152:**38-52.
19. Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP: **Property-based design of GPCR-targeted library.** *J Chem Inf Comput Sci* 2002, **42(6):**1332-1342.
20. Klabunde T: **Chemogenomics Approaches to Ligand Design.** In *Ligand Design for G Protein-coupled Receptors* Great Britain: Wiley-VCH; 2006:115-135.
21. Schuffenhauer A, Zimmermann J, Stoop R, Vyver JJ van der, Lecchini S, Jacoby E: **An ontology for pharmaceutical ligands and its application for in silico screening and library design.** *J Chem Inf Comput Sci* 2002, **42(4):**947-955.
22. Frimurer TM, Ulven T, Elling CE, Gerlach LO, Kostenis E, Högberg T: **A physicogenetic method to assign ligand-binding relationships between 7TM receptors.** *Bioorg Med Chem Lett* 2005, **15(16):**3707-3712.
23. 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(2):**391-405.
24. Bock JR, Gough DA: **Virtual screen for ligands of orphan G protein-coupled receptors.** *J Chem Inf Model* 2005, **45(5):**1402-1414.
25. Lapinsh M, Prusis P, Uhlén S, Wikberg JES: **Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions.** *Bioinformatics* 2005, **21(23):**4289-4296.
26. Freyhult E, Prusis P, Lapinsh M, Wikberg JES, Moulton V, Gustafsson MG: **Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling.** *BMC Bioinformatics* 2005, **6:**50.
27. Erhan D, L'heureux PJ, Yue SY, Bengio Y: **Collaborative filtering on a family of biological targets.** *J Chem Inf Model* 2006, **46(2):**626-635.
28. Jacob L, Vert JP: **Efficient peptide-MHC-I binding prediction for alleles with few known binders.** *Bioinformatics* 2008, **24(3):**358-366.
29. Jacob L, Vert JP: **Protein-ligand interaction prediction: an improved chemogenomics approach.** *Bioinformatics* 2008 [http://bioinformatics.oxfordjournals.org/cgi/reprint/btn409].
30. Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G: **GLIDA: GPCR-ligand database for chemical genomic drug discovery.** *Nucleic Acids Res* 2006:D673-D677.
31. Caldwell J, Gardner I, Swales N: **An introduction to drug disposition: the basic principles of absorption, distribution, metabolism, and excretion.** *Toxicol Pathol* 1995, **23(2):**102-114.
32. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 2001, **46(1–3):**3-26.
33. Egan WJ, Merz KM, Baldwin JJ: **Prediction of drug absorption using multivariate statistics.** *J Med Chem* 2000, **43(21):**3867-3877.
34. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD: **Molecular properties that influence the oral bioavailability of drug candidates.** *J Med Chem* 2002, **45(12):**2615-2623.
35. Martin YC: **A bioavailability score.** *J Med Chem* 2005, **48(9):**3164-3170.
36. Bock JR, Gough DA: **Predicting protein-protein interactions from primary structure.** *Bioinformatics* 2001, **17(5):**455-460.
37. Schölkopf B, Tsuda K, Vert JP: *Kernel Methods in Computational Biology* The MIT Press, Cambridge, Massachussetts: MIT Press; 2004.
38. Kashima H, Tsuda K, Inokuchi A: **Marginalized Kernels between Labeled Graphs.** In *Proceedings of the Twentieth International Conference on Machine Learning* Edited by: Faucett T, Mishra N. New York, NY, USA: AAAI Press; 2003:321-328.
39. Gärtner T, Flach P, Wrobel S: **On graph kernels: hardness results and efficient alternatives.** In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory and the Seventh Annual Workshop on Kernel Machines, Volume 2777 of Lecture Notes in Computer Science* Edited by: Schölkopf B, Warmuth M. Heidelberg: Springer; 2003:129-143.
40. Mahé P, Ueda N, Akutsu T, Perret JL, Vert JP: **Graph kernels for molecular structure-activity relationship analysis with support vector machines.** *J Chem Inf Model* 2005, **45(4):**939-51.
41. Todeschini R, Consonni V: *Handbook of Molecular Descriptors* New York: Wiley-VCH; 2002.
42. Gasteiger J, Engel T, (Eds): *Chemoinformatics: a Textbook* New York, NY, USA: Wiley; 2003.
43. Azencott CA, Ksikes A, Swamidass SJ, Chen JH, Ralaivola L, Baldi P: **One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties.** *J Chem Inf Model* 2007, **47(3):**965-974.
44. Kashima H, Tsuda K, Inokuchi A: **Kernels for graphs.** In *Kernel Methods in Computational Biology* Edited by: Schölkopf B, Tsuda K, Vert J. The MIT Press, Cambridge, Massachussetts: MIT Press; 2004:155-170.
45. Mahé P, Ralaivola L, Stoven V, Vert JP: **The Pharmacophore Kernel for Virtual Screening with Support Vector Machines.** *J Chem Inf Model* 2006, **46(5):**2003-2014.
46. Ralaivola L, Swamidass SJ, Saigo H, Baldi P: **Graph kernels for chemical informatics.** *Neural Netw* 2005, **18(8):**1093-1110.
47. Boström J, Greenwood JR, Gottfries J: **Assessing the performance of OMEGA with respect to retrieving bioactive conformations.** *J Mol Graph Model* 2003, **21(5):**449-462.
48. Jaakkola T, Diekhans M, Haussler D: **A Discriminative Framework for Detecting Remote Protein Homologies.** *J Comput Biol* 2000, **7(1,2**95-114 [http://www.cse.ucsc.edu/research/compbio/discriminative/Jaakola2-1998.ps].
49. Leslie C, Eskin E, Noble W: **The spectrum kernel: a string kernel for SVM protein classification.** In *Proceedings of the Pacific Symposium on Biocomputing 2002* Edited by: Altman RB, Dunker AK, Hunter L, Lauerdale K, Klein TE. Singapore: World Scientific; 2002:564-575.
50. Tsuda K, Kin T, Asai K: **Marginalized Kernels for Biological Sequences.** *Bioinformatics* 2002, **18:**S268-S275.
51. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS: **Mismatch string kernels for discriminative protein classification.** *Bioinformatics* 2004, **20(4):**467-476.

52. Saigo H, Vert JP, Ueda N, Akutsu T: **Protein homology detection using string alignment kernels.** *Bioinformatics* 2004, **20(11):**1682-1689.
53. Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, Leslie C: **Profile-based string kernels for remote homology detection and motif extraction.** *J Bioinform Comput Biol* 2005, **3(3):**527-550.
54. Cuturi M, Vert JP: **The context-tree kernel for strings.** *Neural Netw* 2005, **18(8):**1111-23.
55. Dobson P, Doig A: **Predicting enzyme class from protein structure without alignments.** *J Mol Biol* 2005, **345:**187-199.
56. Borgwardt K, Ong C, Schönauer S, Vishwanathan S, Smola A, Kriegel HP: **Protein function prediction via graph kernels.** *Bioinformatics* 2005, **21(Suppl 1):**i47-i56.
57. Qiu J, Hue M, Ben-Hur A, Vert JP, Noble WS: **A structural alignment kernel for protein structures.** *Bioinformatics* 2007, **23(9):**1090-1098.
58. Vert JP: **A tree kernel to analyze phylogenetic profiles.** *Bioinformatics* 2002, **18:**S276-S284.
59. Evgeniou T, Micchelli C, Pontil M: **Learning multiple tasks with kernel methods.** *J Mach Learn Res* 2005, **6:**615-637 [http://jmlr.csail.mit.edu/papers/volume6/evgeniou05a].
60. Kratochwil NA, Malherbe P, Lindemann L, Ebeling M, Hoener MC, Mühlemann A, Porter RHP, Stahl M, Gerber PR: **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* 2005, **45(5):**1324-1336.
61. Surgand JS, Rodrigo J, Kellenberger E, Rognan D: **A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.** *Proteins* 2006, **62(2):**509-538.
62. Okada T, Sugihara M, Bondar AN, Elstner M, Entel P, Buss V: **The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure.** *J Mol Biol* 2004, **342(2):**571-583.
63. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC: **High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.** *Science* 2007, **318(5854):**1258-1265.
64. Russell RB, Barton GJ: **Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.** *Proteins* 1992, **14(2):**309-323.
65. Lefkowitz RJ, Sun JP, Shukla AK: **A crystal clear view of the beta2-adrenergic receptor.** *NatBiotechnol* 2008, **26(2):**189-191.
66. Avlani VA, Gregory KJ, Morton CJ, Parker MW, Sexton PM, Christopoulos A: **Critical role for the second extracellular loop in the binding of both orthosteric and allosteric G protein-coupled receptor ligands.** *J Biol Chem* 2007, **282(35):**25677-25686.
67. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31(13):**3497-3500.
68. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89(22):**10915-10919.
69. Mirzadegan T, Benkö G, Filipek S, Palczewski K: **Sequence analyses of G-protein-coupled receptors: similarities to rhodopsin.** *Biochemistry* 2003, **42(10):**2759-2767.
70. Kristiansen K, Dahl SG, Edvardsen O: **A database of mutants and effects of site-directed mutagenesis experiments on G protein-coupled receptors.** *Proteins* 1996, **26:**81-94.
71. Kobilka BK: **G protein coupled receptor structure and activation.** *Biochim Biophys Acta* 2007, **1768(4):**794-807.
72. Yao X, Parnot C, Deupi X, Ratnala VRP, Swaminath G, Farrens D, Kobilka B: **Coupling ligand structure to specific conformational switches in the beta2-adrenoceptor.** *Nat Chem Biol* 2006, **2(8):**417-422.
73. Chen JZ, Wang J, Xie XQ: **GPCR structure-based virtual screening approach for CB2 antagonist search.** *J Chem Inf Model* 2007, **47(4):**1626-1637.
74. Deupi X, Dölker N, Lòpez-Rodrìguez ML, Campillo M, Ballesteros JA, Pardo L: **Structural models of class a G protein-coupled receptors as a tool for drug design: insights on transmembrane bundle plasticity.** *Curr Top Med Chem* 2007, **7(10):**991-998.
75. Hopkins AL, Groom CR: **The druggable genome.** *Nat Rev Drug Discov* 2002, **1(9):**727-730.
76. Argyriou A, Evgeniou T, Pontil M: **Multi-task feature learning.** In *Adv Neural Inform Process Syst 19* Edited by: Schölkopf B, Platt J, Hoffman T. Cambridge, MA: MIT Press; 2007:41-48.
77. Bonilla E, Chai KM, Williams C: **Multi-task Gaussian Process Prediction.** In *Advances in Neural Information Processing Systems 20* Edited by: Platt J, Koller D, Singer Y, Roweis S. Cambridge, MA: MIT Press; 2008.
78. Abernethy J, Bach F, Evgeniou T, Vert JP: **A new approach to collaborative filtering: operator estimation with spectral regularization.** *J Mach Learn Res* 2008 in press.
79. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics.** *J Mol Graph* 1996, **14:**33-8, 27–8.