Medical & Biological Engineering & Computing manuscript No.
(will be inserted by the editor)

The Influence of Alignment-Free Sequence Representations on the Semi-Supervised Classification of Class C G Protein-Coupled Receptors

Semi-supervised Classification of class C GPCRs

Raúl Cruz-Barbosa · Alfredo Vellido · Jesús Giraldo

Received: date / Accepted: date

Abstract G protein-coupled receptors (GPCRs) are integral cell membrane proteins of relevance for pharmacology. The tertiary structure of the transmembrane domain, a gate to the study of protein functionality, is unknown for almost all members of class C GPCRs, which are the target of the current study. As a result, their investigation must often rely on alignments of their amino acid sequences. Sequence alignment entails the risk of missing relevant information. Various approaches have attempted to circumvent this risk through alignment-free transformations of the sequences on the basis of different amino acid physicochemical properties. In this paper, we use several of these alignment-free methods, as well as a basic amino acid composition representation, to transform the available sequences. Novel semi-supervised statistical machine learning methods are then used to discriminate the different

Raúl Cruz-Barbosa
Computer Science Institute, Universidad Tecnológica de la Mixteca, Huajuapan, Oaxaca, México
E-mail: rcruz@mixteco.utm.mx
Present address: Institut de Neurociències and Unitat de Bioestadística, Universitat Autònoma de Barcelona, Bellaterra, Spain

Alfredo Vellido
Departament de Ciències de la Computació, Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, Spain
E-mail: avellido@cs.upc.edu

Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain

Jesús Giraldo
Institut de Neurociències and Unitat de Bioestadística, Universitat Autònoma de Barcelona, Bellaterra, Spain
Tel.: +34-93-5813813
Fax: +34-93-5812344
E-mail: Jesus.Giraldo@uab.es

class C GPCRs types from the transformed data. This approach is relevant due to the existence of orphan proteins to which type labels should be assigned in a process of deorphanization or reverse pharmacology. The reported experiments show that the proposed techniques provide accurate classification even in settings of extreme class-label scarcity and that fair accuracy can be achieved even with very simple transformation strategies that ignore the sequence ordering.

Keywords Class C G protein-coupled receptors · Semi-supervised learning · Alignment-free sequence representations

1 Introduction

G protein-coupled receptors (GPCRs) are integral cell membrane proteins of great relevance to normal physiology and pathology due to their role in transducing a wide range of extracellular signals after specific ligand binding. It has been reported that 36% of all drugs that were approved by the US Food and Drug Administration during the past three decades target GPCRs [31]. For this reason, and beyond basic research, GPCRs have become the subject of a vast research effort from the pharmaceutical industry [28].

The first GPCR crystal structure, that of rhodopsin, was fully-determined in 2000 [29], and it is only in recent years that the structures of some other distinct receptors, most of them belonging to GPCR class A, have been determined [19].

An alternative way to work on GPCR structural models, when the tertiary 3D crystal structures are not available, is the investigation of their functionality through the analysis of their primary structure: the amino acid sequences, which are well-documented and of which databases are publicly available. This assumes that protein function is encoded in its sequence.

The research reported in this study focuses on class C GPCRs, a receptor family that has of late become an increasingly important target for new therapies [21]. Their remarkable complexity and their sequence diversity makes them an especially attractive system for sequence classification. The discrimination of class C GPCR sequences into the seven different types this class consists of in situations of partial label availability is the ultimate aim of our investigation. The correct discrimination of class C into types may constitute a first step in the study of the molecular processes involved in receptor signalling, both in normal and pathological conditions.

Many of the existing classification approaches use aligned versions of protein sequences [18]. Sequence alignment allows the application of more conventional quantitative analysis techniques, but at the price of risking the loss of relevant information contained in the discarded sequence fragments.

Different approaches have attempted the analysis of alignment-free sequences on the basis of their transformation according to the amino acid physicochemical properties (for a recent review see, for instance, [23]). In the

current study, we use several of them to transform the data sequences, including the Auto-Cross Covariance (ACC) transform [22] and the physicochemical distance transformation (PDT: [23]). In a less complex procedure described in [10] and [26], each GPCR is represented using the means of five z-scale physicochemical descriptors over the sequence. Finally, we also use a further and very simple amino acid sequence transformation that consists on considering only the relative frequencies of appearance of the 20 amino acids in the sequence (thus ignoring the sequential order).

Statistical machine learning (SML) techniques have only recently begun to be applied in bioinformatics and in the -omics sciences [2]. SML models aim to couple the theoretical soundness of probability theory with the data analysis flexibility provided by machine learning, in the common goal of pattern recognition.

In this paper, semi-supervised SML generative models of the manifold learning family [7,8] are applied to the analysis of alignment-free sequences of class C GPCRs, transformed according to the physicochemical properties of their constituent amino acids. The rationale behind the use of semi-supervised approaches is the fact that they can deal with a very common problem in real scientific databases: the only partial availability of class labels. In the case of protein sequential data, this would help the analyst to typify sequences with missing class labels according to the available ones. This is the scenario of, for instance, sequence deorphanization, the process (also known as reverse pharmacology) for which an orphan receptor (a receptor of an initially unknown ligand) matches its natural ligand.

The results of the experiments reported in this paper, which build on those preliminarily presented in [9], indicate that semi-supervised methods working on the physicochemical properties of alignment-free class C GPCR sequences can quite accurately discriminate between the seven types that constitute this class, even in settings of extreme class-label scarcity. Amongst these methods, semi-supervised Generative Topographic Mapping (SS-GTM) consistently yielded the best accuracy results. The use of the ACC data transformation is also shown to provide the most accurate classification. A further interesting finding is that fair accuracy can be achieved even with a very simple transformation that completely ignores the ordering in the protein sequence. The novel inductive version of the SS-GTM proposed in this study is shown in our experiments to infer the most probable class C type labels for several orphan sequences.

2 Materials and Methods

2.1 Materials

GPCRs are the most abundant family of membrane-bound receptors, with more than 800 members in the human proteome [13]. From the determination of the first GPCR crystal structure, that of rhodopsin, in 2000 [29], only in

recent years the structures of some other 15 distinct receptors, all belonging to GPCR class A, have been determined [19]. Having solved in part some of the most stringent difficulties of GPCR crystallography, the number of receptor structures is rapidly growing. Thus, 9 structures from the former list of 15 reported in [19] were published just in 2012, while two new class A-ones [37, 38] and, for the first time [39], one not belonging to this family but to the Frizzled class [12] and two to class B [16,33] were reported in 2013. At the time of writing, the first structures of the 7TM domains of two class C receptors have just been published [42,11].

The characterization of GPCR structure is being made possible thanks to efforts from different laboratories, following diverse approaches [19]. In particular, one of the most active initiatives in the field, the GPCR Network (responsible for the determination of 12 out of 21 current crystal GPCR structures) aimed at achieving 40-60% structural coverage of non-olfactory receptors for the 2010-2015 period by a combination of experimentally solved structures and computationally predicted GPCR 3D-models, if a 35% sequence identity is established as a threshold for GPCR accurate homology modeling [34].

The current study focuses on class C GPCRs, which have become an increasingly important target for new therapies, particularly in areas such as Fragile-X syndrome, schizophrenia, Alzheimer's disease, Parkinson's disease, epilepsy, L-DOPA-induced dyskinesias, generalized anxiety disorder, migraine, chronic pain, gastroesophageal reflux disorder, hyperparathyroidism and osteoporosis[21].

Because of its specificity, data were taken from GPCRDB, which is defined [36] as a molecular-class information system that collects, combines, validates and disseminates large amounts of heterogenous data on GPCRs. GPCRDB divides the GPCR superfamily in 5 families: the class A Rhodopsin like, the class B Secretin like, the class C Metabotropic glutamate/pheromone, Vomeronasal receptors (V1R and V3R) and Taste receptors T2R.

Class C GPCRs were selected for analysis because of (i) their structural complexity, (ii) high sequence-length variability and (iii) therapeutic relevance. Briefly, (i) whereas all GPCRs are characterized by sharing a common seven-transmembrane (7TM) domain, responsible of G protein/β-arrestin activation, most class C GPCRs include, in addition, an extracellular large domain, the Venus Flytrap (VFT) and a cysteine rich domain (CRD) connecting both [30]. To date, no class C-GPCR 7TM domain has been characterized structurally, although some authors anticipate some progress in this respect, even in the short term [34]. (ii) The full or partial presence of the whole domain structure confers a high sequence-length variability to this family. (iii) The involvement of class C GPCRs in many neurological disorders, as previously mentioned, makes this class an attractive target for drug discovery and development.

Class C is, in turn, subdivided into seven types: *Metabotropic glutamate*, *Calcium sensing*, *GABA-B*, *Vomeronasal*, *Pheromone*, *Odorant* and *Taste*. The investigated dataset consists of a total of 1,510 class C GPCR sequences, obtained from GPCRDB, version 11.3.4 as of March 2011. Their distribution into types is as follows: 351 *Metabotropic glutamate*, 48 *Calcium sensing*, 208

GABA-B, 344 *Vomeronasal*, 392 *Pheromone*, 102 *Odorant* and 65 *Taste*. The lengths of these sequences varied from 250 to 1,995 amino acids.

2.2 Methods

2.2.1 Alignment-Free GPCR Representations

A very common preprocessing step for protein classification is multiple alignment. When this is used, the protein classification results strongly depend on the characteristics of the information provided by the alignment. For GPCRs, some disadvantages of this process are: a) that it restricts the analysis strictly to transmembrane domains (loops and N- and C-terminal regions of GPCR proteins are excluded) and, consequently, relevant biological information is lost; b) that the generation of reliable multiple alignments is difficult to obtain when divergent protein sequences are included.

To avoid these drawbacks, and as an attempt not to renounce to any relevant information that might be conveyed by an amino acid sequence, alignment-free protein representations have been defined in the literature. Among these, some rely on transformations based on the amino acid physicochemical characteristics, such as the ACC transformation [22,41] and the mean transformation [26]. Some advantages of these representations are that: a) they can help to discover divergent receptor genes [20]; b) they do not require a homologous relationship among similar sequences to be assumed; and c) their transformed output can directly be used by standard pattern recognition and machine learning methods.

In this paper, we consider a total of four alignment-free data transformations to obtain fixed-length vectors as input data for semi-supervised SML algorithms.

The first and most simple one reflects the amino acid composition (AAcomp) of the primary sequence: the relative frequencies of occurrence of the 20 amino acids are calculated for each sequence resulting in a $N \times 20$ matrix, where N is the number of sequences in the data set. This transformation does not take into account the relative position of amino acids in the sequence.

The remaining three are physicochemical transformations. The second and third, related by the descriptors obtained in [32], are the mean (MeanT) and ACC transformations, respectively.

MeanT, applied in [26] and [10], consists on describing each amino acid sequence using five z-scales (descriptors) and then using their averages as the final feature vector. That is, each of the resulting feature vectors consists on five values calculated as $(\frac{1}{n}\sum_{i=1}^{n} z1_i, \frac{1}{n}\sum_{i=1}^{n} z2_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} z5_i)^T$, where n is the length of the amino acid sequence and $z1_i, \ldots, z5_i$ are the z-values for the i-th amino acid. An extension of the mean transformation (xMeanT) was developed in [10], in which information about the N- and C-terminus of each protein (sequence) is taken into account in order to improve accuracy results. Out of the 15 elements of its resulting feature vector, the first five are obtained

as in MeanT; the next five elements are the z-score mean from the first 150 amino acids of a sequence (N-terminus); and the last five are obtained from the z-score mean of the last 150 amino acids of the sequence (C-terminus).

For the more sophisticated ACC transformation [22,41], time series models are applied to the protein sequences in order to extract their sequential patterns and, consequently, the extracted information is sequence-order dependent. This representation was originally developed in [41] and then applied and modified in [22] and [27].

The ACC transformation can be described as follows: each sequence is first translated into physicochemical descriptors by representing each amino acid with the five z-scales derived in [32], where these scales are in turn obtained from 26 physicochemical properties. The Auto Covariance (AC) and Cross Covariance (CC) variables are then computed from the transformed sequences. The AC measures the correlation of the same descriptor, d, between two residues separated by a lag, l, along the sequence, and it can be calculated as

$$AC_d(l) = \sum_{i=1}^{n-l} \frac{(v_{d,i} - \bar{v}_d)(v_{d,i+l} - \bar{v}_d)}{(n-l)^p}. \tag{1}$$

The CC variable measures the correlation of two different descriptors between two residues separated by a lag along the sequence, and it can be computed as

$$CC_{dd'}(l) = \sum_{i=1}^{n-l} \frac{(v_{d,i} - \bar{v}_d)(v_{d',i+l} - \bar{v}_{d'})}{(n-l)^p}, \tag{2}$$

where $l = 1, \ldots, L$ and L is the maximal lag, which must be lesser than the length of the shortest sequence in the dataset; n is the total number of amino acids in the sequence; $v_{d,i}$ is the value of descriptor $d = 1, \ldots, D$ ($D = 5$) of an amino acid in a sequence at position i; $\bar{v}_d$ is the mean value of descriptor d across all positions; and p is a degree of normalization.

From these, the ACC fixed-length vectors are obtained: First, the AC and CC terms from eqs. 1 and 2 are concatenated for each lag ($C(l) = [AC(l)\ CC(l)]$) and then the ACC is obtained for a maximum lag L by concatenating the $C(l)$ terms, that is, $ACC(L) = [C(1), \ldots, C(L)]$. Details of this procedure can be found in [22,27].

The fourth and last representation used in our experiments, namely PDT, was recently proposed in [23]. This is an extension of ACC, where the AC, described above, is extended to use 531 physicochemical properties instead of 26. This extended AC-like measure, understood as the distance between two residues separated by a lag, l, along a sequence using the same descriptor, d, is computed as

$$\delta_d(l) = \frac{\sum_{i=1}^{n-l}(I_d(A_i) - I_d(A_{i+l}))^2}{n-l} = \sum_{i=1}^{n-l} \frac{(I_d(A_i) - I_d(A_{i+l}))^2}{n-l}, \tag{3}$$

where $I_d(A_i)$ and $I_d(A_{i+l})$ are the normalized physicochemical property values of amino acid A_i and A_{i+l}, respectively, for index d, which can be calculated

as

$$I_d(A_i) = \frac{\hat{I}_d(A_i) - \sum_{m=1}^{20} \frac{\hat{I}_d(R_m)}{20}}{\sqrt{\frac{\sum_{k=1}^{20} \left(\hat{I}_d(R_k) - \sum_{m=1}^{20} \frac{\hat{I}_d(R_m)}{20} \right)^2}{20}}} \tag{4}$$

and $\hat{I}_d(A_i)$ is the raw value of the physicochemical property of amino acid A_i for index $d = 1, \ldots, D$ ($D = 531$); R_m ($m = 1, \ldots, 20$) represents the 20 standard amino acids.

The fixed length vector provided by PDT has $531 * L$ dimensions, where L is a maximal distance or lag ($l = 1, \ldots, L$, as in the ACC transformation). This poses a potential problem in the form of a very high data dimensionality, even for small values of L. It is noteworthy that the resulting PDT variables do not include, unlike ACC, any CC information of the descriptors based on the 531 physicochemical properties.

2.2.2 Semi-supervised Generative Topographic Mapping

GTM [4] is a latent variable model in which a sample of K regularly-spaced points $\mathbf{u}_k$ residing in a low-dimensional space are mapped into the usually high-dimensional observed data space, each of them defining a prototype point. This prototype $\mathbf{y}_k$ is the image of the former according to the mapping function that takes the form,

$$\mathbf{y}_k = \mathbf{W}\Phi(\mathbf{u}_k), \tag{5}$$

where Φ is a set of M nonlinear basis functions ϕ_m, and $\mathbf{W}$ is a matrix of adaptive weights that defines the specific characteristics of the mapping. The prototype vector $\mathbf{y}_k$ can be seen as a representative of those data points $\mathbf{x}_n$ which are closer to it than to any other prototype and, thus, can also be seen as a cluster centroid. GTM performs a type of vector quantization that is similar to that of Self-Organizing Maps.

The set of prototypes $\mathbf{y}_k$ belongs to an intrinsically low-dimensional smooth manifold that wraps around the observed data $X = \{\mathbf{x}_n\}_{n=1}^N$. In this way, GTM becomes a manifold learning method. If we assume that the observed data lie close to the manifold, the conditional distribution of the observed data variables, given the latent variables, $p(\mathbf{x}|\mathbf{u})$ can be described as a noise model:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi} \right)^{D/2} \exp \left\{ -\frac{\beta}{2} \sum_{d=1}^{D} \left(x^d - y^d(\mathbf{u}) \right)^2 \right\}, \tag{6}$$

with variance β^{-1} and D is the data space dimensionality. From this, we can integrate the latent variables out, obtaining the likelihood of the model, and use maximum likelihood to estimate the adaptive parameters. Details of this procedure can be found in [4].

In many real settings, and orphan GPCRs are an example of this, class labels may not be readily available for all cases. If ultimately interested in the

classification of cases, we are faced with a semi-supervised learning problem in which missing case labels must be inferred on the basis of the available ones.

Recently, GTM was redefined in a semi-supervised setting [7] as SS-Geo-GTM. For this, and understanding the model prototypes and manifold as the elements of a proximity graph, existing label propagation algorithms [44,15] were adapted to a variant of GTM (namely Geo-GTM) in which Euclidean distances were replaced by approximations of geodesic distances along the GTM manifold.

A label vector $\mathbf{T}_k \in [0,1]^c$ (where c are the classes) is associated to each Geo-GTM prototype $\mathbf{y}_k$. The weights of the edges are derived from the graph distances d_g between prototypes. The edge weight between nodes k and k' is calculated as

$$w_{kk'} = \exp(-d_g^2(k,k')/\sigma^2). \tag{7}$$

The available label information of $\mathbf{x}_n \in X$ with class assignment $c(\mathbf{x}_n) = C_t \in \{C_1, \ldots, C_c\}$ is used to fix the label vectors of the prototypes to which they are assigned, so that $T_{k,j} = 1$ if $j = t$, and $T_{k,j} = 0$ otherwise. Unlabeled prototypes will then update their label by propagation, according to $\mathbf{T}_k^{new} = \sum_{k'} w_{kk'} \mathbf{T}_{k'} / \sum_{k'} w_{kk'}$. Unlabeled data items are finally labeled by assignment to the class of highest prevalence on the label vector of the prototype $\mathbf{y}_k$ that bears the highest responsibility for them, according to $c(\mathbf{x}_n) = \arg\max_{C_j \in \{C_1,\ldots,C_k\}} T_{k,j}$.

A detailed description of SS-Geo-GTM can be found in [7], whereas a practical application to a problem in the field of neuro-oncology is described in [8]. In the following experiments, a version of this model that employs Euclidean instead of geodesic distances, namely the SS-GTM, is also used.

2.2.3 Performance Assessment Measures

The semi-supervised multi-class classification results reported in this paper were assessed using two performance measures: the accuracy and the Matthews Correlation Coefficient (MCC) [24]. The former is widely known and used as the proportion of correctly classified cases. The latter is of common use in the bioinformatics field as a performance measure when the analyzed datasets are class-unbalanced. It has of late attracted the attention of the machine learning community due to its inclusion in recent versions of the widely used Weka data mining toolkit. Both accuracy and MCC measures can be naturally extended from the binary to the multi-class context [14].

Let us assume a classification problem with $\mathcal{S}$ samples and G classes, and two functions defined as $tc, pc : \mathcal{S} \to \{1, \ldots, G\}$, where $tc(s)$ and $pc(s)$ return the true and the predicted class of s, respectively. The corresponding square confusion matrix C is:

$$C_{ij} = |\{s \in \mathcal{S} : tc(s) = i \; AND \; pc(s) = j\}|, \tag{8}$$

in which the ij-th entry of C is the number of cases of true class i that have been assigned to class j by the classifier. Then, the confusion matrix notation

can be used to define both the accuracy and the MCC as:

$$accuracy = \frac{\sum_{k=1}^{G} C_{kk}}{\sum_{i,j=1}^{G} C_{ij}}, \tag{9}$$

$$MCC = \frac{\sum_{k,l,m=1}^{G} C_{kk}C_{ml} - C_{lk}C_{km}}{\sqrt{\sum_{k=1}^{G}\left[\left(\sum_{l=1}^{G} C_{lk}\right)\left(\sum_{f,g=1 f\neq k}^{G} C_{gf}\right)\right]}\sqrt{\sum_{k=1}^{G}\left[\left(\sum_{l=1}^{G} C_{kl}\right)\left(\sum_{f,g=1 f\neq k}^{G} C_{fg}\right)\right]}}. \tag{10}$$

MCC takes values in the interval $[-1, 1]$, where 1 means complete correlation (perfect classification), 0 means no correlation (all samples have been classified to be of only one class) and -1 indicates a negative correlation (extreme misclassification case). The MCC measure was originally extended to the multi-class problem in [14]. Recently, and following a comparison between MCC and Confusion Entropy [40] reported in [17], MCC was recommended as an optimal tool for practical tasks, since it presents a good trade-off among discriminatory ability, consistency and coherent behavior with varying number of classes, unbalanced datasets and randomization.

3 Results and Discussion

The experiments whose results are reported next aim to gauge classification performance in situations of only partial class label availability, using semi-supervised methods. Therefore, the protein sequences described in Materials, transformed according to the methods outlined in Methods, were used as an input to the SS-Geo-GTM and SS-GTM models summarily described in the previous section, but also, alternatively and for comparison, to a semi-supervised SVM model for manifold learning (SS-SVMan, [43]). The latter is a variant of the widely used SVM, in which the learning process is modified to accommodate manifold consistency and the hinge loss of class prediction (an approximation to misclassification error). The result is an SVM-like process. There are three parameters involved in the choice of the SS-SVMan model: C, γ, and ρ. The last one is a coefficient that guarantees the invertibility of an expression leading to the obtention of the objective function. The other two parameters, typical of an SVM, are chosen for our experiments as indicated in [43].

All transformed datasets were first modeled using GTM and Geo-GTM, where the latent grid was fixed to a square layout of approximately $(N/2)^{1/2} \times (N/2)^{1/2}$ and N is the number of points in the data set. SS-Geo-GTM, SS-GTM and SS-SVMan were all implemented in MATLAB®. For the experiments reported next, the matrix $\mathbf{W}$ and the inverse variance β in SS-Geo-GTM and SS-GTM were initialized according to a standard procedure described in [4], which ensures the replicability of the results.

The goal of the experiments was threefold. Firstly, we aimed to gauge the influence of the alignment-free amino acid sequence representations in the semi-supervised classification of class C GPCR subfamilies. Secondly, we aimed to compare the performance of the three semi-supervised models in terms of classification performance. In both cases, we were interested not only in overall classification results, but also in the individual results for each of the seven class C GPCR types described in Materials. The latter should shed light on the possibility that some of these types might be easier to discriminate from the rest than others. Thirdly, we aimed to apply a semi-supervised method for simulation purposes of a deorphanization process of some orphan GPCR's.

Since unaligned amino acid sequences have varying lengths and the semi-supervised methods use vectors of common dimensionality as input, data from the seven types of class C GPCRs were first transformed according to the alignment-free representations already described. Data normalization (or standardization) can be applied in such a way that the columns of the data matrix all have zero mean and unit standard deviation. For comparison purposes, experiments were carried out using both normalized and unnormalized versions of the transformed data, and only the best results are reported in the following figures.

The figures of merit used to assess the performance of the semi-supervised models, described in the previous section, are the average classification accuracy over 100 runs and the MCC for multi-class problems. Labels were available for all sequences in the sample originally extracted from the GPCRDB. In order to evaluate the models in a semi-supervised setting, labels were therefore randomly removed (thus becoming *missing* values) in every run of the experiments. The class label effective availability was made to vary from a very extreme (1%) to a relatively relaxed (30%) setting.

The average classification results for the dataset obtained using the sequence order-unrelated AAcomp transformation are graphically displayed in Fig. 1. From these results, the SS-GTM is shown to outperform the rest of methods, both in terms of accuracy and MCC, in the most extreme settings up to 10% labeled data availability, which means that the unsupervised nature of GTM-based models can help to discover the class structure in a better way when very few labeled data are available. On the contrary, when the label availability condition is relaxed, the SS-SVMan model outperforms the GTM-based methods, which means that the supervised nature of SVM-based models is likely to better reveal the class structure only when enough labeled data (as much as 30% in this dataset) are available.

The results for the (also sequence order-unrelated) mean and extended mean transformations are presented in Figs. 2 and 3. Here, the restriction that the first and the last 150 amino acids in a sequence must be taken into account reduces the dataset from 1510 to 1494 items: those with a length greater or equal to 300. In both figures, the performance of the three semi-supervised methods follow the same patterns found in Fig. 1. Despite the fact that xMeanT yields better results than MeanT, confirming the results

reported in [10], both are clearly less discriminative than the more simple AAcomp transformation.

The ACC transformation uses two parameters that must be set to adequate values prior to classification: the maximum lag L and the degree of normalization p. In this study, their optimal values were experimentally chosen by investigating the impact of multiple combinations of possible values on classification accuracy. Previous experiments, as those reported in [22] and [27], have shown that the maximum lag is to be found in the range [1,160]. Following [27], and the rationale that a large maximum lag implies high data dimensionality and the problems that come with it (curse of dimensionality), we searched for L in the range of 1 to 30. The p parameter was set at different values, including: 0, 0.5 and 1.0. The average classification results were computed using 30% of labeled data for each combination. As an illustration, the classification accuracy results for p values of 0.5 and 1.0 are shown in Online Resource 1 (the results for $p = 0$ were not stable). Out of these, $p = 0.5$ provides the best results. A classification accuracy of around 85% was achieved with a lag of 7 and results stabilized from a value of 13 onwards. For computational time expediency, a maximal lag of 13 was thus selected for the complete ACC experiments.

The average classification results for the ACC representation, with $L = 13$ and $p = 0.5$, are shown in Fig. 4. The performance of the three analyzed methods follows the same pattern found in Fig. 1 for AAcomp, but this time more pronounced in favour of GTM-based models. The results for SS-GTM and SS-Geo-GTM are very similar and these, in turn, clearly outperform the SS-SVMan using from 1% to 20% labeled data. SS-SVMan is competitive only when enough labeled data (30% in this case) are available. The accuracy and MCC results are, again, quite consistent, which suggests that the models have not been strongly affected by class-unbalance. Notice though that the result for SS-SVMan at 1% label availability is an exception to this pattern, suggesting that the model has been very negatively affected in this setting.

Given that the PDT transformation is an extension of the ACC, the L (maximal distance) parameter must also be tuned. Following [23], this parameter was experimentally chosen by investigating the impact of varying its values, in the range of 1 to 10, on the resulting classification accuracy. The average classification results were computed using 30% of labeled data for each value, as shown in Online Resource 2. A lag of value 8 was selected because a maximal classification accuracy was achieved with this value. The corresponding average classification results using the PDT representation with $L = 8$ are shown in Fig. 5. The performance of the analyzed models is again consistent with that reported in Figs. 1 and 4, with GTM-based models yielding the best results when a low percentage of labels is available and the SVM-based model becoming the most efficient for label availability of 20% and 30%. A detailed account of the means and standard deviations of the accuracy and MCC results for all data representations used in this section can be found in Online Resources 3 and 4.

As could be expected, the ACC transformation yields consistently better classification than the far simpler AAcomp data representation. Surprisingly though, the differences are only modest (in the range of 3-4% average accuracy and 0.02-0.06 MCC) when enough labels become available (30%). This means that the frequency of amino acid occurrence within the sequence conveys, by itself, relevant discriminatory information. It also means that the sequence ordering itself adds only limited type-discriminatory information. Alternatively, it could also mean that much of the sequence order information is already implicitly present in the frequency of amino acid occurrence.

Interestingly, the PDT transformation does not provide any advantage for the analyzed data when compared to ACC in terms of discrimination for the GTM-based models. In fact, its performance is only slightly better than that obtained with the AAcomp transformation. It is only when used with SS-SVMan that this performance is superior to the best obtained with ACC using GTM-based models. This relatively poor performance for the analyzed data might be explained by three factors (acting separately or in combination):

- The high-dimensionality of the data resulting from the PDT transformation, which could hamper classification due to excessive sparsity.
- The possibility that the supplementary physicochemical information considered by PDT is either highly redundant or uninformative for type-discrimination purposes.
- The PDT transformation, unlike ACC, does not include the descriptor cross-covariance information of the physicochemical properties. This might be counteracting the potentially beneficial effect of a richer physicochemical information.

The power of the ACC representation can be further appreciated in the following variation on the previous experiments. Instead of considering a complete 20 amino acid *vocabulary* to represent the class C GPCR sequences, we now consider a simplified vocabulary of 7 "grouped residues". Using a group definition based on that described in [5], the following 7 groups of amino acids were considered: *aliphatic hydrophobic* (alanine, valine, isoleucine, leucine, methionine, proline), *aromatic hydrophobic* (phenylalanine), *positively charged* (lysine, arginine), *negatively charged* (aspartic acid, glutamic acid), *aliphatic polar* (serine, threonine, cysteine, asparagine, glutamine), *aromatic polar* (tyrosine, histidine, tryptophan) and *glycine*. Then, for each sequence, each amino acid was replaced by its corresponding group. Obviously, some biological information must be lost in this simplified translation.

In the first step of the transformation, each sequence is translated into physicochemical descriptors by representing each amino acid with the five z-scales derived in [32]. To do this in the newly coded sequences, each "grouped residue" was represented by the five z-scales averaged over the amino acids that belong to it. Then, the ACC transformation was carried out. The L and p parameters were again experimentally selected according to the results in Online Resource 5, which show that a maximum classification accuracy was achieved with a lag of 15 and $p = 1.0$.

The classification results for this simplification of the ACC representation are summarized in Fig. 6. They are indeed consistently lower than those of the original ACC, but very similar to those obtained with the PDT transformation even if obtained from a far less rich physicochemical information content.

3.1 Class C GPCR Type-Specific Classification and its Contribution to Overall Classification

As previously stated, we are interested in finding to what extent each of the seven class C GPCR types described in Materials can be discriminated from the rest and how each of them influences the overall classification performance. Previous research, using fully unsupervised kernel visualization methods applied to aligned class C GPCRs, has shown that some of these types might be easier to discriminate from the rest than others [35]. More specifically, it has been shown that the *GABA-B* type is clearly distinct from the rest (this is not unexpected, as GABA-B is structurally different from the rest of class C GPCRs in that, while including the VFT and 7TM domains, lacks of the connecting CRD), whereas Vomeronasal, Pheromone, and Odorant are difficult to discriminate from each other (again, not unexpected given their shared relationship to the odor function). Somewhere in between, *Metabotropic glutamate*, *Calcium sensing*, and *Taste* have fairly specific features according to their aligned representation. Furthermore, these different levels of discriminability have been shown to be consistent with the standard phylogenetic tree representations of their aligned sequences [6].

The classification accuracies as a function of the percentage of available labels are presented, for all class C GPCR types, in Fig. 7 for each of the most discriminative data transformation method (i.e. excluding the variants of mean transformation). Following the same order as for the overall accuracy, we first display the results for AAcomp in the first row of Fig. 7. The results for ACC, PDT and the "grouped residues" simplification of ACC are displayed, in turn, in the second, third and fourth row.

At a glance, it is clear from these figures that the overall pattern of semi-supervised classification is quite stable across representation methods. The main differences are actually observed not between different sequence representations but between GTM-based and SVM-based methods. For both, some types of class C GPCRs achieve high classification accuracy even with low label percentages and increase it steadily thereafter. Other types, though, require many more labels to increase their accuracy in any significant way.

Two individual results require further clarification: SS-Geo-GTM for AAcomp and SS-SVMan for ACC with "residues groups" are atypical as they both show the *Metabotropic glutamate* type to be almost perfectly classified even for the extreme 1% label-availability limit, while, simultaneously, the rest of types yield uncharacteristically poor accuracies than only recover *normal* values at the highest levels of label availability. The most likely reason for this anomaly (which also shows *Metabotropic glutamate* losing accuracy as label

availability increases, something that is counterintuitive) is that, in the absence of enough labels, the underlying models are completely biased towards a type of high prevalence (*Metabotropic glutamate* is the 2^{nd} most represented type in the dataset) that is also relatively easy to discriminate. These models unduly favour *Metabotropic glutamate* in detriment of the rest of types.

For GTM-based methods, *GABA-B* yields the best accuracies across all representation methods, with the aforementioned single exception of AAcomp with SS-GeoGTM. These results agree completely with the aligned-sequence cluster structure visualization reported in [35,6]. *Metabotropic glutamate* is second best in most cases, hitting an accurate classification upper limit quite early in terms of label availability. *Taste* and *Calcium sensing* achieve similar accuracies to *Metabotropic glutamate*, but they require a sizeable proportion of labels for that. On the contrary, their classification deteriorates alarmingly when class labels are scarce. This might reflect either the fact that these are the types with lower prevalence in the analyzed dataset, or the intrinsic heterogeneity of these types. *Vomeronasal*, *Pheromone* and, most notably, *Odorant* are the most difficult types to discriminate. Again, this is in accordance with the results reported in [35,6].

The type-specific results are different for SS-SVMan. In this case, *Metabotropic glutamate* is clearly the most differentiated type for all representations, with high accuracies even in extreme settings of label availability. On the opposite side, *Taste* and, specially, *Odorant* yield the worst accuracies. In between, the rest of types are fairly easy to discriminate, at least in the presence of a big enough number of labels. Across representations, though, SS-SVMan completely fails to recognize most types at low levels of label availability, reaching accuracies lower than 10% for *Odorant*. This explains the low global accuracies achieved by SS-SVMan in this setting.

The failure of SS-SVMan at low levels of label availability and the differences with GTM-based methods could at least be partially explained by the sensitivity of the former to differences in type prevalence. In other words, the SS-SVMan might be failing to classify those types that are less represented in the dataset. According to the figures in Materials, *Odorant* and *Taste* are two of the three least numerous types. The third one is *Calcium sensing*, for which reasonably good accuracies are obtained. The reason for the latter result might be the fact that this type is very homogeneous (low within-type variability) and different enough from the rest, as reported in [6]. The sensitivity to relative type-prevalence might also be behind the unusually good performance of *Pheromone* and *Vomeronasal*, because these are both highly represented in the analyzed dataset.

3.2 Deorphanization Experiment using an Inductive Version of SS-GTM

In this section, we simulate a process of deorphanization of some sequences which are known to belong to the class C GPCR family [30,1] but for whom the corresponding natural ligand is unknown. For this, we developed a novel

inductive version of the naturally transductive SS-GTM model. Inductive models have the advantage of using the best previously trained model for testing a new or unknown observation (a sequence, in this case) and classify it in the corresponding class. In contrast, transductive models for classification of new or unknown observations need to retrain the previously obtained models including the new observations, a procedure that is computationally expensive.

Following [3], we transform the transductive SS-GTM model presented in section 2.2.2 into an inductive version. This can be accomplished by using the prototype points, $\mathbf{y}_k$, (Eq.5) with their corresponding label vectors, $\mathbf{T}_k$, as well as the σ parameter found by the transductive version, as described next.

First, the weight vector $\mathbf{W}$ of the edges between the new observation, $\mathbf{x}^{new}$, and the prototype points, $\mathbf{y}_k$, is computed as in Eq.7, which now uses the Euclidean distance, as SS-GTM requires it. Then, the corresponding label vector for $\mathbf{x}^{new}$ is calculated as:

$$\mathbf{T}^{new} = \frac{\sum W_{x^{new},\mathbf{y}_k} \mathbf{T}_k}{\sum_k W_{x^{new},\mathbf{y}_k} + \epsilon}, \tag{11}$$

where ϵ is a regularization term used for numerical stability, as suggested in [3], and the $\sum$ symbol (without subindex) denotes sums along the columns of the matrix formed by $W_{x^{new},\mathbf{y}_k} \mathbf{T}_k$. Finally, the class label asigned to $\mathbf{x}^{new}$ is given by $c(\mathbf{x}^{new}) = \arg\max_{C \in \{1,\ldots,c\}} \mathbf{T}^{new}$.

We can now apply this inductive version of SS-GTM in some illustrative experiments concerning the deorphanization of sequences GPRC5A, GPRC5B, GPRC5C and GPRC5D, which are known to belong to the class C GPCR family and can be accessed at the UniProt knowledge base[1] with identifiers: Q8NFJ5, Q9NZH0, Q9NQ84 and Q9NZD1, respectively. In this database, Q9NZH0 and Q9NQ84 are provided with two isoform sequences and Q9NZD1, with three. The resulting eight sequences were analyzed using the ACC transformation where the necessary parameters were set as in section 3.

The results of their semi-supervised classification using the inductive SS-GTM for deorphanization are summarized in Table 1. We observe that these results are consistent in the sense that the isoform sequences are classified within the same subfamily in all cases. The results in section 3.1 indicated that the metabotropic glutamate subfamily is one of the two subfamilies best discriminated by SS-GTM, which reinforces the reliability of the assignment of the isoform sequences of Q9NZH0 and Q9NQ84 to the metabotropic glutamate subfamily. Given that the same results indicate that the Odorant subfamily is the worse discriminated by SS-GTM, the assignment of the isoform sequences of Q9NZD1 to the subfamily of Odorant should be considered with more caution. Note that all these orphan GPCRs have been linked with the *GABA-B* subfamily with which they have been reported to share high levels of sequence homology, despite not having long N-termini for ligand binding [25]. Indeed, these preliminary semi-supervised subfamily assignments should be further confirmed by pharmacological laboratory tests.

[1] http://www.uniprot.org/

Tables and Figures

Fig. 1 (a) Average accuracy and (b) MCC results using the AAcomp representation

Fig. 2 (a) Average accuracy and (b) MCC results using the MeanT representation

Fig. 3 (a) Average accuracy and (b) MCC results using the xMeanT representation

Fig. 4 (a) Average accuracy and (b) MCC results using the ACC representation

Fig. 5 (a) Average accuracy and (b) MCC results using the PDT representation

Fig. 6 (a) Average accuracy and (b) MCC results using the ACC representation for "grouped residues"

Fig. 7 Class-specific percentage of contribution to overall classification for (first-row) AA-comp representation, (second-row) ACC, (third-row) PDT and (fourth-row) ACC representation of "grouped residues". Left-column: SS-GTM model; center-column: SS-Geo-GTM (with label simplified as SS-Geo); right-column: SS-SVMan. In the inlaid legends, mGluR stands for *Metabotropic glutamate* and CaSR stands for *Calcium sensing*

Table 1 The semi-supervised classification results by inductive SS-GTM for some orphan GPCR sequences

Sequence ID	UniProt ID	Assigned Subfamily
GPRC5A	Q8NFJ5	Taste
GPRC5B	Q9NZH0-1	Metabotropic glutamate
	Q9NZH0-2	Metabotropic glutamate
GPRC5C	Q9NQ84-1	Metabotropic glutamate
	Q9NQ84-2	Metabotropic glutamate
GPRC5D	Q9NZD1-1	Odorant
	Q9NZD1-2	Odorant
	Q9NZD1-3	Odorant

4 Conclusions

The discovery of the tertiary structures of GPCRs has of late quickened its pace, mostly thanks to several innovative protein engineering techniques and crystallography methods. Despite this, an overwhelming majority of these advances relate to class A receptors. Given the interest of class C receptors in pharmacology, and in the absence of much knowledge regarding their 3D crystal structures, the investigation of their functionality can be approached through the analysis of their primary structure in the form of amino acid sequences.

Alignment-free representations of these sequences ensure that no relevant primary information is lost. Several state-of-the-art as well as basic sequence representations of this type have been investigated in this paper in the context of semi-supervised classification of class C GPCR types. Among these, it has been shown that the sequence order-dependent ACC transformation (based on physicochemical properties of the amino acids) captures the most discriminative characteristics of the class C sequences. Furthermore, it has been found that accuracy does not decrease dramatically even when a very simple order-independent amino acid composition transformation is used, suggesting that most information is encoded in the own proportion of amino acids present in the sequence.

In addition to this, type-specific classification results have shown that the discriminative ability of the classifiers for each type varies according to the utilized data representation, but keeping, in general, a stable and consistent classification pattern across all representations. Moreover, and importantly for the problem of deorphanization in reverse pharmacology, the experimental results indicate that semi-supervised methods working on the physicochemical properties of alignment-free class C GPCR sequences can quite accurately discriminate between the seven types that constitute this class in settings of extreme class-label scarcity. Amongst these methods, SS-GTM consistently yielded the best accuracy results. Some preliminary experiments using a novel and fully inductive version of SS-GTM have yielded promising results that open the door to an at least partially automated quantitative procedure for sequence deorphanization.

Given the need for robust bioinformatics tools well-suited for classification and functional analysis of GPCRs, we expect semi-supervised methods, including clustering and graphical network-based techniques, to become useful alternatives for the challenges involved in these tasks.

Acknowledgements R. Cruz-Barbosa acknowledges the Mexican National Council for Science and Technology for his postdoctoral fellowship. This research is partially funded by Spanish research projects TIN2012-31377, SAF2010-19257, Fundació La Marató de TV3 (110230) and RecerCaixa 2010ACUP 00378.

References

1. Alexander, S.P.H., Benson, H.E., Faccenda, E., Pawson, A.J., Sharman, J.L., Spedding, M., Peters, J.A., Harmar, A.J., CGTP-Collaborators: The concise guide to pharmacology 2013/14: G protein-coupled receptors. Br J Pharmacol **170**, 1459–1581 (2013)
2. Aliferis, C.F., Statnikov, A., Tsamardinos, I.: Challenges in the analysis of mass-throughput data: A technical commentary from the statistical machine learning perspective. Cancer Inform **2**, 133–162 (2006)
3. Bengio, Y., Delalleau, O., Roux, N.L.: Semi-Supervised Learning, chap. Label Propagation and Quadratic Criterion, pp. 193–216. The MIT Press (2006)
4. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Comput **10**, 215–234 (1998)
5. Branden, C., Tooze, J.: Introduction to Protein Structure. Garland Publishing (1991)
6. Cárdenas, M.I., Vellido, A., Olier, I., Rovira, X., Giraldo, J.: Complementing kernel-based visualization of protein sequences with their phylogenetic tree. In: Lecture Notes in Bioinformatics (LNCS/LNBI), Vol.7548, pp. 136–149 (2012)
7. Cruz-Barbosa, R., Vellido, A.: Semi-supervised geodesic Generative Topographic Mapping. Pattern Recognit Lett **31**, 202–209 (2010)
8. Cruz-Barbosa, R., Vellido, A.: Semi-supervised analysis of human brain tumours from partially labeled MRS information, using manifold learning models. Int J Neural Syst **21**, 17–29 (2011)
9. Cruz-Barbosa, R., Vellido, A., Giraldo, J.: Advances in semi-supervised alignment-free classification of G protein-coupled receptors. In: Procs. of the International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO'13), pp. 759–766 (2013)
10. Davies, M.N., Secker, A., Freitas, A.A., Mendao, M., Timmis, J., Flower, D.R.: On the hierarchical classification of G protein-coupled receptors. Bioinformatics **23**(23), 3113–3118 (2007)
11. Doré, A.S., Okrasa, K., Patel, J.C., Serrano-Vega, M., Bennett, K., Cooke, R.M., Errey, J.C., Jazayeri, A., Khan, S., Tehan, B., Weir, M., Wiggin, G.R., Marshall, F.H.: Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. Nature **551**, 557–562 (2014)
12. Foord, S.M., Bonner, T.I., Neubig, R.R., Rosser, E.M., Pin, J.P., Davenport, A.P., Spedding, M., Harmar, A.J.: International Union of Pharmacology. XLVI. G protein-coupled receptor list. Pharmacol Rev **57**(2), 279–288 (2005)
13. Fredriksson, R., Lagerström, M.C., Lundin, L.G., Schiöth, H.B.: The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol **63**, 1256–1272 (2003)
14. Gorodkin, J.: Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem **28**, 367–374 (2004)
15. Herrmann, L., Ultsch, A.: Label propagation for semi-supervised learning in self-organizing maps. In: Procs. of the 6^{th} International Workshop on Self-Organizing Maps (WSOM) (2007)
16. Hollenstein, K., Kean, J., Bortolato, A., Cheng, R.K., Doré, A.S., Jazayeri, A., Cooke, R.M., Weir, M., Marshall, F.H.: Structure of class B GPCR corticotropin-releasing factor receptor 1. Nature (2013). DOI 10.1038/nature12357. Available online
17. Jurman, G., Riccadonna, S., Furlanello, C.: A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE **7**(8), e41,882 (2012)
18. Karchin, R., Karplus, K., Haussler, D.: Classifying G-protein coupled receptors with support vector machines. Bioinformatics **18**, 147–159 (2002)
19. Katritch, V., Cherezov, V., Stevens, R.C.: Structure-function of the G protein-coupled receptor superfamily. Annu Rev Pharmacol Toxicol **53**, 531–556 (2013)
20. Kim, J., Moriyama, E.N., Warr, C.G., Clyne, P.J., Carlson, J.R.: Identification of novel multi-transmembrane proteins from genomic databases usin quasi-periodic structural properties. Bioinformatics **16**, 767–775 (2000)
21. Kniazeff, J., Prézeau, L., Rondard, P., Pin, J.P., Goudet, C.: Dimers and beyond: The functional puzzles of class C GPCRs. Pharmacol Ther **130**, 9–25 (2011)

22. Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., Wikberg, J.E.S.: Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci **11**, 795–805 (2002)
23. Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X.: Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS ONE **7**, e46,633 (2012)
24. Matthews, B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta - Protein Structure **405**, 442–451 (1975)
25. Oh, D.Y., Kim, K., Kwon, H.B., Seong, J.Y.: Cellular and molecular biology of orphan G protein-coupled receptors. Int Rev Cytol **252**, 163–218 (2006)
26. Opiyo, S.O., Moriyama, E.N.: Protein family classification with partial least squares. J Proteome Res **6**, 846–853 (2007)
27. Otaki, J.M., Mori, A., Itoh, Y., Nakayama, T., Yamamoto, H.: Alignment-free classification of G-protein-coupled receptors using self-organizing maps. J Chem Inf Model **46**, 1479–1490 (2006)
28. Overington, J.P., Al-Lazikani, B., Hopkins, A.L.: How many drug targets are there? Nat Rev Drug Discov **5**, 993–996 (2006)
29. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., et al.: Crystal structure of rhodopsin: a G protein-coupled receptor. Science **289**, 739–45 (2000)
30. Pin, J.P., Galvez, T., Prézeau, L.: Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. Pharmacol Ther **98**, 325–354 (2003)
31. Rask-Andersen, M., Sällman-Almén, M., Schiöth, H.B.: Trends in the exploitation of novel drug targets. Nat Rev Drug Discov **10**, 579–590 (2011)
32. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S.: New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem **41**, 2481–2491 (1998)
33. Siu, F.Y., He, M., de Graaf, C., Han, G.W., Yang, D., Zhang, Z., Zhou, C., Xu, Q., Wacker, D., Joseph, J.S., Liu, W., Lau, J., Cherezov, V., Katritch, V., Wang, M.W., Stevens, R.C.: Structure of the human glucagon class B G-protein-coupled receptor. Nature (2013). DOI 10.1038/nature12393. Available online
34. Stevens, R.C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H., Wüthrich, K.: The GPCR network: a large-scale collaboration to determine human GPCR structure and function. Nat Rev Drug Discov **12**, 25–34 (2013)
35. Vellido, A., Cárdenas, M.I., Olier, I., Rovira, X., Giraldo, J.: A probabilistic approach to the visual exploration of G protein-coupled receptor sequences. In: Procs. of the 19^{th} European Symposiun on Artificial Neural Networks (ESANN 2011), pp. 233–238 (2011)
36. Vroling, B., Sanders, M., Baakman, C., Borrmann, A., Verhoeven, S., Klomp, J., Oliveira, L., de Vlieg, J., Vriend, G.: GPCRDB: information system for G protein-coupled receptors. Nucleic Acids Res **39(suppl.1)**, D309–319 (2011)
37. Wacker, D., Wang, C., Katritch, V., Han, G.W., Huang, X.P., Vardy, E., McCorvy, J.D., Jiang, Y., Chu, M., Siu, F.Y., Liu, W., Xu, H.E., Cherezov, V., Roth2, B.L., Stevens, R.C.: Structural features for functional selectivity at serotonin receptors. Science **340**, 615–619 (2013). DOI 10.1126/science.1232808
38. Wang, C., Jiang, Y., Ma, J., Wu, H., Wacker, D., Katritch, V., Han, G.W., Liu, W., Huang, X.P., Vardy, E., McCorvy, J.D., Gao, X., Zhou, E.X., Melcher, K., Zhang, C., Bai, F., Yang, H., Yang, L., Jiang, H., Roth, B.L., Cherezov, V., Stevens, R.C., Xu, H.E.: Structural basis for molecular recognition at serotonin receptors. Science **340**, 610–614 (2013). DOI 10.1126/science.1232807
39. Wang, C., Wu, H., Katritch, V., Han, G.W., Huang, X.P., Liu, W., Siu, F.Y., Roth, B.L., Cherezov, V., Stevens, R.C.: Structure of the human smoothened receptor bound to an antitumour agent. Nature (2013). DOI 10.1038/nature12167
40. Wei, J.M., Yuang, X.J., Hu, Q.H., Wang, S.Q.: A novel measure for evaluating classifiers. Expert Syst Appl **37**, 3799–3809 (2010)
41. Wold, S., Jonsson, J., Sjöström, M., Sandberg, M., Rännar, S.: DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta **277**, 239–253 (1993)

42. Wu, H., Wang, C., Gregory, K.J., Han, G.W., Cho, H.P., Xia, Y., Niswender, C.M., Katritch, V., Meiler, J., Cherezov, V., Conn, P.J., Stevens, R.C.: Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science **344**(6179), 58–64 (2014)
43. Wu, Z., Li, C.H., Zhu, J., Huang, J.: A semi-supervised SVM for manifold learning. In: Procs. of the 18^{th} International Conference on Pattern Recognition (ICPR) (2006)
44. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. Rep. CMU-CALD-02-107, Carnegie Mellon Univ. (PA) USA (2002)