Noname manuscript No.
(will be inserted by the editor)

Systematic analysis of primary sequence domain segments for the discrimination between class C GPCR subtypes

Caroline König · René Alquézar · Alfredo Vellido · Jesús Giraldo

Received: date / Accepted: date

Abstract G protein-coupled receptors (GPCRs) are a large and diverse super-family of eukaryotic cell membrane proteins that play an important physiological role as transmitters of extracellular signal. In this paper, we investigate Class C, a member of this super-family that has attracted much attention in pharmacology. The limited knowledge about the complete 3-D crystal structure of Class C receptors makes necessary the use of their primary amino acid sequences for analytical purposes. Here, we provide a systematic analysis of distinct receptor sequence segments with regard to their ability to differentiate between seven class C GPCR subtypes according to their topological location in the extracellular, transmembrane or intracellular domains. We build on the results from previous research that provided preliminary evidence of the potential use of separated domains of complete class C GPCR sequences as the basis for subtype classification. The use of the extracellular N-terminus domain alone was shown to result in a minor decrease in subtype discrimination in comparison to the complete sequence, despite discarding much of the sequence information. In this paper, we describe the use of Support Vector Machine-based classification models to evaluate the subtype discriminating capacity of the specific topological sequence segments.

Caroline König · René Alquézar · Alfredo Vellido
Univ. Politècnica de Catalunya, UPC BarcelonaTech, 08034, Barcelona, Spain E-mail: {ckonig, alquezar, avellido}@cs.upc.edu

Alfredo Vellido
Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès - Spain

Jesús Giraldo
Institut de Neurociències - Unitat de Bioestadìstica, Univ. Autònoma de Barcelona, 08193, Cerdanyola del Vallès , Spain
Network Biomedical Research Center on Mental Health (CIBERSAM)
E-mail: jesus.giraldo@uab.es

Keywords G-Protein coupled receptors · biocuration · pharmacoproteomics · machine learning · Support Vector Machines

1 Introduction

Research in biology and in the omics sciences in particular is quickly evolving towards increasingly data-driven models [1], in such a way that the creation, management and maintenance of publicly-available databases becomes paramount. Such databases are usually curated by specialized scientists that carry out some of these tasks in a process that has come to be known as *biocuration* [2,3].

In this study, we analyze G protein-coupled receptors (GPCRs), which are proteins of the eukaryotic cell membrane, where they have a function as transmitters of extracellular signals to the inside of the cell. This key physiological functionality has made them an interesting target for drugs in current pharmacological research [4,5].

While GPCRs embrace a heterogenous super-family of receptors, the current study specifically investigates its class C [6] (defined in accordance with the IUPHAR[1] convention). Class C receptors are relevant to the investigation of therapies for neurological diseases [7]. The functionality of proteins is known to be determined primarily by their 3-D structural configuration, which defines their ability for ligand binding. Despite recent discoveries of 3-D GPCR crystal structures [8,9], the knowledge about tertiary and quaternary structures is extremely limited in the case of class C GPCRs [10,11]. In the face of such lack of knowledge, information about their primary amino acid sequences (in this case widely known and available from publicly, web-accessible databases) is often a complementary approach for the investigation of receptor functionality.

The unambiguous identification and characterization of biological entities is an important challenge for biocurators. Addressing this challenge, the GPCRs analyzed in this paper are characterized according to subtype labels at different levels of organization. In previous research, we investigated the feasibility of discrimination between the seven defined class C GPCR subtypes using supervised machine learning classification approaches. These classifiers used different alignment-free sequence transformations, including transformations based on the physicochemical properties of the amino acids [12] and on short n-gram features [13]. These experiments showed a reasonably clear differentiation between subtypes, but also an evident upper threshold to classification accuracy, as well as some consistent misclassification patterns [14]. Note that these former experiments were based on the entire and unaligned primary sequences of the receptors.

The GPCRs are built according to different structural domains, including a seven-helix transmembrane (7TM) domain, as well as extracellular (N-

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

terminus) and intracellular (C-terminus) domains. For Class C GPCRs in particular, the extracellular domain of the receptor is large and consists of the Venus Flytrap (VFT) domain, where the endogenous agonists bind, and a cysteine-rich domain (CRD) connecting VFT and 7TM domains in many of their subtypes [15].

In previous research, we also analyzed whether the extracellular N-terminus domain of the sequences sufficed to distinguish between class C GPCR subtypes [16]. These experiments revealed that, although the classification models built using the isolated N-terminus domain did not achieve the subtype discrimination capabilities of the entire sequence in full, the observed reduction of classification performance was not too significant. Such results revealed the importance of investigating the subtype-discrimination capacity of separate structural domains further. Therefore, we build on these preliminary results in the current paper to provide a systematic analysis of such capacity for the complete set of different topological locations in the class C sequences (that is, in the extracellular, transmembrane and intracellular domains), including their combinations. The classification performances achieved with the sequence segments and the entire sequence are compared. These analyses should be understood as part of a GPCR characterization process for biocuration assistance.

Note also that the research reported in this paper is a direct extension of preliminary work reported in [17]. The remainder of the paper is structured as follows: the analyzed class C GPCR data are summarily described in section 2, followed, in section 3, by the description of the supervised classification strategy, the sequential data transformation methods, the criteria for partition of the sequence in domains and sub-domains and, lastly, the classification performance metrics. Next, the experimental results are reported and discussed. This is followed by a summary of conclusions.

2 Materials

The investigated data were gathered from a GPCR-specific curated information repository, the GPCRdb [18]. This repository was created in 1993 and it is now part of the GPCR Consortium[2], an industry-academia partnership and also part of the GLISTEN EU COST Action for the creation of a pan-European multidisciplinary research network.

GPCRdb characterizes the GPCR superfamily as the union of five major families (namely, A to E) based on functionality, ligand types and sequence similarities. As previously introduced, the current paper only investigates one of the GPCR families, namely class C, which has become popular in current pharmaco-proteomics research due to the selection of some of its members as drug development targets for human central nervous system therapies in areas such as pain, anxiety, or neurodegenerative disorders [6, 19].

Class C of GPCRs is further subdivided into seven subtypes: Metabotropic Glutamate (MG) receptors, Calcium sensing (CS), GABA-B (GB), Vomeronasal

[2] URL: http://gpcrconsortium.org

(VN), Pheromone (Ph), Odorant (Od) and Taste (Ta). The analyzed data set from version 11.3.4, released on March 2011, contains a total of 1,510 sequences from those seven subtypes. We limited our analyses to the subset of 1,252 sequences (approximately 83% of the total) that contain information of the complete 7-TM domain. The distribution of sequences per subtype is shown in Table 1, both for the original data set and for the subset comprising only sequences with complete 7-TM structure.

Table 1 Number of sequences per subtype available in the original data set and in the subset of sequences with complete 7-TM structure.

Class C subtype	**♯ sequ. original dataset**	**♯ sequ. compl. 7-TM structure**
MG	351	282
CS	48	45
GB	208	156
VN	344	293
Ph	392	323
Od	102	90
Ta	65	62
	1510	1252

3 Methods

3.1 Supervised classification techniques

Class C GPCR subtype discrimination is addressed here as a supervised classification problem in which class labels are the assignments of each of the sequences to one of the seven existing subtypes according to the information available in the database. The first phase of the experiments reported in Section 4 involved the use of several models for the classification of the alignment-free complete sequences. These results were used to choose which classifier was most adequate for the rest of analyses. The comparison was performed with a similar selection of classifiers to that used in a previous study [12] and included Naïve Bayes (NB) [20], Random Forest (RF) [21] and Support Vector Machine (SVM) [22] classifiers.

NB is a simple probabilistic classifier that applies Bayes' theorem under the assumption of attribute independence, creating a probabilistic model for class prediction. This is the baseline against which the other models' performance is compared.

RF is an ensemble based learning method [23] in which each of the elements of the ensemble is a decision tree [24] and the classification decision is the result of an internal voting system.

SVMs have been widely used in different variants in previous research for protein classification from their primary sequences; some examples include [25], [26], or [27]. Their underlying principles stem from statistical learning theory [22]. They map the D-dimensional vectors $\mathbf{x}_i, i = 1, \ldots, N$, where $\mathbf{x}_i \in R^D$ and N is the number of instances, into possibly higher-dimensional feature spaces through a function ϕ . The use of non-linear kernel functions allows SVMs to separate input data in higher dimensional spaces, in a way that would not be possible in the observed data space.

This is a multi-class classification problem and the *libsvm* implementation [28] of the SVM models was used. It entails a one-*vs*-one classification approach and the use of the nonlinear radial basis function (RBF) kernel: $K(x_i, x_j) = e^{(-\gamma||x_i - x_j||)}$. The use of the RBF kernel requires adjusting two parameters (the error penalty C and the γ parameter of the kernel) through grid search.

The classification results for all classifiers were obtained employing a 5-fold cross validation (5-CV) procedure with stratification for fold generation. This procedure was chosen due to limited data availability. In an ideal situation with abundant data available for analysis, the classification results would have more appropriately been evaluated using a test set of previously unseen data. In the scenario of our analyses, though, a test set could only be obtained through random sampling from the limited available data. This would be a fairly arbitrary procedure, and the evaluation of our experiments on the basis of such test would be a far less statistically reliable procedure than the 5-CV procedure proposed.

3.2 Alignment free sequence transformations

The use of the supervised classification models described in the previous section requires transforming the unaligned amino acid primary sequences of varying length into fixed-size matrix representations. In previous research [12], we used transformations based on the physicochemical properties of the amino acids that have been widely employed in proteomics research [29,26]. In the current study, we use transformations that have their foundations in the field of symbolic language analysis instead. They treat protein sequences as text from a 20 amino acid alphabet [30,31]. Here, short sequence fragments known as n-grams are understood as “words”. In [32], a successful application of class A GPCR classification using text classification methods was reported. This study used a discretization of n-gram features. In our research, we follow a similar strategy and calculate the relative frequency of occurrence of n-grams of sizes one and two, in which we call, in turn, *AA* and *Digram* transformations. In previous research, these n-gram-based transformations achieved relatively high classification performances in the analysis of the complete sequences of the original data set [13]. In this study, we go one step further and do not only calculate the frequencies of AA and Digram for all sequence segments (called *appended frequencies*), but also the *accumulated frequencies*, which are calculated as the occurrence of AA or Digram in all the segments under study, divided by the sum of the lengths of these segments.

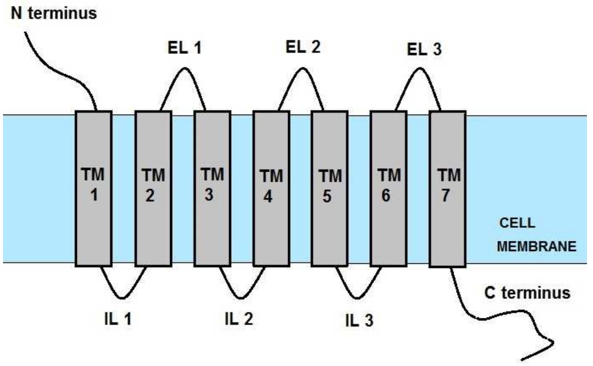

Fig. 1 Graphical representation of the common structure of GPCRs.

3.3 Segmentation of Structural Sequence Domains

Class C GPCRs have a common complex structure due to their transmembrane location: an extracellular domain comprising the N-terminus and 3 extracellular loops (EL), the 7TM, and an intracellular domain consisting of three intracellular loops (IL) and the C-terminus. Complete sequences, in accordance to this catalogue of structural domains, are partitioned into 15 segments. For this, we employed the *Phobius* transmembrane detection tool [33]. Table 2 summarizes some general information about the lengths (in number of amino acids) of these segments.

Table 2 Statistical information in reference to the length of the segments.

Segment	Mean	Min	Max	StDev
Complete Sequence	861.7	250	1,768	181
N-terminus	532.2	6	1,502	148.3
EL1	11.6	5	329	10.4
EL2	27	5	70	10.4
EL3	9	5	31	3.9
TM1	24.7	16	34	1.9
TM2	21.8	17	31	1.7
TM3	23.5	17	34	2.3
TM4	22.3	18	33	2.9
TM5	23.5	17	34	2.3
TM6	21.3	17	27	1.3
TM7	23.6	16	31	1.6
IL1	17	6	567	39.9
IL2	18.9	11	69	4.2
IL3	11.9	6	85	3.3
C-terminus	73	0	1,044	113

3.4 Performance Metrics

Several metrics were used to evaluate the classification models in the reported experiments. First, at the subtype level, Precision (Prec), Recall (Rec) and Matthews Correlation Coefficient (MCC) were used to evaluate the binary classifier for each subtype (See Table 3). The MCC is considered to be a more complete figure of merit as it encompasses all elements of the confusion matrix [34] and is most robust for unbalanced data sets [35]. Being calculated as the correlation coefficient between the observed and the predicted classification, it ranges from -1 (for complete misclassification) to 1 (for perfect classification). *Prec* describes the correctness of the predicted positives. It is therefore a measure of quality, because it measures to which degree all predicted positives are true. Finally, *Rec* measures the rate of discovery of true positives. It is therefore a measure of completeness, because it measures to which degree all true positives are detected.

At the global level, the quality of the multi-class models was evaluated using classification accuracy, which is the proportion of correctly classified receptors, and multi-class MCC (See Table 4).

Table 3 Performance measures for binary classifiers: These metrics build on the notion of true and false predictions in binary classification with "positive" and "negative" classes [36]. True positives (tp) and true negatives (tn) are correctly classified cases of the positive and negative classes respectively. Correspondingly, false positives (fp) an false negatives (fn) are misclassified cases of the negative and positive classes on the other hand.

Measure	Formula	Meaning
Precision	$\frac{tp}{tp+fp}$ (1)	Measure of quality
Recall	$\frac{tp}{tp+fn}$ (2)	Measure of completeness
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$ (3)	Measure of correctness
MCC	$\frac{tp*tn-fp*fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$ (4)	Correlation coefficient

4 Experiments

4.1 Experimental settings

Details of the choice of parameter values for the classifiers are provided here. The RF was run with 10 trees of unlimited lengths. The SVM used a one-*vs*-one classification approach. An RBF kernel ($K(x_i, x_j) = e^{(-\gamma||x_i - x_j||)}$) was employed. It requires adjusting two parameters (the error penalty C and the γ parameter of the kernel) through grid search. Such grid search was carried

Table 4 Performance measures for multi-class classifiers: tp_i, tn_i, fp_i and fn_i stand for tp, tn, fp and fn for class i [36]. The multi-class MCC involves all the entries of the confusion matrix $C_{K\times K}$ of all K classes.[37]. The ij^{th} entry (c_{ij}) describes the number of instances of the true class i assigned to the class j by the classifier.

Measure	**Formula**
Accuracy	$\frac{\sum_{i=1}^{K} \frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{K} (5)$
MCC	$\frac{\sum\limits_{k,l,m=1}^{K} C_{kk}C_{ml}-C_{lk}C_{km}}{\sqrt{\sum\limits_{k=1}^{K}[(\sum\limits_{l=1}^{K}C_{lk})(\sum\limits_{f,g=1 f\neq k}^{K}C_{gf})]}\sqrt{\sum\limits_{k=1}^{K}[(\sum\limits_{l=1}^{K}C_{kl})(\sum\limits_{f,g=1 f\neq k}^{K}C_{fg})]}} (6)$

out with C values ranging from 1 to 5 and γ values from 2^{-16} to 2^3. Table 5 compiles the best parameter settings found through the grid search and used to implement the SVM models in the reported experiments for the different sequence segments.

4.2 Classifier performance comparison with complete sequences

As stated in Section 3.1, we first used several supervised models for the classification of the complete sequences in order to select the most adequate classifier. Table 6 shows the classification performance for the different classifiers (best results are highlighted in bold). The results reveal that, when compared with RF and NB, the best classification performance was achieved by SVM, both for the AA and Digram transformations. For this reason, SVM was used in the subsequent experiments.

Table 7 details the underlying subtype classification results by reporting the per-subtype *Prec*, *Rec* and MCC obtained by the SVM classifier from the Digram data. The best results were obtained for subtypes MG, CS, GB and Ta, while the results for subtypes VN, Ph and Od were less accurate. The Od subtype, in particular, yielded very poor results. Overall, these results are, in any case, in line with those obtained in previous research [12], [13].

4.3 Experiments with topological sequence segments

The experiments reported in this Section concern the SVM classification models built for the different topological segments and their combinations. Table 8 shows the classification results for the segments in the extracellular domain. Table 9 corresponds to the 7TM, and Table 10, in turn, to the four intracellular regions IL1, IL2, IL3 and the C-terminus. Table 11, on the other hand,

Table 5 Parameter settings for the SVM experiments: For each experiment on a sequence segment, the error penalty C and the γ parameter of the RBF kernel are reported.

	AA		**Digram**	
Segment	**C**	γ	**C**	γ
Complete Sequence	4	2^{-4}	5	2^{-9}
N-Terminus	5	2^{-4}	5	2^{-9}
EL1	5	2^{-4}	4	2^{-11}
EL2	4	2^{-4}	4	2^{-10}
EL3	5	2^{-4}	5	2^{-11}
All EL appended freq.	5	2^{-6}	5	2^{-12}
All EL accum. freq.	5	2^{-4}	5	2^{-11}
(Nterm + EL) app. freq.	5	2^{-7}	5	2^{-12}
(Nterm + EL) accum. freq.	5	2^{-10}	5	2^{-10}
TM1	5	2^{-4}	5	2^{-9}
TM2	5	2^{-4}	4	2^{-10}
TM3	5	2^{-4}	5	2^{-9}
TM4	5	2^{-4}	4	2^{-10}
TM5	4	2^{-4}	4	2^{-11}
TM6	5	2^{-3}	5	2^{-10}
TM7	5	2^{-3}	5	2^{-9}
TM append. frequency	4	2^{-9}	5	2^{-13}
TM accum. frequency	4	2^{-4}	5	2^{-11}
IL1	4	2^{-4}	5	2^{-9}
IL2	4	2^{-4}	5	2^{-11}
IL3	5	2^{-4}	5	2^{-10}
C-terminus	5	2^{-5}	5	2^{-12}
(IL+ C-term.) append. freq.	4	2^{-8}	4	2^{-13}
(IL + C-term.) accum. freq.	4	2^{-4}	4	2^{-10}
(7TM+NT) append. freq.	5	2^{-9}	5	2^{-13}
(7TM+NT) accum. freq.	4	2^{-4}	5	2^{-9}
(15 Segments) append. freq.	5	2^{-9}	5	2^{-14}
(15 Segments) accum. freq.	4	2^{-4}	5	2^{-9}

summarizes the classification results for the N-terminus combined with the 7TM region. Finally, Table 12 shows the classification results for all 15 segments of the complete sequence. Each table displays the name of the segments considered in the experiment, the size of the feature set and the classification performance as measured by MCC and accuracy.

Table 6 Classification results for the complete sequences according to classifier.

	AA			**Digram**		
Classifier	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
NB	20	0.625	0.703	400	0.792	0.834
RF	20	0.657	0.726	400	0.656	0.724
SVM	20	**0.838**	**0.873**	400	**0.917**	**0.934**

Table 7 Subtype classification results achieved with SVM from the Digram data transformation.

Class C subtype	**MCC**	**Prec**	**Recall**
MG	0.946	0.975	0.949
CS	0.951	0.911	0.927
GB	1.0	0.981	0.989
VN	0.936	0.932	0.913
Ph	0.897	0.922	0.875
Od	0.810	0.675	0.722
Ta	1.0	1.0	1.0

Table 8 Classification results for the extracellular segments.

	AA			**Digram**		
Segments	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
N-terminus	20	0.792	0.835	400	0.901	0.920
EL1	20	0.802	0.842	390	0.786	0.831
EL2	20	0.798	0.839	386	0.825	0.861
EL3	20	0.779	0.825	327	0.769	0.816
All EL appended freq.	60	0.839	0.873	1103	0.873	0.880
All EL accum. freq.	20	0.804	0.845	398	0.844	0.875
(Nterm + EL) app. freq.	80	**0.878**	**0.904**	1502	0.889	0.912
(Nterm + EL) accum. freq.	20	0.8089	0.849	400	**0.901**	**0.921**

4.4 Subtype specific classification results of topological sequence segments

In this Section, we extend the previous analysis by reporting the per-subtype classification results for the sequence segments (and its combinations) found to perform best as detailed in the previous sub-section. Table 13 shows these subtype classification results for the concatenation of all 15 segments (MCC=0.914, Accu=0.932), the N-terminus (MCC=0.901, Accu=0.92), the extracellular segments, i.e. N-terminus + EL (MCC=0.901, Accu=0.921), the N-terminus + 7TM (MCC=0.909, Accu =0.928), the 7TM segments (MCC=0.873, Accu=0.902) and the intracellular segments, i.e. IL+C-terminus (MCC=0.88, Accu=0.906). For each data set, the types of transformation and frequency are reported in the table.

Table 9 Classification results for the transmembrane segments.

	AA			**Digram**		
Segments	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
TM1	20	0.741	0.794	321	0.778	0.823
TM2	20	0.809	0.850	298	0.806	0.847
TM3	20	0.829	0.866	290	0.846	0.878
TM4	20	0.776	0.822	320	0.822	0.860
TM5	20	0.8181	0.859	293	0.817	0.856
TM6	20	0.794	0.836	262	0.81	0.848
TM7	20	0.755	0.808	281	0.801	0.843
TM append. frequency	140	**0.873**	**0.902**	2066	**0.871**	**0.900**
TM accum. frequency	20	0.847	0.879	384	0.864	0.894

Table 10 Classification results for the intracellular segments.

	AA			**Digram**		
Segments	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
IL1	20	0.777	0.825	398	0.739	0.795
IL2	20	0.815	0.853	388	0.837	0.872
IL3	20	0.817	0.857	304	0.789	0.834
C-terminus	20	0.74	0.793	400	0.753	0.805
(IL+ C-term.) append. freq.	80	**0.880**	**0.906**	1490	**0.874**	**0.895**
(IL + C-term.) accum. freq.	20	0.795	0.837	400	0.854	0.885

Table 11 Classification results for the N-terminus concatenated with the 7TM regions.

	AA			**Digram**		
Segments	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
appended frequency	160	**0.897**	**0.919**	2467	0.889	0.915
accumulated frequency	20	0.830	0.866	400	**0.909**	**0.928**

Table 12 Classification results for the concatenation of all 15 segments.

	AA			**Digram**		
Segments	**Size**	**MCC**	**Accu**	**Size**	**MCC**	**Accu**
appended frequency	300	**0.905**	**0.925**	5058	0.888	0.911
accumulated frequency	20	0.840	0.875	400	**0.914**	**0.932**

4.5 Discussion

The results of our experiments for the sequence segments and their combinations reveal a neat pattern of progressive deterioration of classification perfor-

Table 13 Subtype classification results for different sequence segments and transformation as described in the header. MCC best results over segment choices for each subtype shown in bold.

	Concaten. 15 segments (Digram accum. frequ.)			**N-terminus (Digram)**		
Class C subtype	**Prec**	**Recall**	**MCC**	**Prec**	**Recall**	**MCC**
MG	0.947	0.982	0.953	0.962	0.951	0.943
CS	0.951	0.933	0.939	1.0	0.911	**0.952**
GB	1.0	0.981	**0.989**	1.0	0.968	0.982
VN	0.939	0.929	**0.913**	0.919	0.918	0.893
Ph	0.894	0.922	**0.875**	0.88	0.916	0.859
Od	0.853	0.688	0.722	0.751	0.725	0.712
Ta	1.0	1.0	**1.0**	1.0	0.967	0.982
	N-terminus + EL (Digram accum. frequ.)			**N-terminus + 7TM (Digram app. frequ.)**		
Class C subtype	**Prec**	**Recall**	**MCC**	**Prec**	**Recall**	**MCC**
MG	0.968	0.961	**0.954**	0.922	0.986	0.939
CS	0.980	0.889	0.928	0.933	0.889	0.906
GB	0.993	0.974	0.982	1.0	0.962	0.978
VN	0.912	0.908	0.882	0.917	0.894	0.877
Ph	0.865	0.919	0.851	0.878	0.904	0.851
Od	0.752	0.688	0.70	0.873	0.70	**0.764**
Ta	1.0	0.9372	0.966	1.0	0.983	0.991
	Transmembrane (AA app. frequ.)			**IL + C-terminus (AA app. frequ.)**		
Class C subtype	**Prec**	**Recall**	**MCC**	**Prec**	**Recall**	**MCC**
MG	0.926	0.986	0.94	0.953	0.975	0.953
CS	0.899	0.933	0.912	0.939	0.889	0.908
GB	1.0	0.968	0.982	0.982	0.9811	0.978
VN	0.89	0.894	0.859	0.879	0.918	0.867
Ph	0.873	0.883	0.833	0.884	0.898	0.85
Od	0.667	0.5	0.549	0.75	0.513	0.592
Ta	1.0	0.954	0.974	0.986	0.985	0.984

mance as we remove more parts of the sequence. It is nevertheless remarkable that the classification performance never decreases below 0.75 (neither in MCC nor in accuracy), even for very short segments, and seldom below 0.8. These results thus reveal a notable conservation of the subtype discriminability capabilities throughout the sequence.

For the entire sequence, the best classification was found for the Digram representation, which yielded an MCC of 0.917 and an accuracy of 0.934, which is a similar performance to that of its partition into 15 segments, with an MCC of 0.914 and accuracy of 0.932 for the Digram representation and accumulated frequencies (see Table 12). Note that by using the segmentation of the entire sequence, the classification results of the AA transformation were clearly improved, as the entire sequence achieved an MCC of 0.838 and accuracy of 0.873 using 20 attributes, while the appended frequency of the 15 segments yielded an MCC of 0.905 and accuracy of 0.925 using 300 attributes. This result validates the approach consisting on the combination of complete sequence segmentation and use of appended frequencies.

The analysis of the extracellular segments revealed that the classification performance using the N-terminus alone, or combined with the extracellular loops (see Table 8) decreases just over one percentage point, both in MCC and accuracy, when compared to the performance of the complete sequence and the Digram transformation. The combination of the N-terminus with the 7TM provided similar classification performances as well (see Table 11).

The experiments corresponding to the extracellular loops, transmembrane and intracellular segments show less accurate classification compared to those of the entire sequence or the N-terminus. At large, the combination of topologically-alike segments improves the classification results obtained using single segments (with the aforementioned exception of the N-terminus). Note as well that some very short sequence segments such as IL2, EL2, TM3 and TM4 (several of them comprising no more than 2.2% of the sequence) barely drop more than 6% in classification performance when compared with the best results. This is a somewhat surprising outcome that indicates that subtype differences are deeply embedded even in such small segments.

Regarding the type of transformation, Digram yielded the best results in general, with two interesting exceptions, namely for the 7TM regions and for the IL + C-terminus for the appended frequencies. The comparison between the use of appended frequencies and accumulated frequencies reveals that the former achieve better results with the AA transformation, whereas the latter perform better with Digram.

The per-subtype classification results reported in Table 13 are consistent with the results obtained for the entire sequence (See Table 7), as all data sets achieve better results for subtypes MG, CS, GB and Ta, while subtypes Vn, Ph and Od show the worst performance.

A detailed comparison of the subtype classification results shows that the entire sequence and the concatenation of its 15 segments provide the best performance for subtypes GB (MCC=0.989), Vn (MCC=0.913) , Ph (MCC=0.875) and Ta (MCC=1.0). In turn, the best results for MG were found for the entire sequence (MCC=0.953) and N-terminus + EL (MCC=0.954). For subtype CS, the best result was found for the N-terminus (MCC=0.952), while Od performed best for the combination of N-terminus + 7TM (MCC=0.764).

The overall good behavior of those sequences including the N-terminus is consistent with the fact that this domain contains the binding sites for the endogenous ligands responsible for the activation of class C GPCRs. Thus, the AAs present in the N-terminus determine the recognition of glutamate in MG receptors, GABA in GABA-B receptor, Ca^{2+} in CS receptor, *etcetera*. As a consequence, the N-terminus conveys most of the discriminatory elements for the classification of GPCR class C sequences. However, GPCRs and particularly their class C are complex entities both at the structural and functional levels. GPCRs are allosteric machines and the binding sites for the transducer G proteins are located at the intracellular part of the receptors far away from the ligand binding sites. This may explain the contribution of the ILs in our analysis. Moreover, the 7TM domain needs to be activated for G protein binding and then contributions of this structural domain for sequence classification are expected. Inasmuch as allosteric cooperativity interactions between the 7TM and VFT domains have been also reported [38], it is expected that segments including these domains appear in our study. Finally, ELs are involved in 7TM domain flexibility and cooperativity interactions, which justify their putative discriminative power.

As a whole, these results provide a complete and detailed landscape of the relative capabilities of different sequence segments (from different GPCR domains and in different combinations) in the task of discriminating between the seven subtypes of class C GPCRs. This detailed landscape should help database biocurators in their tasks.

5 Conclusions

The research reported in this paper is based on the web-accessible and public protein databases of the GPCRdb consortium. Biocurators of this type of databases face the non-trivial challenge of unambiguously identifying and characterizing GPCRs. In this database, receptors are characterized according to subtype labels at different levels of organization. In previous research, the analysis of the N-terminus of the extracellular domain provided some preliminary evidence of the potential use of individual domains of complete class C GPCR sequences as the foundation for subtype classification.

We have performed a systematic analysis of the classification performance of each of the individual sequence segments in which the sequence can be divided in each of its structural domains, as well as the performance of several of their combinations. The experimental results revealed that none of them reached the classification performance of the complete sequence or the concatenation of its 15 constituent segments. However, the segments of the extracellular domain, the N-terminus in combination with the 7TM and, to some degree, the intracellular domain, have all performed almost as well as the complete sequence.

The identification of the most discriminative segments should be the starting point for future work focusing on these separate regions. Such future research should involve feature selection starting from these segments as a way to discover specific motifs with subtype discriminative capabilities and potential functional roles.

Acknowledgments

This research was partly financed by the Spanish SAF2014-58396-R and MINECO TIN2016-79576-R projects.

References

1. Leonelli S (2016) Data-centric biology: a philosophical study. University of Chicago Press, Chicago (IL), U.S.A.
2. Howe D et al (2008) Big data: The future of biocuration. Nature 455(7209):47-50.
3. Baxevanis AD, Bateman A (2006) The importance of biological databases in biological discovery. Curr Protoc Bioinform 50:1.1.1-1.1.8.
4. Rask-Andersen M, Almén MS, Schiöth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10(8):579.
5. Santos R et al (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16(1):19-34.
6. Leach K, Gregory KJ (2016) Molecular insights into allosteric modulation of Class CG protein-coupled receptors. Pharmacol Res 116:105–118.
7. Kniazeff J, Prézeau L, Rondard P, Pin JP, Goudet C (2011) Dimers and beyond: The functional puzzles of class C GPCRs. Pharmacol Therapeut 130(1):9-25.
8. Alexander SP et al (2015) The concise guide to PHARMACOLOGY 2015/16: G protein coupled receptors. Brit J Pharmacol 172(24):5744-5869.
9. Cooke RM, Brown AJ, Marshall FH, Mason JS (2015) Structures of G protein-coupled receptors reveal new opportunities for drug discovery. Drug Discov Today 20(11):1355-1364.
10. Wu H et al (2014) Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science 344(6179):58-64.
11. Dore AS et al (2014) Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. Nature 511(7511):557.
12. König C, Cruz-Barbosa R, Alquézar R, Vellido A (2013) SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences. In: Petrosino A, Maddalena L, Pala P (eds) New Trends in Image Analysis and Processing. Proceedings of the International Conference on Image Analysis and Processing (ICIAP 2013). Lecture Notes in Computer Science, vol 8158. Springer, Berlin, Heidelberg, pp 336-343.
13. König C, Alquézar R, Vellido A, Giraldo J (2014) Reducing the n-gram feature space of class C GPCRs to subtype-discriminating patterns. J Integr Bioinform 11(3):99-115.
14. König C, Cárdenas MI, Giraldo J, Alquézar R, Vellido A (2015) Label noise in subtype discrimination of class CG protein-coupled receptors: A systematic approach to the analysis of classification errors. BMC Bioinformatics 16(1):314.
15. Pin JP, Galvez T, Prézeau L (2003) Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. Pharmacol Therapeut 98(3):325-354.
16. König C, Alquézar R, Vellido A, Giraldo J (2015) The extracellular N-terminal domain suffices to discriminate class C G Protein-Coupled Receptor subtypes from n-grams of their sequences. In: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1-7.
17. König C, Alquézar R, Vellido A, Giraldo J (2017) Topological sequence segments discriminate between class C GPCR subtypes. In Fdez-Riverola F, Mohamad M, Rocha M, De Paz J, Pinto T (eds) Proceeings of the 11^{th} International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB). Advances in Intelligent Systems and Computing 616. Springer, pp 164-172.
18. Isberg V et al (2016) GPCRdb: an information system for G protein-coupled receptors. Nucleic Acids Res 44:D356-64.
19. Pin JP, Bettler B (2016) Organization and functions of mGlu and GABAB receptor complexes. Nature 540(7631):60-68.
20. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11^{th} Conference on Uncertainty in Artificial Intelligence (UAI). Morgan Kaufmann Publishers Inc. pp 338-345.
21. Breiman L (2001) Random forests. Mach Learn 45(1):5-32.
22. Vapnik VN (1998) Statistical learning theory (Vol 1) Wiley, New York, U.S.A.
23. Dietterich TG (2000) Ensemble methods in machine learning. In: Kittler J, Roli F (eds) Proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000) Cagliari, Italy. Lecture Notes in Computer Science 1857, Springer, pp 1-15.

24. Quinlan JR (2014) C4.5: programs for machine learning. Elsevier.
25. Karchin R, Karplus K, Haussler D (2002) Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18(1):147-159.
26. Liu B, Wang X, Chen Q, Dong Q, Lan X (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. PloS ONE 7(9):e46633.
27. Meng FR, You ZH, Chen X, Zhou Y, An JY (2017) Prediction of DrugTarget Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures. Molecules 22(7):1119.
28. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM T Intel Syst Tec 2(3):27.
29. Opiyo SO, Moriyama EN (2007) Protein family classification with partial least squares. J Proteome Res 6(2):846-853.
30. Caragea C, Silvescu A, Mitra P (2011) Protein sequence classification using feature hashing. In: Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp 538-543.
31. Mhamdi F, Elloumi M, Rakotomalala R (2004) Textmining, feature selection and datamining for proteins classification. In: Proceedings of the 2004 International Conference on Information and Communication Technologies: From Theory to Applications. IEEE, pp 457-458.
32. Cheng BYM, Carbonell JG, Klein-Seetharaman J (2005) Protein classification based on text document classification techniques. Proteins 58(4):955-970.
33. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783-795.
34. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA-Protein Struct 405(2):442-451.
35. Bekkar M, Djemaa HK, Alitouche TA (2013) Evaluation measures for models assessment over imbalanced data sets. Journal Of Information Engineering and Applications 3(10).
36. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inform Process Manag 45(4):427-437.
37. Jurman G, Riccadonna S, Furlanello C (2012) A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE 7(8):e41882.
38. Rovira X et al (2015) Overlapping binding sites drive allosteric agonism and positive cooperativity in type 4 metabotropic glutamate receptors. FASEB J 29(1):116-130.