The Extracellular N-terminal Domain Suffices to Discriminate Class C G Protein-Coupled Receptor Subtypes from *n*-Grams of their Sequences

Caroline König,
René Alquézar
and Alfredo Vellido
Computer Science Department
Universitat Politècnica de Catalunya (UPC BarcelonaTech)
Barcelona, 08034, Spain
Email: {ckonig, alquezar, avellido}@cs.upc.edu

Jesús Giraldo
Institut de Neurociències - Unitat de Bioestadìstica,
Universitat Autònoma de Barcelona (UAB),
Cerdanyola del Vallès, 08193, Spain
Email: jesus.giraldo@uab.es

***Abstract*—The investigation of protein functionality often relies on the knowledge of crystal 3-D structure. This structure is not always known or easily unravelled, which is the case of eukaryotic cell membrane proteins such as G Protein-Coupled Receptors (GPCRs) and specially of those of class C, which are the target of the current study. In the absence of information about tertiary or quaternary structures, functionality can be investigated from the primary structure, that is, from the amino acid sequence. In previous research, we found that the different subtypes of class C GPCRs could be discriminated with a high level of accuracy from the n-gram transformation of their complete primary sequences, using a method that combined two-stage feature selection with kernel classifiers. This study aims at discovering whether subunits of the complete sequence retain such discrimination capabilities. We report experiments that show that the extracellular N-terminal domain of the receptor suffices to retain the classification accuracy of the complete sequence and that it does so using a reduced selection of n-grams whose length of up to five amino acids opens up an avenue for class C GPCR signature motif discovery.**

I. Introduction

GPCRs are eukaryotic cell membrane proteins with several biologically relevant roles due to their ability to transmit extracellular signals, activating intra-cellular signal transduction pathways. Their biomedical relevance comes from the fact that they have become, at present, the target of up to 36% of all drugs approved by the US Food and Drug Administration during the past three decades [1], making them of obvious interest in pharmacology.

This study involves class C, a family of GPCRs that has become the center of much investigation in new therapies for (amongst others) neurological pathologies [2]. The determination of the 3-D structure of full-length class C GPCRs has proven to be a particularly challenging task and it was only in 2014 that the first structures of the seven transmembrane (7TM) domains of two class C receptors were discovered [3], [4]. For this reason, the definition of direct quantitative strategies for the analysis of the primary sequence of class C GPCRs is a relevant problem in bioinformatics. Some recent approaches have even suggested the possibility of generating 3-D models of receptors *de novo* by defining restraints on structural proximity of residue pairs through amino acid evolutionary covariation analysis [5].

We are specifically interested in finding out which characteristics of the receptor sequence facilitate the discrimination between the known subtypes of class C. For that, we can start from sequence transformations to align the sequences, which is a common strategy in the field, or from alignment-free transformations of different types, which is our strategy of choice to allow retaining as much of the available information as possible. This GPCR discrimination task has been addressed in the past using supervised [6], semi-supervised [7] and even fully unsupervised [8], [9] modelling approaches.

The current paper focuses on a specific type of unaligned sequence transformation, namely *n*-grams, which are specific amino acid subsequences of length n that can also be understood as sequence *motifs*. In previous research, we found that the different subtypes of class C GPCRs could be discriminated with a high level of accuracy from the n-gram transformation of their complete primary sequences, using a method that combined two-stage feature selection with kernel classifiers.

The GPCR, though, is a highly structured protein with very clearly differentiated domains, including the 7TM domain, with seven transmembrane α-helices and three extra-cellular and three intra-cellular loops connecting them; the extracellular N-terminus and an intracellular C-terminus [10]. This study aims at discovering whether subunits of the complete sequence retain the subtype-discrimination capabilities of the complete sequence. For that, we extended the combination of feature selection and classification method to lengthier n-grams of class C GPCR sequence parts.

In the following, we report experiments that show that the extracellular N-terminal domain of class C GPCRs suffices to retain the classification accuracy of the complete sequence. Bearing in mind that the objectives of this study go beyond discriminability assessment to also address the discovery of those subsequences that are the most responsible for achieving class C GPCR subtype discrimination, the selection of the latter from n-grams of length of up to five amino acids opens up an avenue for GPCR signature motif discovery.

II. Materials and Methods

The raw class C GPCR sequences under study, which belong to an open access curated database, are described next. This is followed by a brief introduction to the n-gram technique for the transformation of these symbolic primary amino acid sequences into real-valued features. Next, we describe some basic techniques to reduce the dimensionality of the resulting features and the subsequent method for classification.

A. The Class C GPCR Database

As described in the introduction, GPCRs are a large family of integral cell membrane receptor proteins, mediating signal transmission from the extracellular to the intracellular domains and thus prompting cellular response. The data analyzed in our experiments was extracted from GPCRDB [11], which is a publicly accessible, curated database including heterogeneous but thoroughly structured information about GPCRs at large. GPCRDB divides this superfamily into several major classes (namely, A to C, plus Vomeronasal receptors, cAMP receptors and Taste receptors T2R) based on ligand types, functions, and sequence similarity. Our investigations concerns only class C, which is of medical interest for its increasingly important role in the development of new therapies in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics.

The originally sampled data set (from version 11.3.4 as of March 2011) comprised 1,510 class C GPCR sequences, sub-classified into seven subtypes: 351 metabotropic glutamate (mG), 48 calcium sensing (CS), 208 GABA-B (GB), 344 vomeronasal (VN), 392 pheromone (Ph), 102 odorant (Od) and 65 taste (Ta). Given that this study focuses on the investigation of how the information in different sequence parts and domains could be independently used to discriminate between these class C subtypes, these sequences were filtered to remove all those from which no 7TM domain information was available, using a transmembrane topology prediction tool (Phobius [12]). As a result, the final data set included 1,252 sequences: 282 mG, 45 CS, 156 GB, 293 VN, 333 Ph, 80 Od and 63 Ta.

At a structural level, class C GPCRs have seven transmembrane helices; an extracellular domain, the N-terminus; and the C-terminus. The N-terminus contains the Venus Flytrap (VFT), which is a part of special relevance as it contains the recognition site where the endogenous agonist binds [13]. Besides the VFT, the N-terminus also contains a cystein-rich domain (CRD), which is present in almost all classes (with the exception of subtype GB [14]) and connects the VFT to the first transmembrane helix. Due to the importance of the N-terminus for the protein activation, this part of the receptor was isolated in all sequences and saved for analysis.

B. The n-Gram GPCR Sequence Transformation

The n-grams are, in general, contiguous specific amino acid subsequences of length n. This concept is commonly used in text and natural language processing, but it has also been used in the context of protein analysis [15], [16], [17], [18], even by direct transposition of text classification methods for the classification of GPCRs [19].

In most cases, GPCR classification from n-grams has been accomplished for $n = 2$, that is, for digrams, also known as dipeptide composition [20]. This choice is not made on biological grounds, but, usually for the sake of analytical and even computational tractability: for a 20 amino acid alphabet there are 400 digram and 8,000 trigram possible combinations; the large feature space dimensionality of longer n-grams would compromise such tractability.

In this study, we use the relative frequencies of occurrence of the n-grams, which are continuous, real-valued variables. Each feature corresponds to the measurement of an n-gram relative frequency in the sequence. The n-grams can be understood as receptor sequence deterministic motifs. They can be either *contiguous*, if there are no gaps between the amino acids that constitute de n-gram, or *gap* motifs, if such gaps (filled with any amino acid of the analyzed alphabet, known as a *wild-card*) are allowed.

For the experiments reported in the following sections, we considered a combination of *contiguous* and *rigid gap* motifs or n-grams of up to five amino acids. The latter are *rigid* in the sense that there is a fixed number of gaps in between the n-gram amino acids [21].

As stated in [22], the sequence space of proteins is redundant and, therefore, the use of amino acid grouping strategies based on physicochemical similarity is advisable, in order to decrease the granularity of the feature space and therefore alleviate the computational burden and potential inconsistencies involved in the analysis of very high dimensional datasets. In this study, and besides the 20-amino acid alphabet (as displayed in Table I), we used the Sezerman (SEZ) alphabet to create the n-grams. It includes 11 groups of amino acids of length 1 to 4, namely: [A], [W], [C], [P], [DE], [QN], [ST], [GT], [YF], [RKH] and [IVLM]. The performance of SEZ has been positively evaluated in [23] in the classification of GPCRs into their major classes.

TABLE I. List of the 20 amino acids in the AA alphabet.

Amino acid name	Symbol	Amino acid name	Symbol
Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Methionine	M
Aspartate	D	Phenylalanine	F
Cysteine	C	Proline	P
Glutamate	E	Serine	S
Glutamine	Q	Threonine	T
Glycine	G	Tryptophan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V

C. Two-stage Feature Selection with SVM-Based Classification

The n-gram transformation of the GPCR sequences is likely to yield many features that are not relevant in terms of class C subtype discrimination. These irrelevant n-gram frequencies may have a negative impact (or at best a negligible one) in this classification process and, therefore, we aim to investigate whether a subset of relevant frequencies retains the subtype classification capabilities. Indirectly, we also want to investigate the selected n-grams for hitherto unknown signature motifs. One criterion of significance is their statistical or informative performance, which is related to biological significance [24]. It has been suggested that motif over-representation maybe due to evolutionary preservation of sequence segments, suggesting their structural and functional roles [21]. This should make n-gram frequencies informative measures in terms of functionality exploration.

Two complementary feature selection approaches were used in this study, following a two-step strategy previously tested for complete sequences in [6]: two-sample t-tests among the class C GPCR subtypes for feature filtering followed by sequential forward feature selection with a Support Vector Machine (SVM) classifier. In a similar approach, a t-test with SVM classifiers was used in [25].

The two-sample t-tests in the first step are a somehow crude evaluation of the discriminating power of individual features. This univariate statistical test analyzes whether there are foundations to consider two independent samples as coming from populations (normal distributions) with unequal means by analyzing the values of the given feature. In our case, we used t-tests with 0.01 significance level.

Given the multi-class nature of the problem, t-tests were run for the 21 two-class combinations of the 7 class C subtypes. The two-sample t-test values were calculated at this detail because the multi-class SVM implementation internally performs class-vs-class comparisons. Therefore, the t-test analyzes the data in each binary classifier, making the ranking of the features possible according to their overall significance (i.e., according to how many binary classifiers a feature is significant in).

The second dimensionality reduction step starts from the selection performed through the t-tests and involves a sequential forward selection algorithm [26] operating in wrapper mode, that is, each feature subset is tested as part of the classification procedure [27]. It starts from an empty candidate feature set and adds, in each iteration, the feature which most improves the SVM classifier accuracy in a 5-fold cross-validation (5-CV). Note that an ideal or *signature* motif, and thus a candidate for potential structural and functional roles, has been described to be one "that matches all the sequences of the target family and no other sequence outside this family" [21].

SVMs have often been the type of classifier of choice for the analysis of GPCRs from different types of transformations of their primary sequences (see, for instance, [20], [28]). As previously mentioned, SVMs were used in the current study for the supervised classification of the alignment-free amino acid sequence transformations into the seven subtypes of class C GPCRs described in Section II-A. Given the multi-class problem setting, the LibSVM implementation [29] was employed.

III. EXPERIMENTS

A. Results

1) Comparative Classification of the N-terminal Domain: We built our SVM-based classification models using the n-grams from the N-terminus for each of the two alphabets under consideration: the complete amino acid alphabet (AA) and SEZ. In previous research [6], we analyzed the amino acid frequencies (1-grams) and digrams from the complete sequence. For comparative purposes, Table II shows the classification results, as measured by accuracy, for each alphabet using n-grams of length 1 and 2, for both approaches.

TABLE II. N-GRAM COMPARATIVE CLASSIFICATION RESULTS FOR THE N-TERMINUS AND THE COMPLETE SEQUENCE, WHERE D IS THE SIZE OF A FEATURE SET AND ACC STANDS FOR CLASSIFICATION ACCURACY (RATIO OF CORRECTLY CLASSIFIED SEQUENCES).

	N-terminus				**Complete Sequence**			
	AA		**SEZ**		**AA**		**SEZ**	
N-GRAM	D	**ACC**	D	**ACC**	D	**ACC**	D	**ACC**
1-gram	20	0.84	11	0.78	20	0.87	11	0.82
2-gram	400	0.92	121	0.91	400	0.93	121	0.93

2) t-Test Filtering and Sequential Forward Feature Selection: Supported by the previous results, we then proceeded to apply the proposed two-step feature selection with SVM-based classification to the combination of *contiguous* and *rigid gap motifs* (n-grams) of lengths three to five. Their very high dimensionality makes them difficult to use with SVMs and, therefore, t-test filtering was used to generate a first crude ranking of features.

Table III shows this ranking according to the overall significance of the attributes. This means that, for each alphabet, we counted how many features were significant (column D) in at least 20,19,18,17, etc., class-vs-class tests (bear in mind that there are 21 possible combinations of the 7 class C subtypes). The ACC values shown for each subset are the classification accuracies of the SVM built from each feature set. For comparison, we also show, besides the results obtained with *rigid gap motifs* for the N-terminus, the corresponding results from previous research [6] in which *continuous* (n-grams) of lengths one to three were calculated from the complete sequence.

TABLE III. N-GRAM COMPARATIVE CLASSIFICATION RESULTS AFTER T-TEST, WHERE D IS THE SIZE OF A FEATURE SET AND ACC STANDS FOR CLASSIFICATION ACCURACY (RATIO OF CORRECTLY CLASSIFIED SEQUENCES).

	N-Terminus				**Complete Sequence**			
	AA		**SEZ**		**AA**		**SEZ**	
SIGNIF	D	**ACC**	D	**ACC**	D	**ACC**	D	**ACC**
20	-	-	-	-	1	0.37	2	0.5
19	4	0.55	11	0.76	15	0.88	8	0.77
18	25	0.87	42	0.88	49	0.93	39	0.9
17	97	0.92	133	0.915	105	0.93	79	0.92
16	268	0.92	331	0.92	212	0.94	149	0.93
15	600	0.93	649	0.92	357	0.94	253	0.94
14	1187	0.93	1185	0.92	585	0.94	386	0.93

The filtering method described in Section II-C found feature subsets with high classification accuracy. Nevertheless, their dimensionality is still quite high, which is the reason we applied the more nuanced second step of dimensionality reduction consisting on SVM-based sequential forward selection. Table IV shows, for each alphabet, the results of applying this method starting from the n-gram subset that is significant in 16 class-vs-class problems, as reported in Table III. The initial number of features (FEAT), the final number of selected features (D) and the corresponding classification accuracies are displayed.

TABLE IV. CLASSIFICATION RESULTS FOR THE AA AND SEZ ALPHABETS, USING SEQUENTIAL FORWARD FEATURE SELECTION STARTING FROM THE FIRST STAGE, T-TEST-BASED SELECTION THAT IS SIGNIFICANT IN 16 CLASS-*vs*-CLASS T-TESTS.

AA			SEZ		
FEAT	D	**ACC**	**FEAT**	D	**ACC**
268	45	0.91	331	43	0.90

B. Discussion

From Table II, it seems clear that the classification analysis using only the N-Terminus almost completely retains the accuracies obtained using the complete sequences, specially for the digram representation. This is consistent with the fact that the VFT, included in the N-terminus, contains the orthosteric binding site that, because it differentiates between different endogenous ligands, should also help to differentiate between the different class C subtypes. From a practical viewpoint, this result potentially simplifies the search for signature motifs by restricting it to the extracellular domain, while making the analysis more computationally tractable.

We are, in any case, interested in the analysis of longer n-grams. The classification results for n-grams of lengths between three and five, reported in Table III, provide evidence of the usefulness of this simple ranking approach based on filtering: the n-gram representation of the AA alphabet retains an accuracy of 0.92 with 268 attributes, while the n-gram representation of the SEZ alphabet achieves the same accuracy with 331.

The subsequent second-step, SVM-based forward selection process, starting from the optimal t-test selection was quite successful at reducing the number of attributes, while maintaining an accuracy of approximately 0.91 in the case of the AA alphabet for 45 features (a 83% reduction of the dimensionality) and a very reasonable 0.90 in the case of SEZ for 43 features (a 87% reduction), as seen in Table IV. In the case of the AA alphabet, the algorithm selects 6 *contiguous* and 39 *rigid gap* n-grams. For the SEZ alphabet, the 43 n-grams include 12 *contiguous* and 31 *rigid gap* ones. Table V lists all these n-grams in the order they were selected by the sequential forward procedure.

This list should be the starting point for proteomics experts to investigate the involvement of specific n-grams in structural and functional roles of the receptor. For class C GPCRS, this entails investigating motifs potentially related to the orthosteric site at the VFT, that is, the binding site of a ligand. The standing hypothesis for our study is that the n-grams shown to have the ability to discriminate between class C subtypes might be related to these binding sites, because the latter are meant to be subtype-specific in as much as each subtype binds to different ligands.

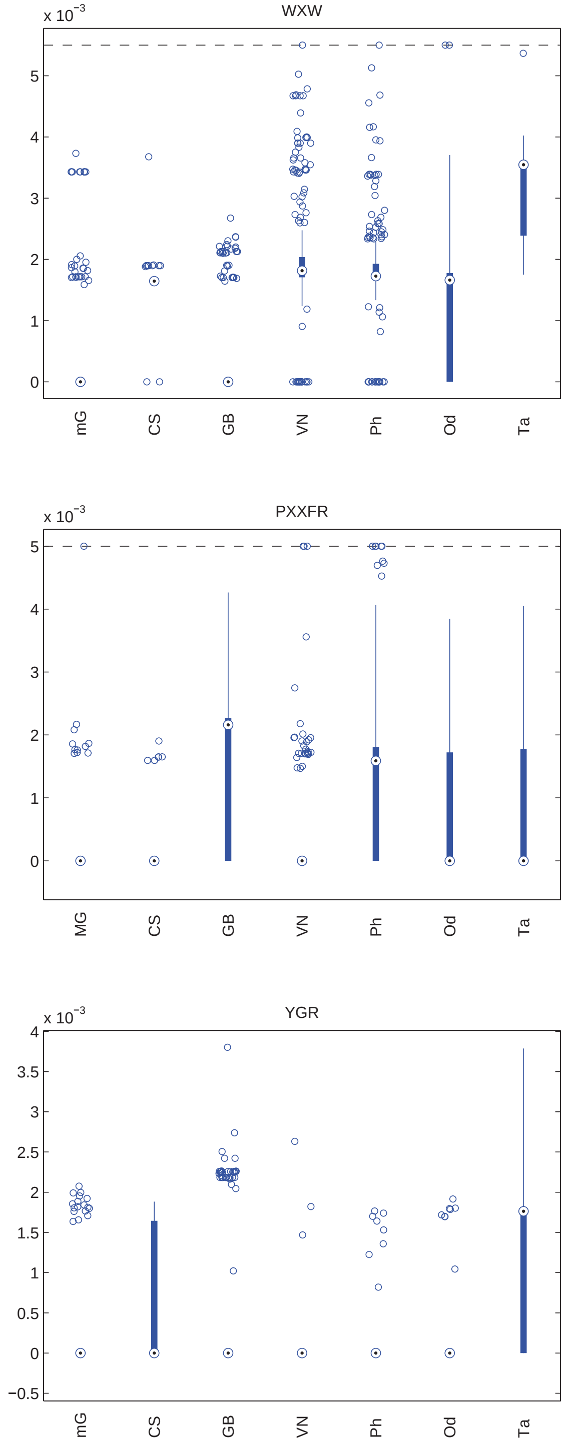
Note that we have not only provided a selected list of n-grams with the ability to discriminate the most between class C GPCR subtypes, but also an explicit ranking of relevance for these n-grams that experts can resort to. For obvious space limitations, we only show in some detail the three n-grams from each alphabet at the top of this ranking.

TABLE V. LISTS OF n-GRAMS, FROM ALPHABETS AA (LEFT COLUMN) AND SEZ (RIGHT COLUMN), RANKED BY RELEVANCE ACCORDING TO THE SEQUENTIAL FORWARD FEATURE SELECTION PROCEDURE FOR SVM CLASSIFIERS. FOR EACH ALPHABET, THE RANKING ORDER (♯), THE SYMBOLIC SUBSEQUENCE(SEE TABLE I), WHERE X IS THE WILDCARD RESIDUE IN *rigid gap* n-GRAMS, AND THE NUMBER OF BINARY CLASSIFIERS IN WHICH THE n-GRAM WAS FOUND TO BE SIGNIFICANTLY DISCRIMINANT (SIGN), ARE DISPLAYED.

AA			SEZ		
♯	n-**gram**	**SIGN**	♯	n-**gram**	**SIGN**
1	WXW	18	1	WXXW	16
2	PXXFR	16	2	G[DE]X[RKH]	16
3	YGR	17	3	[ST]XX[QN][ST]	16
4	WXWXG	17	4	GXCC	16
5	CIA	16	5	CX[IVLM]	16
6	YXI	16	6	[QN]XWG	16
7	AXXL	16	7	[ST][QN]A[RKH][IVLM]	17
8	TGXE	19	8	W[QN]X[QN]	18
9	GXXG	16	9	[ST][QN][RKH][ST]	16
10	GEXXN	17	10	PPX[ST]	17
11	DCXXG	16	11	W[IVLM][QN][RKH][DE]	16
12	FPXH	16	12	[IVLM][IVLM][IVLM][ST]W	17
13	PNXXL	18	13	[QN]X[QN]XW	16
14	WXL	17	14	[QN]GW[QN]	16
15	QXMXF	16	15	[QN]X[IVLM]XC	16
16	CXG	17	16	[IVLM]GXXC	16
17	IPG	16	17	[ST][QN]W[QN]	16
18	HXXF	17	18	[IVLM]X[ST]XC	16
19	CXXGT	17	19	[RKH]WX[IVLM]	19
20	YXKXG	17	20	[QN][ST]W	16
21	DYG	16	21	[ST][DE][ST]	16
22	PXIXY	16	22	PX[DE][ST]	16
23	WXXV	16	23	[IVLM]XW	16
24	YXXXY	16	24	AXXX[ST]	16
25	CXEXC	16	25	C[RKH]XG	17
26	VXXLL	16	26	[IVLM][IVLM]XW	16
27	SNXXD	16	27	CXAX[RKH]	16
28	SXKXQ	16	28	[QN]XGX[QN]	16
29	CXDG	17	29	[IVLM]XC[QN]	16
30	IXR	17	30	W[ST]XX[IVLM]	16
31	WXXXL	16	31	WX[RKH]W	16
32	AWXXS	16	32	W[QN]P	16
33	AXXSS	16	33	[DE]CXXC	17
34	PGXXK	16	34	[QN]CC	16
35	GXRK	16	35	[ST]XWW	16
36	PNXT	16	36	[ST]X[ST]X[QN]	16
37	VXCXD	16	37	[QN][QN]XX[ST]	16
38	GXXY	19	38	GXC[RKH]	16
39	DCLP	16	39	W[RKH]X[IVLM]	16
40	GXCXA	16	40	[QN][RKH][ST][RKH][IVLM]	16
41	IXWH	16	41	[ST]X[RKH][ST]	16
42	CXXGT	17	42	[RKH]XGXA	16
43	CXAXS	16	43	[QN]XWX[ST]	16
44	YXD	16			
45	VVFS	16			

In the case of the AA alphabet, we consider the *rigid gap* n-grams WXW (which is significant in 18 t-tests) and PXXFR (significant in 16 t-tests) and the *contiguous* YGR (significant in 17 t-tests). Figure 1 shows the corresponding relative frequencies per subtype as boxplot diagrams.

Figure 2 shows the corresponding boxplots for the three most discriminant n-grams from the SEZ alphabet. They are WXXW, G[DE]X[RKH] and [ST]XX[QN]ST, all of which are significant in 16 tests.



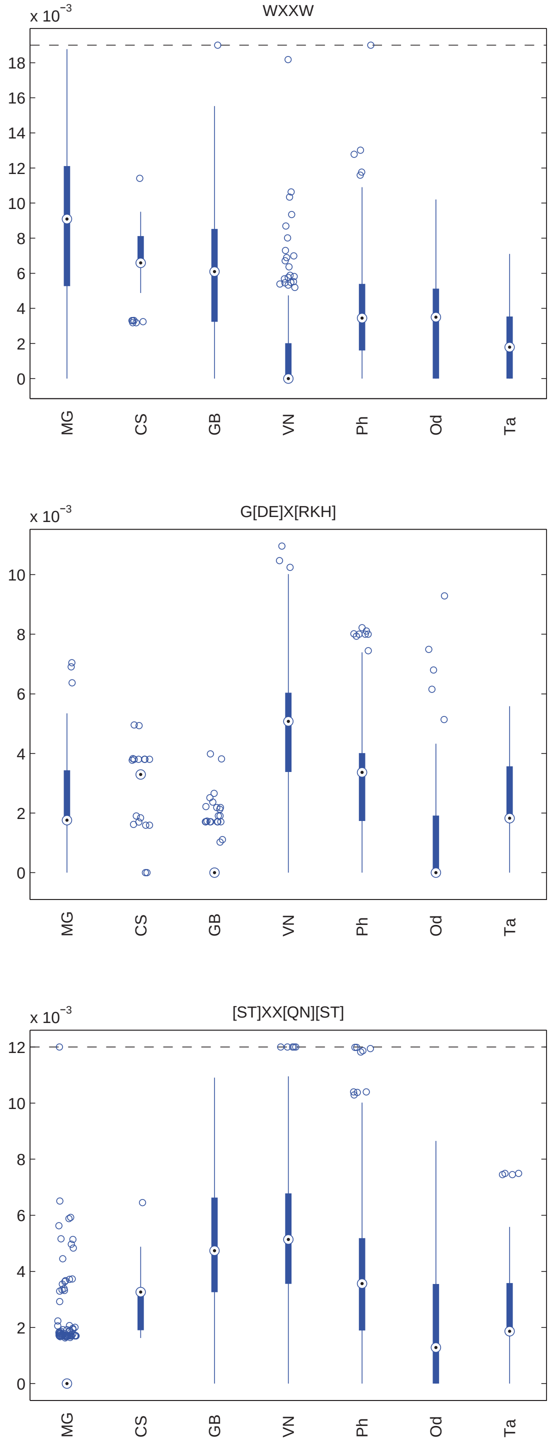


Fig. 1. Frequencies boxplots of the three n-grams of the AA alphabet ranked as the most discriminative in the classification of the 7 class C GPCR subtypes.

Fig. 2. Frequencies boxplots of the three n-grams of the SEZ alphabet ranked as the most discriminative in the classification of the 7 class C GPCR subtypes.

The AA n-grams discrimination capabilities seem to be mainly based on their existence, or lack of it, in sequences of different subtypes. This is consistent with the restrictive idea that a *signature* motif should be characterized as one that matches all the sequences of a given family and no sequence outside this family [21]. WXW seems to be mostly absent in two of the main subtypes, namely mG and GB, whereas PXXFR appearance seems mostly restricted to GB and Ph, and YGR restricted to Ta. Note also that the grouping of some frequency values for some subtypes beyond the main quartiles of the boxplots is a hint to the existence of grouping structure within subtypes. For instance, YGR seems to be present with very specific frequencies not only in Ta, but also in small subgroups of mG and GB.

The SEZ alphabet discrimination capabilities, instead, seem to be rather more subtle, as they are less based on the lack of a given n-gram than on a more gradual differentiation of the range of their frequencies. This a somehow natural consequence of their flexibility of sequence instantiation, resulting from the less granular use of the amino acid alphabet. WXXW seems very frequent in mG but infrequent in VN, Ph, Od and Ta. Instead, G[DE]X[RKH] is most frequent in VN and least in GB and Od, while [ST]XX[QN][ST] is mostly absent from mG (again with the exception of an eccentric but tight subgroup), but most frequent in GB and VN.

IV. Conclusion

Class C GPCRs are the largest class of current drug targets, with a direct involvement in a wide array of pathologies. This makes them relevant both for pharmacology and for medicine at large. Their full tertiary structure is unknown, making their functional study more complicated than that of other families of GPCRs such as, for instance, class A.

In this study, we have analyzed class C amino acid primary sequences from their *contiguous* and *rigid gap* n-gram frequencies, using a combination of feature selection and classification. This analysis involved class C subtype discrimination and aimed at identifying those n-grams most relevant to such task as candidate signature motifs. Motif over-representation in the sequence maybe the result of evolutionary preservation, which might be a lead to potential structural and functional roles. The selected discriminant n-grams may be related to the orthosteric sites at the VFT of the N-Terminal domain, given that these sites bind to different ligands for different subtypes and are thus subtype-specific.

Our previous research, using the frequencies of n-grams of length up to three obtained from the complete sequences, reported class C subtype classification accuracies that have been matched in the current study using the frequencies of a parsimonious selection of n-grams of length up to five obtained from just the N-terminal domain. Such results reinforce the interest of this extracellular domain in class C GPCR functional investigation. Of note also that the list of relatively long selected n-grams should be more effective than shorter ones as the starting point for proteomics experts to investigate motifs potentially related to the orthosteric site of the VFT, an investigation with clear potential in pharmacological research.

Acknowledgment

This research was partially funded by Spanish MINECO TIN2012-31377 and SAF2010-19257, as well as Fundació La Marató de TV3 110230 projects.

References

[1] M. Rask-Andersen, M. Sällman-Almén, and H. B. Schiöth, "Trends in the exploitation of novel drug targets," *Nat. Rev. Drug Discov.*, vol. 10, pp. 579–590, 2011.

[2] J. Kniazeff, L. Prézeau, P. Rondard, J. P. Pin, and C. Goudet, "Dimers and beyond: The functional puzzles of class C GPCRs," *Pharmacol. Ther.*, Vol. 130, pp. 9–25, 2011.

[3] H. Wu, C. Wang, K. J. Gregory, G. W. Han, K. P. Cho, Y. Xia, C. M. Niswender, V. Katritch, J. Meiler, V. Cherezov, P. J. Conn, and R. C. Stevens, "Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator," *Science*, Vol. 344, no. 6179, pp. 58–64, 2014.

[4] A. S. Doré, K. Okrasa, J. C. Patel, M. Serrano-Vega, K. Bennett, R. M. Cooke, J. C. Errey, A. Jazayeri, S. Khan, B. Tehan, M. Weir, G. R. Wiggin, and F. H., Marshall, "Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain," *Nature*, Vol. 551, pp. 557–562, 2014.

[5] T. A. Hopf, S. Morinaga, S. Ihara, K. Touhara, D. S. Marks, and R. Benton, "Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors," *Nat. Commun.*, Vol. 6, p.6077, 2015.

[6] C. König, R. Alquézar, A. Vellido, and J. Giraldo, "Finding class C GPCR subtype-discriminating n-grams through feature selection," *J. Integr. Bioinform.*, Vol. 11, no. 3, pp. 254, 2014.

[7] R. Cruz-Barbosa, A. Vellido, and J. Giraldo, "The influence of alignment-free sequence representations on the semi-supervised classification of class C G protein-coupled receptors,'" *Med. Biol. Eng. Comput.*, Vol. 53, no. 2, pp. 137–149, 2015.

[8] M. I. Cárdenas, A. Vellido, and J. Giraldo, "Visual interpretation of class C GPCR subtype overlapping from the nonlinear mapping of transformed primary sequences," in *Proc.* 2^{nd} *International Conference on Biomedical and Health Informatics (IEEE BHI'14)* pp. 764–767, 2014.

[9] M. I. Cárdenas, A. Vellido, C. König, R. Alquézar, and J. Giraldo, "Exploratory visualization of misclassified GPCRs from their transformed unaligned sequences using manifold learning techniques," in F. Ortuño, I. Rojas (eds.): *Proc.* 2^{nd} *International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2014)*, pp. 623–630, 2014.

[10] V. Katritch, V. Cherezov, and R. C. Stevens, "Structure-function of the G proteincoupled receptor superfamily," *Annu. Rev. Pharmacol. Toxicol.* Vol. 53, pp. 531-56, 2013.

[11] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg, and G. Vriend, "GPCRDB: information system for G protein-coupled receptors," *Nucleic Acids Res.*, Vol. 39(suppl 1), pp. D309–D319, 2011.

[12] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak, "Improved prediction of signal peptides: SignalP 3.0." *J. Mol. Biol.*, Vol. 340, no. 4, pp. 783–795, 2004.

[13] J. P. Pin, T. Galvez, and L. Prezeau, "Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors," *Pharmacol. Therapeut.*, Vol. 98, no. 3, pp. 325–354, 2003.

[14] P. Rondard, J. P Pin, and C. Goudet, "Dimers and beyond: The functional puzzles of class C GPCRs," *Pharmacol. Therapeut.*, Vol. 130, no. 1, pp. 9–25, 2011.

[15] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, "Protein superfamily classification using fuzzy rule-based classifier," *IEEE T. Nanobiosci.*, Vol. 8, no. 1, pp. 92–99, 2009.

[16] C. Caragea, A. Silvescu, and P. Mitra, "Protein sequence classification using feature hashing," in *proc. 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.538–543, 2011.

[17] J. Cao, and L. Xiong, "Protein sequence classification with improved Extreme Learning Machine algorithms," *BioMed Res. Int.*, ID103054, 2014.

[18] M. J. Iqbal, I. Faye, B. B. Samir, and A. M. Said, “Efficient feature selection and classification of protein sequence data in bioinformatics,” *ScientificWorldJournal*, ID173869, 2014.

[19] B. Cheng, J. Carbonell, and J. Klein-Seetharaman, “Protein classification based on text document classification techniques,” *Proteins*, Vol. 58, no. 4, pp. 955–970, 2005.

[20] M. Bhasin, and G. P. S. Raghava, “GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors,” *Nucleic Acids Res.*, Vol. 32(suppl 2), pp. W383-W389, 2004.

[21] P. G. Ferreira, and P. J. Azevedo, “Evaluating deterministic motif significance measures in protein databases,” *Algorithm. Mol. Biol.*, Vol. 2, no. 1, p. 16, 2007.

[22] A. Albayrak, H. H. Otu, and U. O. Sezerman, “Clustering of protein families into functional subtypes using relative complexity measure with reduced amino acid alphabets,” *BMC Bioinformatics*, Vol. 11, no. 1, p. 428, 2010.

[23] M. Can Cobanoglu, Y. Saygin, and U. O. Sezerman, “Classification of GPCRs using family specific motifs,” *IEEE ACM T Comput. Bi.*, Vol. 8, no. 6, pp. 1495–1508, 2011.

[24] R. K. Hart, A. K. Royyuru, G. Stolovitzky, and A. Califano, “Systematic and fully automated identification of protein sequence patterns,” *J. Comput. Biol.*, Vol. 7, no. 3-4, pp. 585–600, 2000.

[25] A. Albayrak, and U. O. Sezerman, “Discrimination of thermophilic and mesophilic proteins using reduced amino acid alphabets with n-grams,” *Curr. Bioinform.*, Vol. 7, no. 2, pp. 152–158, 2012.

[26] J. Kittler, “Feature set search algorithms,” in *Pattern Recognition and Signal Processing*, C. H. Chen, ed., pp. 41–60. Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1978.

[27] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, Vol. 23, no. 19, pp. 2507–2517, 2007.

[28] Y. Z. Guo, M. Li, M. Lu, Z. Wen, K. Wang, G. Li, and J. Wu, “Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform,” *Amino Acids*, Vol. 30, no. 4, pp. 397–402, 2006.

[29] C. Chang, and C. Lin, “LIBSVM: A library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, Vol 2, no. 3, pp. 1–27, 2011.