Reducing the n-gram feature space of class C GPCRs to subtype-discriminating patterns

Caroline König[1,*], René Alquézar[1,2], Alfredo Vellido[1,3] and Jesús Giraldo[4]

[1]Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, BarcelonaTech, 08034, Barcelona, Spain,
http://www.lsi.upc.edu/recerca-en/soco

[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, 08034, Barcelona,Spain

[3]Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), 08193, Cerdanyola del Vallès, Spain

[4]Institut de Neurociències - Unitat de Bioestadística, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Spain

Summary

G protein-coupled receptors (GPCRs) are a large and heterogeneous superfamily of receptors that are key cell players for their role as extracellular signal transmitters. Class C GPCRs, in particular, are of great interest in pharmacology. The lack of knowledge about their full 3-D structure prompts the use of their primary amino acid sequences for the construction of robust classifiers, capable of discriminating their different subtypes. In this paper, we investigate the use of feature selection techniques to build Support Vector Machine (SVM)-based classification models from selected receptor subsequences described as n-grams. We show that this approach to classification is useful for finding class C GPCR subtype-specific motifs.

1 Introduction

G protein-coupled receptors (GPCRs) are cell membrane proteins with a key role in regulating the function of cells due to their transmembrane location. This strategic location between extra- and intracellular media, together with an evolutionary-optimized 3D structure, confers them the ability to transmit extracellular signals, activating intra-cellular signal transduction pathways, making them particularly attractive for pharmacological research.

The functionality of a protein depends at large on its structural configuration in 3D, which determines its ability for ligand recognition. GPCR crystallization has been a challenging task riddled by technical obstacles until recently. The first GPCR crystal 3D structure was not fully-determined until 2000 [15] and, despite active research, only the structure of approximately

*To whom correspondence should be addressed. Email: ckonig@lsi.upc.edu

12% of the human GPCR superfamily, most of them belonging to class A, has been determined so far [9]. The transmembrane domains (TM) of the other GPCR classes are less represented, with, currently, two structures for class B, two for class C and one for class F (see [1] for comparison between classes A, B and F, and [22] and [10] for the crystal structures of the TM domain of the two class C receptors).

The current investigation focuses precisely on the characterization of class C, which is one of the five GPCR families. As the full 3D structure of their members is widely unknown, an alternative approach for feature selection relies on the analysis of GPCR primary structure, i.e. of the amino acid (AA) sequences, which are publicly available from several databases

The research whose results are reported in this paper specifically focuses on the class C subset of a publicly available GPCR database. These data were analyzed in a previous study [12] using a supervised, multi-class classification approach that yielded relatively high accuracies in the discrimination of the seven constituting subtypes of the class. This previous work used several transformations of the unaligned sequences based on the physicochemical properties of their AAs. In the current study, we go one step further and apply feature selection prior to classification with SVMs from *n-gram* subsequence features.

A key relevant objective of this work is the analysis of the constructed classifiers in order to find subfamily-specific motifs that might reveal information about ligand binding processes. A further motivation for this study is both the scarcity of structural information for the TM domain of class C GPCRs and the functional complexity found in some members of this family, for which minor changes in ligand structure lead to sharp changes in receptor selectivity and pharmacological profiles [21].

The remaining of the paper is structured as follows: Section 2 provides a brief description of the investigated receptors: both a general description of GPCRs and the specific analyzed database. This is followed, in section 3, by a description of the classifiers, the data transformation methods and the feature selection techniques. Experimental results are summarily reported and discussed in section 4. The paper wraps up with a conclusions section.

2 Materials

2.1 Class C GPCRs

GPCRs are cell membrane proteins with the key function of transmitting signals from outside to inside the cell. For this reason, they are involved in many physiological functions both in health and disease and, as a consequence, they are of special relevance in pharmacology. GPCRDB[20], the widely used database of GPCRs that was employed in our experiments, divides the GPCR superfamily into five major classes based on the ligand types, functions, and sequence similarities.

The current study concerns class C of these receptors. This class has become an increasingly important target for new therapies, particularly in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics. They are also important from structural and mechanistic

viewpoints.

Whereas all GPCRs are characterized by sharing a common seven-transmembrane helices (7TM) domain, most class C GPCRs include, in addition, an extracellular large domain, the Venus Flytrap (VFT) and a cystein rich domain (CRD) connecting both [16] . The extracellular domain contains the orthosteric site where the endogenous neurotransmitter binds whereas the 7TM domain contains allosteric sites where synthetic allosteric modulators bind and the intracellular region where cytosolic signaling proteins such as G-proteins or β-arrestin bind. Allosteric modulators are currently at the center of pharmaceutical research as they offer some advantages over orthosteric ligands, including higher subtype selectivity because of the greater sequence divergence of allosteric sites relative to orthosteric sites [23].

Class C is further subdivided into seven types: Metabotropic glutamate (mG), Calcium sensing (CS), GABA-B (GB), Vomeronasal (VN), Pheromone (Ph), Odorant (Od) and Taste (Ta), which will be the classes considered in our classification analysis.

2.2 Analyzed data

The data analyzed in this study were extracted from GPCRDB[1][20], a curated and publicly accessible database of GPCRs. The investigated dataset (version 11.3.4 as of March 2011) comprises a total of 1,510 class C GPCR sequences, belonging to seven subfamilies and including: 351 mG, 48 CS, 208 GB, 344 VN, 392 Ph, 102 Od and 65 Ta. The lengths of these sequences vary from 250 to 1,995 AAs. The mean lengths for each subfamily are 904 (mG), 949 (CS), 893 (GB), 826 (VN), 851 (Ph), 611 (Od) and 836 (Ta).

Figure 1 displays the evolutionary relationships between the seven sequence subfamilies using a phylogenetic tree (PT). A PT is a dendrogram-like graphical representation of the evolutionary relationship between the taxonomic groups that share a set of homologous sequence segments. Specifically, Figure 1 shows a Treevolution radial PT plot [18] for the 1,510 GPCR sequences under investigation and their separation into subclasses. This representation provides evidence of the heterogeneity of some of the subfamilies (such as mG, Ph and Od), as they are shown to occupy several different evolutionary branches of the tree. Although less obvious in this particular representation, there is some degree of overlapping between the different subfamilies in their tree representation.

3 Methods

In this study, SVMs [19] were used for the supervised classification of the alignment-free amino acid sequences into the seven subclasses of class C GPCRs. Given the multi-class problem setting, the LIBSVM implementation [4] was used. The AA sequences of varying lengths were first transformed into fixed-size feature representations. In previous work, we used transformations based on the physicochemical properties of the sequences [12] for this purpose. Instead, in this work we use short protein subsequences in the form of n-gram features. The n-grams,

[1] http://www.gpcr.org/7tm

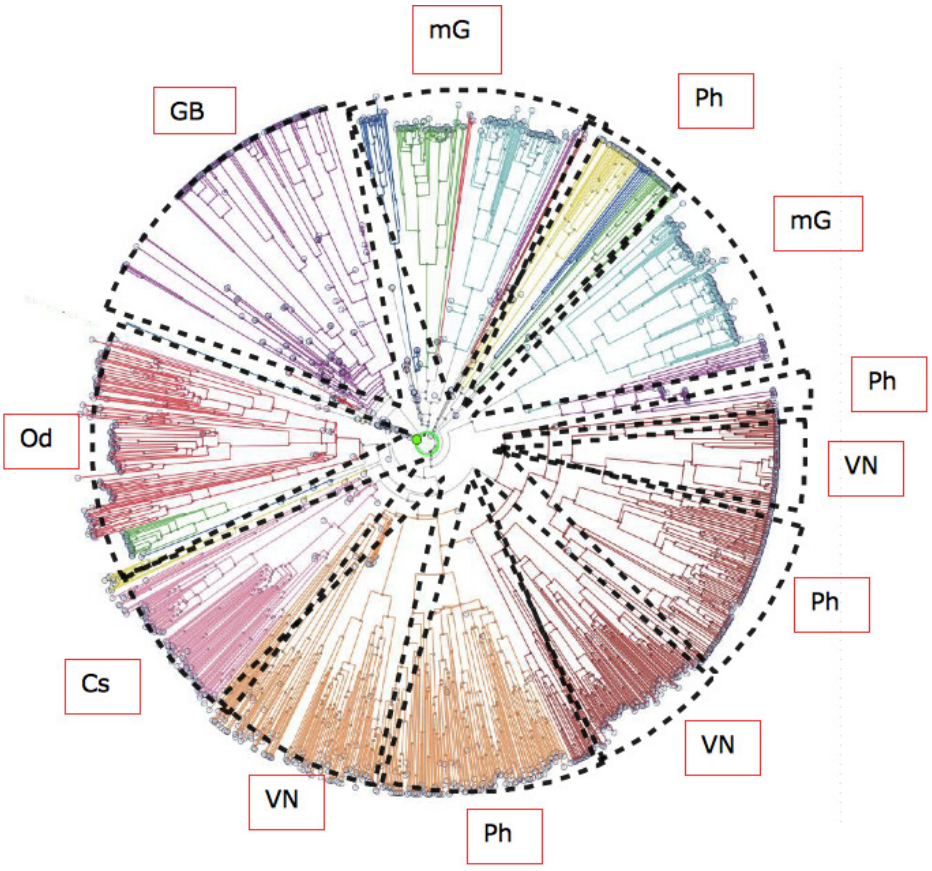

Figure 1: Treevolution radial phylogenetic tree for the 1,510 sequences under investigation [3]. Each outer leaf of each branch corresponds to a single sequence, tree colors represent families of descendant nodes. Subfamilies mG, Ph and Od are shown to cover several unrelated evolutionary branches.

described in section 3.2, were created from three different existing alphabets that have previously been used for the classification of GPCR sequences [6]. Different feature selection methods are also used to reduce the dimensionality of the data with the objective of finding the parsimonious set of n-grams that might best discriminate the class C subfamilies.

3.1 Amino acid alphabets

The 20 standard AAs can be grouped in different ways depending on the criteria used for analyzing the similarities between their physicochemical properties [8]. An appropriate grouping of AAs reduces the size of the alphabet and may decrease noise. Here, besides the basic 20-AA alphabet (AAA), we used two alternative amino acid groupings (See Table 1): the Sezerman (SEZ) alphabet, which includes 11 groups, and the Davies Random (DAV), including 9 groups. They have both been evaluated in the classification of GPCRs into their 5 major classes [6].

Table 1: Amino acid grouping schemes.

	1	2	3	4	5	6	7	8	9	0	X
SEZ	**IVLM**	**RKH**	**DE**	**QN**	**ST**	**A**	**G**	**W**	**C**	**YF**	**P**
DAV	**SG**	**DVIA**	**RQN**	**KP**	**WHY**	**C**	**LE**	**MF**	**T**		

3.2 N-grams

The concept of n-gram has widely been used in protein analysis ([2],[14]). A successful application of text classification methods for the classification of class A GPCRs was presented in [5]. While a discretization of the n-gram features was used in that study, we instead use the relative frequencies of the n-grams, which are non-discrete variables, in our experiments. Therefore, the n-gram feature representation corresponds here to the measurement of the relative frequency of each n-gram in a sequence. Due to the exponential growth of the size of n-grams, we limit the reported research to n-grams of sizes 1, 2 and 3.

3.3 SVMs

SVM classifiers are founded on the statistical learning theory first introduced in [19]. SVMs map the feature vectors $x_i, i = 1, \ldots, N$, where $x_i \, \epsilon \, \mathbb{R}^n$ and N is the number of instances, into possibly higher dimensional spaces by means of a function ϕ . The objective is to find a linear separating hyperplane, which separates the feature vectors according to its class label with a maximal margin, while minimizing the classification error ξ. The use of non-linear kernel functions allows SVMs to separate input data in higher dimensional spaces that would not be separable with linear classifiers in the original input space.

The radial basis function (RBF) kernel, specified as $K(x_i, x_j) = e^{(-\gamma||x_i - x_j||)}$, is a popular non-linear kernel. To use it, the SVM needs to adjust two parameters through grid search: the error penalty parameter C and the parameter γ of the RBF function. The goal of separating the seven subclasses of the class C GPCRs requires the extension of the original two-class classification approach of SVMs to a multi-class classification approach. To that end, we have chosen the "one-against-one" approach to build the global classification model, which is implemented in the LIBSVM[2] library [4].

3.4 Performance measures

Two different measures were used to evaluate the test performance of the multi-class trained classifiers, namely the Accuracy (ACC), which is the ratio of correctly classified instances to all instances, and the Matthews Correlation Coefficient (MCC), which indicates how predictable the target variable is knowing the other variables: its value ranges from -1 to 1, where 1 corresponds to a perfect classification, 0 to a random classification and -1 to complete misclassification. The MCC is usually accepted to be a balanced figure of merit when classes are of

[2]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

unbalanced sizes (although, as described in section 2.2, the classes analyzed in our experiments are not too unbalanced).

For the individual (binary) classification of each subtype, we report the MCC and two further common measures: Precision and Recall. The former is the ratio of cases belonging to a class that are correctly classified to the cases predicted to belong to such class, whereas the later is the ratio of cases belonging to a class that are correctly classified to the cases that actually belong to that class.

The multi-class trained classifier is evaluated through a 5-fold cross-validation (CV) with stratification. The reported measures are the mean values of the respective metric over the five iterations of the 5-CV.

In summary, we measure the Accuracy and MCC at the global level in our experiments, while we measure the Precision, Recall and MCC at class level.

3.5 Feature selection

Many irrelevant features are likely to exist in the different n-gram frequency representations of the data. To ameliorate the classification process by minimizing the negative impact of irrelevant features, we use two different feature selection approaches in this study: sequential forward feature selection with an SVM-classifier and a classifier-independent basic filter method that computes two-sample t-tests among the C GPCR subfamilies.

A sequential forward selection algorithm [11] is used to find the reduced set of features that best discriminated the data subtypes. This kind of algorithm is a so-called wrapper method, where the classification model search is performed within the subset feature search [17].

The algorithm starts from an empty candidate feature set and adds, in each iteration, the feature which most improves the accuracy (i.e., that which minimizes the misclassification rate). The algorithm uses an SVM classifier in which the accuracy is evaluated using a 5-CV to test the candidate feature set. The algorithm stops when the addition of a further feature does not increase the accuracy over a threshold set at $1e^{-6}$.

As an alternative filtering approach, a two-sample t-test was used to evaluate the discriminating power of each feature. 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 confidence. If the t-test suggested that this hypothesis was true (i.e. the null hypothesis was rejected), the feature was considered to significantly distinguish between the two different subtypes of class C GPCRs.

As we face a multi-class classification problem, the t-test results were examined for the 21 feasible two-class combinations of the 7 class C subfamilies. We decided to calculate the two-sample t-test values at this detail because the multi-class LIBSVM implementation internally performs a comparison of the data between each class (one-*vs*-one implementation). Therefore, the t-test exactly evaluates the data considered in each binary classifier, making the ranking of

Table 2: N-gram classification results for the different alphabets without feature selection, where N is the size of a feature set and ACC stands for classification accuracy (ratio of correctly classified sequences to all sequences).

	AAA		SEZ		DAV	
N-GRAM	**N**	**ACC**	**N**	**ACC**	**N**	**ACC**
1-gram	20	0.870	11	0.820	9	0.780
2-gram	400	0.930	121	0.926	81	0.910
1,2-gram	420	0.930	132	0.921	90	0.916

the features possible according to their overall significance (i.e., in how many binary classifiers a feature is significant).

Note that, as explained in the following sections, the t-test filtering method is not used here only as an alternative to dimensionality reduction using a wrapper approach; in fact, it is used as a first classifier-independent feature selection step in problems that are computationally too demanding for classifier-dependent feature selection from very high-dimensional data sets.

4 Experiments

4.1 Classification according to n-gram representation

First, we built classification models with n-grams for each of the three alphabets (AAA, SEZ, DAV). Table 2 shows the classification results obtained when no feature selection method was applied, as well as the size of the feature set for each alphabet. Note that each element in each alphabet is itself considered as a 1-gram, regardless the number of constituent AAs. Obviously, the size of the n-gram feature set increases significantly with the size of the alphabet. Results are shown for 1-grams, 2-grams, and the combination of both.

The construction of an SVM model from 3-grams for all three alphabets was unsuccessful, probably due to the existence of a large set of irrelevant features. This was the primary reason behind the decision of applying feature selection as part of the classification process.

4.2 Sequential forward feature selection

Table 3 shows the classification results when sequential forward selection was performed on each n-gram dataset. For each alphabet (AAA, SEZ and DAV), this table shows a comparison between the original size of the n-grams (N) and the number of selected features found by the algorithm (FS), as well as the corresponding classification accuracy.

The experiments show that the feature selection algorithm was successful, with only one exception: in the case of the 1,2,3-gram feature set (combination of all n-grams) of the AAA-alphabet: due to the large number of features, the computational cost of the forward selection

Table 3: N-gram classification results using sequential forward feature selection, for the three different alphabets.

	AAA			SEZ			DAV		
N-GRAM	**N**	**FS**	**ACC**	**N**	**FS**	**ACC**	**N**	**FS**	**ACC**
1-gram	20	17	0.880	11	10	0.790	9	7	0.770
2-gram	400	48	0.930	121	25	0.906	81	31	0.900
1,2-gram	420	54	0.926	131	37	0.916	90	42	0.920
1,2,3-gram	8420	-	-	1331	34	0.925	818	34	0.923

Table 4: Classification results with t-test-based subset selection, with subsets of features that are significant in a given number of t-tests, from 20 down to 12.

	AAA		SEZ		DAV	
SIGNIF	**N**	**ACC**	**N**	**ACC**	**N**	**ACC**
20	1	0.370	2	0.500	0	-
19	15	0.880	8	0.770	10	0.830
18	49	0.931	39	0.900	23	0.880
17	105	0.933	79	0.922	58	0.910
16	212	0.937	149	0.930	99	0.920
15	357	0.936	253	0.936	164	0.926
14	585	0.943	386	0.935	238	0.933
13	909	0.937	505	0.943	325	0.930
12	1284	0.942	633	0.940	429	0.927

algorithm is too high. In fact, this was the result that prompted us to investigate a classifier-independent filter feature selection method that could provide us with a first rough selection of features to be used as a preliminary step to a subsequent process of forward feature selection.

4.3 t-Test filtering

In order to handle the 1,2,3-gram feature sets, which, due to their size, were either impossible or very difficult to use in the previous methods, we decided to use the t-test filtering method to create a ranking of the features. Table 4 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 N) in at least 20,19,18, etc. of the total 21 two-class tests. The ACC values shown for each subset are the classification accuracies of an SVM-classifier built on each feature set.

4.4 t-Test filtering and forward selection

The filtering method described in the previous section found feature subsets with high classification accuracy. Nevertheless, given their high dimensionality, we decided to apply the forward

Table 5: Classification results with forward selection on top of t-test-based selection for a subset solution in which features are significant in at least 12 of the 21 t-tests.

AAA			SEZ			DAV		
FEAT	**N**	**ACC**	**FEAT**	**N**	**ACC**	**FEAT**	**N**	**ACC**
1284	49	0.939	633	59	0.939	429	60	0.940

selection algorithm to these subsets as a subsequent dimensionality reduction step.

Table 5 shows the results of applying forward selection starting from the n-gram subset reported in the last row of Table 4 (features relevant in at least 12 classifiers), for each alphabet. The initial number of features (FEAT), the number of selected features (N) and the corresponding classification accuracies are shown.

4.5 Discussion

4.5.1 Classification from n-grams with and without feature selection

The results reported in Table 4 provide clear evidence of the usefulness of the t-test-based simple feature ranking method, as parsimonious feature subsets that outperform the classification accuracies obtained without feature selection or with forward selection on its own were found. For example, the 1,2,3-gram representation of the AAA alphabet achieves an accuracy of 0.943 with 585 attributes, improving on the 0.930 accuracy obtained directly with the 2-gram representation using only forward selection (as reported in Table 3). In the case of the SEZ alphabet, the same 0.943 accuracy was obtained with this filtered 1,2,3-gram representation with 505 features; this has again to be compared to the 0.926 obtained with the 2-gram representation (Table 2) and the 0.925 obtained with the 1,2,3-gram representation (Table 3). Using the DAV alphabet, we found a subset with 238 features that yielded a 0.933 accuracy, whereas the 1,2,3-gram representation with forward selection yielded a 0.920 (Table 3).

Nevertheless, the filter selection method on its own still renders rather high-dimensional optimal solutions and the slight classification improvement it generates might not be enough to counter-balance the complexity of the solution. In fact, the most interesting results, as reported in Table 5, come from the application of the classifier-dependent forward selection to the results of the filter method. Results show that this approach was quite successful at reducing the number of attributes by as much as 96% while retaining an accuracy in the area of 0.94 for all three alphabets.

Overall, the experimental results reported in the previous section support the interest of using feature selection on the analyzed n-gram data: data dimensionality has been notably reduced without compromising classification quality. Forward selection has been shown to be an effective method, although it is computationally too costly when the size of the feature set is too high from the onset. In this situation, a fast univariate t-test-based filtering method becomes an appropriate solution to reduce the feature candidate set as a preprocessing step prior to the forward selection algorithm.

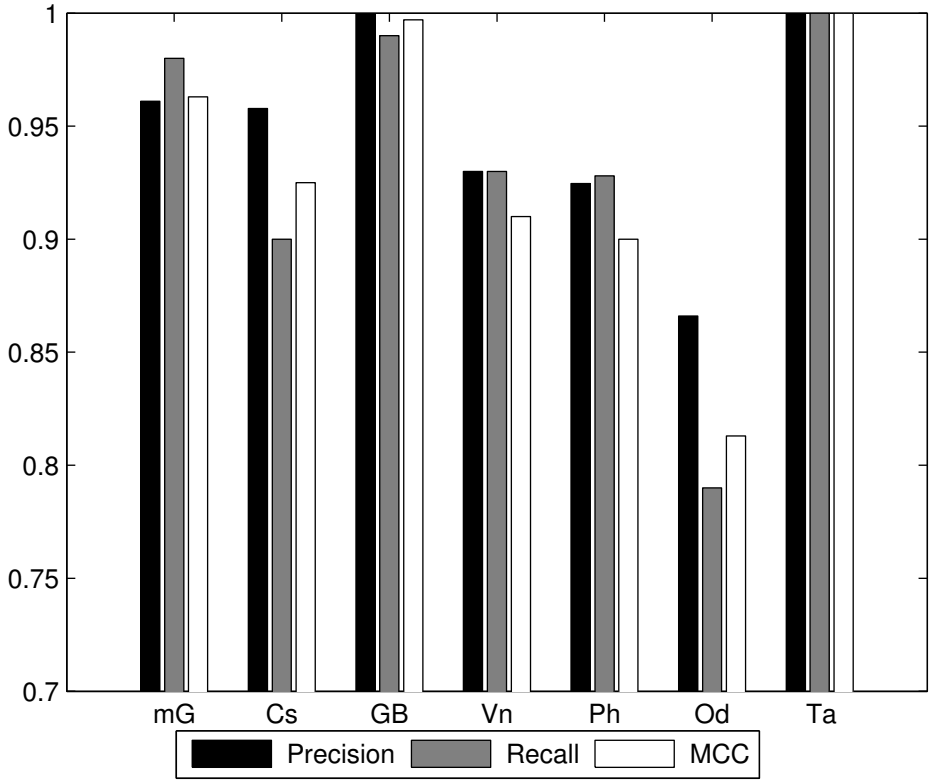

Figure 2: Graphical representation of the Precision, Recall and MCC per subfamily, for the AAA alphabet.

To the best of the authors' knowledge, the reported classification results are the best to date using the class C GPCRDB database, comparing favourably to those in [7, 12, 13]. These results correspond to the reduced SEZ dataset with 505 attributes (See Table 4) and a SVM model with parameter C=2, γ= 2^{-10} achieving a mean accuracy (ACC) of 0.943 and a mean MCC of 0.93. Figure 2 shows the corresponding per-class classification results, i.e. the Precision, Recall and MCC of each binary classification.

4.5.2 Qualifying feature selection from t-test values

An analysis of the t-test values (hypothesis value and p-value) allows measuring to what degree an individual feature discriminates between two classes. Test values are first analyzed to detect the 3-grams with the best discrimination capabilities. We subsequently analyze if these 3-grams may be part of larger n-grams which may also be discriminative.

The close scrutiny of the test values of the reduced feature set of the AAA alphabet (See Table 5: 49 features, including 33 3-grams, 13 2-grams and 3 1-grams) revealed that the 3-grams CSL, ITF and FSM are the most significantly discriminative.

CSL, in particular, is the most significant one according to the t-test values of 20 two-sample tests. This feature was only found not to be significant for the mG *vs.* Ph discrimination.

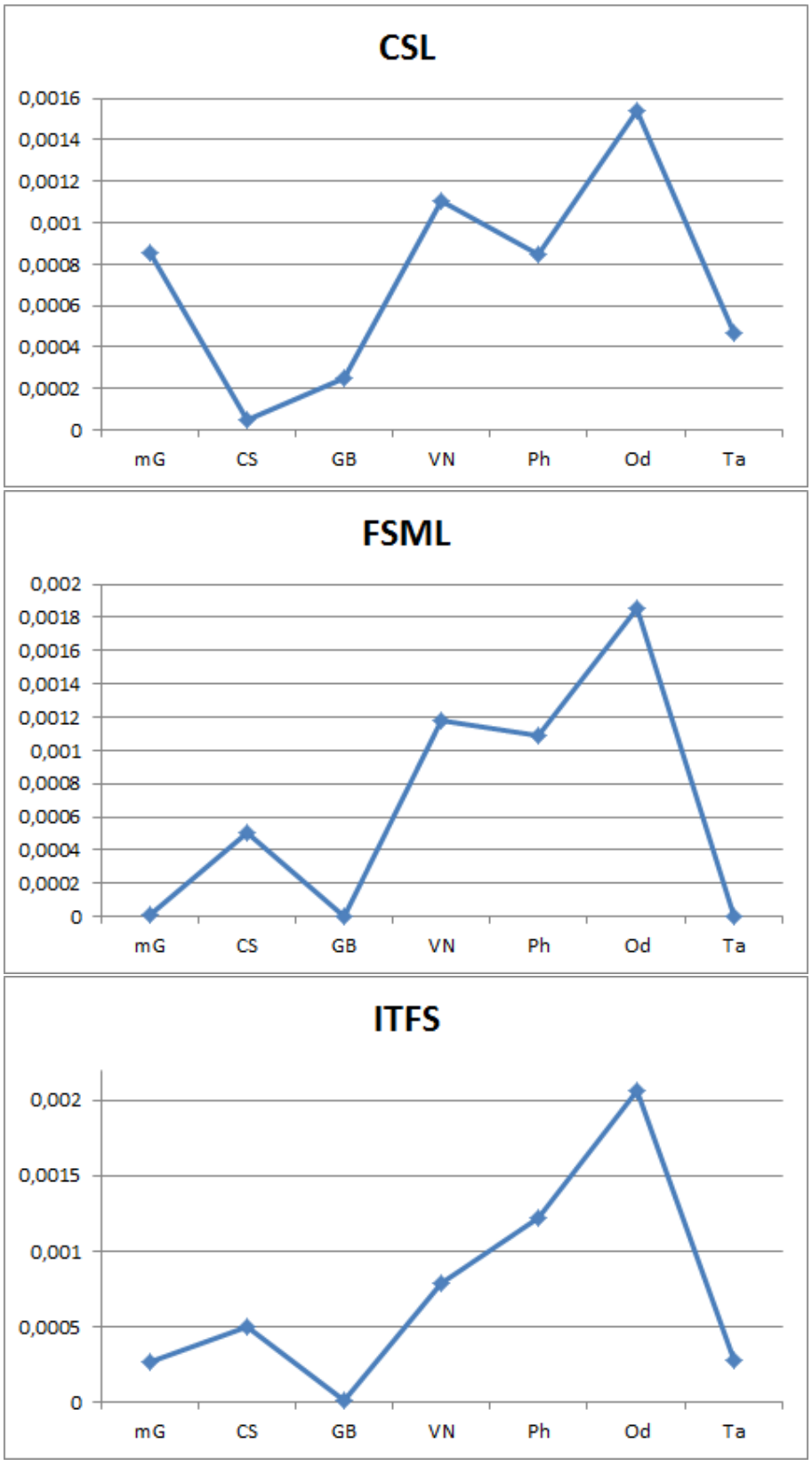

Figure 3: Mean values of CSL (top), FSML (center) and ITFS (bottom) N-gram features for the 7 class C GPCR subfamilies.

The ITF n-gram is deemed to be significant in 18 tests and an analysis of longer n-grams (results not reported) showed that the the ITFS 4-gram is specially discriminating, with a significant impact on the discrimination of 19 binary classifiers (i.e., all but mG *vs.* Ta and CS *vs.* Ta). Furthermore, the ITFSM 5-gram is still highly discriminative, showing significant values for 17 tests.

Another relevant 3-gram is FSM, which is significant for 18 two-class tests. An analysis of longer n-grams showed that the FSML 4-gram is highly discriminative (in 18 tests: all but mG *vs.* GB, mG *vs.* Ta and GB *vs.* Ta). The FSMLI 5-gram was also found to be significant for 15 tests.

Figure 3 shows the mean values of n-gram features CSL, ITFS, and FSML for the 7 class C GPCR subtypes.

Beyond mean values, a statistical analysis of the most discriminative 3-grams, CLS, ITF and FSM, revealed the existence of extreme values in some subfamily distributions, which would require a deeper analysis: Figures 4, 5 and 6 display box plots of the corresponding n-grams. The box describes the range of values between the first and the third quartiles (Q1 and Q3) with the median (Q2) as the horizontal line inside the box. The crosses are data considered to be outliers, which, in this case, are points which fall below Q1-1.5(IQR) or above Q3+1.5(IQR), where IQR is the interquartile range described by the box. The interval in which the data are considered not to be outliers is represented in the plot by the dashed lines stemming from the box.

The n-gram CSL (Figure 4), which was found to be discriminant in 20 two-class tests, has its maximum values for classes Od, VN, Ph and mG, whereas this n-gram is mostly non-existent in classes Cs and GB. The statistical analysis of the distribution of this n-gram confirms that CSL is suitable for the description of nearly all subfamilies (except GB) as only a relative small number of outlier values exist for all of them. In subfamily GB, this n-gram is mostly non-existent, but 17% of the sequences of this subclass appear as outliers (corresponding to sequences containing this n-gram). In consequence, subfamily GB is not well represented by n-gram CSL as its distribution is not uniform. A superficial analysis of the location of the n-grams in the sequence shows that in class Od, this n-gram appears near the middle of the sequences as well as near to their end. In the case of Ta, it appears often near the beginning, while in VN it appears in all positions (beginning, middle and end).

The n-gram ITF (Figure 5) was found to be discriminant in 18 tests and has maximum values for the subfamilies Od, Ph, VN and CS. The data of the corresponding box plot confirms that this n-gram is suitable for the discrimination of these subfamilies as the existence of extreme values is quite low in these cases. For GB and Ta, this n-gram is mostly non-existent as both the median and the IQR are zero and a low number of sequences have a positive frequency of this n-gram. Despite the fact that mG also has a median and IQR with value zero, mG has to be considered a special case as its distribution has approx. 10% of outliers, which correspond to sequences containing this n-gram. Regarding the subsequence specific location of the ITF n-gram, it appears in any position (beginning, middle and end) in class Od, while in Ph and VN, it is predominantly located at the end, and in CS it is found near the middle section.

Finally, n-gram FSM (Figure 6), which was deemed significant in 18 tests, shows maximum

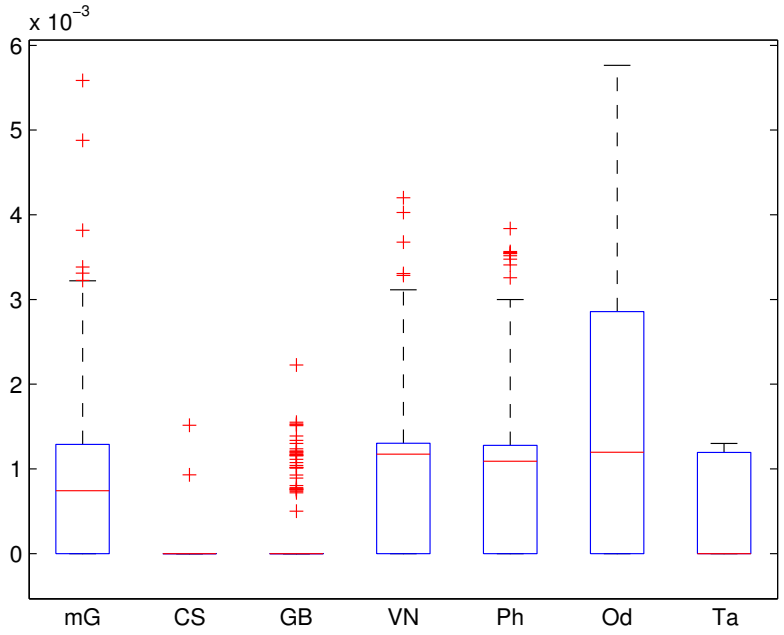

Figure 4: Box plot of the CSL n-gram.

values for subfamilies Od, VN, Ph and CS and is mostly non-existent in subfamilies mG, GB and Ta. Nevertheless, the box plot representation suggests that this n-gram describes properly Od and CS as subclasses with presence of this n-gram and GB and Ta as subfamilies not containing this n-gram. Subfamilies mG, VN and Ph show a higher number of outliers, namely 5%(mG), 14%(VN) and 13%(Ph), which indicates that the appearance of FSM in these subfamilies is not uniform. Regarding the location of the n-gram, FSM appears in the class Od at the middle and at the end of the sequence. In the case of CS, it appears at the middle; in Vn, it appears at the end, and in Ph, both at the end and beginning.

Overall, these n-grams might be the basis for an ulterior investigation of specific motifs in class C GPCR sequences that might provide clues about ligand binding processes.

5 Conclusions

Class C GPCRs, a family of receptors of great interest in pharmacology, are usually investigated from their primary sequences. This study has addressed the problem of class C GPCR subtype discrimination according to a novel methodology that transforms the sequences according to the frequency of occurrence of the low level n-grams of different AA alphabets.

These sequence transformations generate high-dimensional data sets that are likely to include plenty of irrelevant information. For this reason, dimensionality reduction through combination of a two-sample t-test and forward feature selection was implemented as part of classification with SVMs.

Reduced sets of n-grams that yielded similar classification accuracies were found for each of

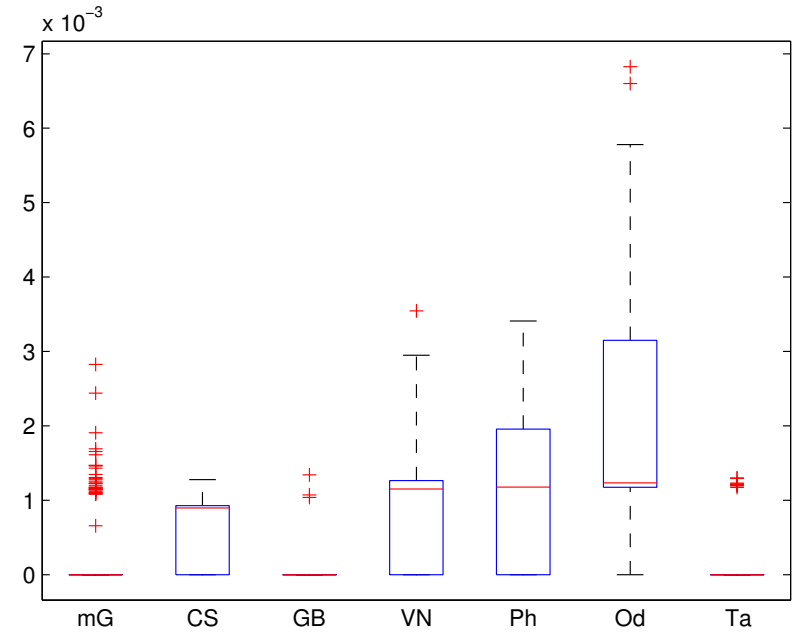

Figure 5: Box plot of the ITF n-gram.

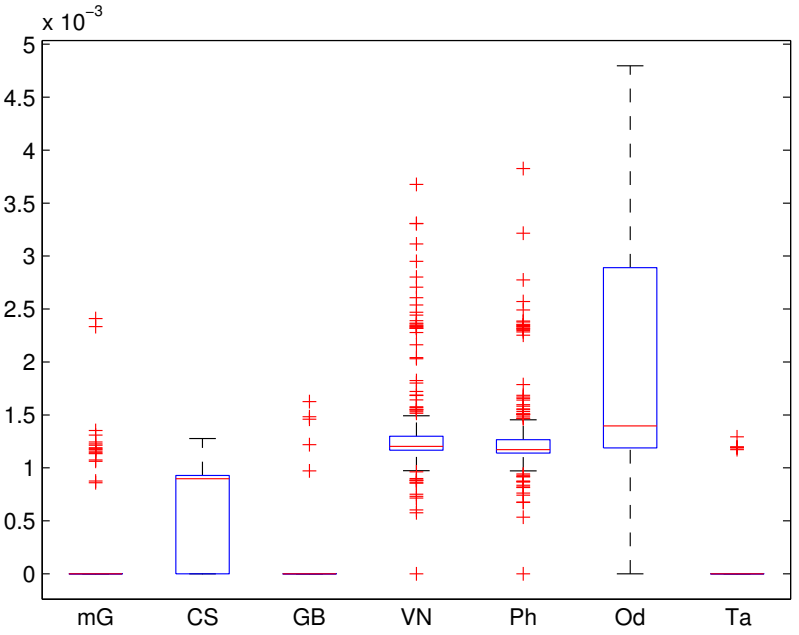

Figure 6: Box plot of the FSM n-gram.

the three transformation alphabets. These results are the best reported to date using the class C data from the GPCRDB database.

The analysis of the features of the AAA alphabet using the values obtained in the t-tests has provided insight about the n-grams that are best at discriminating between the GPCR subfamilies. This might be considered as preliminary evidence of the existence of subfamily-specific motifs that might reveal information about ligand binding processes. For this reason, the proposed method will be extended in future work to the analysis of larger n-grams. From this analysis, we expect to find larger n-grams that might actually be considered as potentially true subtype-specific motifs.

The analysis of the statistical distributions of the attribute values provided further insight about the nature of the analyzed data. Although the highly discriminative n-grams contributed to achieve high classification accuracy, the detected n-grams were not equally suitable to explain the data of all subfamilies. The n-grams were only appropriate to describe the distribution of the values of given subsets of subfamilies. This may be the result of the heterogeneity of some of these subfamilies. As explained in Section 2.2, some subfamilies group nodes which are descendants from evolutionary unrelated proteins leading to separate groups. On the other hand, the data also contains overlapping data as some subclasses contain sequences which are descendants from a common ancestor. This might come to explain why, in this multi-class classification problem, the feature selection algorithm required to reach a certain number of attributes (10-30) to achieve high classification accuracies. In future work, we will address this issue by taking into account the possible subdivisions of the analyzed subfamilies.

The study of the location of n-grams in the sequence revealed that they do not appear at the same locations in different subfamilies. This discovery encourages us to apply the proposed feature selection method to separated sequence segments in order to compare n-grams specific to their subsequence specific location.

Acknowledgements

This research was partially funded by MINECO TIN2012-31377 and SAF2010-19257, as well as Fundació La Marató de TV3 110230 projects and ERA-NET NEURON PCIN-2013-018-C03-02.

References

[1] A. Bortolato et al., Structure of class B GPCRs: new horizons for drug discovery *British Journal of Pharmacology*, 171(13):3132–45, 2014

[2] C. Caragea, A. Silvescu, P. Mitra, Protein Sequence Classification Using Feature Hashing, In *2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 538–543, 2011.

[3] 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 *Procs. of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, pp. 623–630, 2014.

[4] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[5] B. Cheng, J. Carbonell, and J. Klein-Seetharaman, Protein classification based on text document classification techniques, *Proteins: Structure, Function, and Bioinformatics*, 58(4):955–970, 2005.

[6] M. C. Cobanoglu, Y.l Saygin, and U. Sezerman, Classification of GPCRs using family specific motifs, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1495–1508, 2011.

[7] R. Cruz-Barbosa, A. Vellido, J. Giraldo. 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)*, pp. 759–766, 2013.

[8] M. N. Davies, A. Secker, A. Freitas, E. Clark, J. Timmis, and D. R. Flower, Optimizing amino acid groupings for GPCR classification, *Bioinformatics*, 24(18):1980–1986, 2008.

[9] V. Katritch, V. Cherezov, and R. C. Stevens, Structure-function of the G protein coupled receptor superfamily, *Annual Review of Pharmacology and Toxicology*, 53(1):531–556, 2013.

[10] A. S. Dor, 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*, 551:557–562, 2014.

[11] J. Kittler, Feature Set Search Algorithms, *Pattern Recognition and Signal Processing*, pp. 41–60, 1978.

[12] C. König, R. Cruz-Barbosa, R. Alquézar and A. Vellido, SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences, *ICIAP 2013 Workshops, Lecture Notes in Computer Science*, 8158, pp. 336–343, 2013.

[13] C. König, A. Vellido, R. Alquézar and J. Giraldo, Misclassification of class C G-protein-coupled receptors as a label noise problem, In *Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 695–700, 2014.

[14] F. Mhamdi, M. Elloumi, R. Rakotomalala, Textmining, features selection and datamining for proteins classification, In *Proceedings of the 2004 International Conference on Information and Comunication Technologies: From Theory to Applications*, pp. 457–458, IEEE, 2004.

[15] K. Palczewski, T. Kumasaka, T. Hori et al., Crystal structure of Rhodopsin: a G Protein-Coupled Receptor, *Science*, 289(5480):739–745, 2000.

[16] J. P. Pin, T. Galvez, and L. Prezeau, Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors, *Pharmacology & Therapeutics*, 98(3):325 – 354, 2003.

[17] Y. Saeys, I. Inza, and P. Larrañaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23(19):2507–2517, 2007.

[18] R. Santamaría and R. Therón, Treevolution: visual analysis of phylogenetic trees, *Bioinformatics*, 25(15):1970–1971, 2009.

[19] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 1998.

[20] 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 Research*, 39(suppl 1):D309–D319, 2011.

[21] M.R. Wood et al., Molecular switches on mGluR allosteric ligands that modulate modes of pharmacology, *Biochemistry*, 50:2403–2410, 2011

[22] H. Wu, C. Wang, K. J. Gregory, et al., Structure of a class C GPCR Metabotropic Glutamate Receptor 1 bound to an allosteric modulator, *Science*, 344(6179):58-64, 2014.

[23] S. Yin and C. M. Niswender, Progress toward advanced understanding of metabotropic glutamate receptors: structure, signaling and therapeutic indications, *Cellular Signalling*, 26(10):2284–2297, 2014.