Finding class C GPCR subtype-discriminating n-grams through feature selection

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

[1] Departament de Llenguatges i Sistemes Informàtics, Univ. Politècnica de Catalunya, BarcelonaTech, 08034, Barcelona, Spain
[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 Bioestadìstica, Univ. Autònoma de Barcelona, 08193, Cerdanyola del Vallès , Spain
{ckonig, alquezar, avellido}@lsi.upc.edu, jesus.giraldo@uab.es

Abstract. 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 describe 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.

Keywords: G-Protein coupled receptors, pharmaco-proteomics, feature selection, n-grams, support vector machines

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 is the result of their ability to transmit extracellular signals, activating intra-cellular signal transduction pathways, ability that makes them particularly attractive for pharmacological research.

The functionality of a protein depends at large on its structural configuration in 3-D, which determines its ability for a given ligand binding. Despite active research, the 3-D structure is currently only determined in full for approximately a 12% of the human GPCR superfamily [6]. As a result, GPCR classes that lack a known 3-D structure require alternatives such as the analysis of their primary amino acid sequence, which is well-known and reported in many open curated databases.

This paper specifically focuses on the class C subset of a publicly available GPCR database. These data were analyzed in a previous study [8] 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 based on the physicochemical properties of the sequence amino acids. In the current study, we go one step further and apply feature selection prior to classification with SVMs from n-gram subsequence features. A 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 the fact that no major motifs are currently known for class C GPCRs [11].

2 Materials

GPCRs are cell membrane proteins that transmit signals from the extracellular to the intracellular domain, prompting cellular response. This makes them of great relevance in pharmacology. The GPCRDB [12], a popular curated database of GPCRs, divides the superfamily into five major classes (namely, A to E) based on ligand types, functions, and sequence similarities. As stated in the introduction, this study concerns class C, which has of late become an increasingly important target for new therapies, particularly in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics.

The investigated data (from version 11.3.4 as of March 2011) comprises of 1,510 class C GPCR sequences, belonging to seven subfamilies: 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).

3 Methods

In this work, SVMs 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 svmLib implementation [2] was used. The amino acid sequences of varying lengths were first transformed into fixed-size feature representations. For this, we used in previous work transformations based on the physicochemical properties of the sequences [8]. Instead, in this work we use short protein subsequences in the form of n-gram features. The n-grams were created from three different existing alphabets that have previously been used for the classification of GPCR sequences [4]. 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 subtypes.

3.1 Amino acid alphabets

According to [5], many amino acids have similar phisicochemical properties, which makes them equivalent at a functional level. An appropriate grouping of amino acids reduces the size of the alphabet and may decrease noise. In this work, besides the basic 20-amino acid alphabet, 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 [4] in the classification of GPCRs into their 5 major classes.

Table 1: Amino acid grouping schemes

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

3.2 N-grams

The concept of n-grams has widely been used in protein analysis ([1],[9]). A successful application of text classification methods for the classification of class A GPCRs was presented in [3]. While a discretization of the n-gram features was used in that study, we instead used the relative frequencies of the n-grams, which are non-discrete variables. 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 size 1,2 and 3.

3.3 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 used two different feature selection approaches in this study: sequential forward feature selection with an SVM-classifier and a filter method computing two-sample t-tests among the C GPCR subtypes.

A sequential forward selection algorithm [7] was 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 [10].

This 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 $1e^{-6}$.

A two-sample t-test was used to evaluate the discriminating power of each feature as a filtering approach. 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 subtypes. We decided to calculate the two-sample t-test values at this detail because the multi-class svmLib 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 the features possible according to their overall significance (i.e., in how many binary classifiers a feature is significant).

4 Experiments

4.1 N-gram representation

First, we built classification models with n-grams for each of the three alphabets (AA, SEZ, DAV). Table 2 shows the classification results obtained and the size of the feature set for each alphabet. We observe that the size of the n-gram feature set decreases significantly with the size of the alphabet, but that the best classification results are obtained for the AA alphabet, which is the largest. Nevertheless, the construction of an SVM model with 3-grams for all three alphabets was unsuccessful, probably due to the existence of a large set of irrelevant 3-grams. For this reason, feature selection was implemented.

Table 2: N-gram classification results, where N is the size of a feature set and ACC stands for classification accuracy (ratio of correctly classified sequences).

	AA		**SEZ**		**DAV**	
N-GRAM	**N**	**ACC**	**N**	**ACC**	**N**	**ACC**
1-gram	20	0.87	11	0.82	9	0.78
2-gram	400	0.93	121	0.926	81	0.91
1,2-gram	420	0.93	132	0.921	90	0.916

4.2 Sequential Forward Feature Selection

Table 3 shows the results of the sequential forward selection performed on each n-gram dataset. For each alphabet (AA,SEZ,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, as well as the corresponding classification accuracy. The experiments show that the feature selection algorithm was successful, as it was able to find, in almost all cases, a reduced subset of features providing approximately the same prediction accuracy. There were two exceptions: in the case of the 1-grams of the SEZ and DAV subsets, the algorithm was not able to reduce the number of features, probably due to the small size of the feature set. The other exception is the 1,2,3-gram feature set of the AA-alphabet: due to the large number of features the computational cost of the forward selection algorithm is too high. For this reason, we decided to apply a filtering method to reduce the candidate feature subset as a previous step to the forward selection.

Table 3: N-gram classification results using feature selection

	AA			**SEZ**			**DAV**		
N-GRAM	**N**	**FS**	**ACC**	**N**	**FS**	**ACC**	**N**	**FS**	**ACC**
1-gram	20	17	0.88	11	-	-	9	-	-
2-gram	400	48	0.93	121	25	0.906	81	31	0.9
1,2-gram	420	54	0.926	131	37	0.916	90	42	0.92
1,2,3-gram	8420	-	-	1331	34	0.925	818	34	0.923

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 establish 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. two-class tests. The ACC values shown for each subset are the classification accuracies of a SVM-classifier built on each feature set.

These results provide evidence of the usefulness of this simple ranking, as we were able to find subsets that outperform the classification accuracies obtained with the previous methods. For example, the 1,2,3-gram representation of the AA alphabet achieves an accuracy of 0.943 with 585 attributes, whereas the 2-gram representation achieves a 0.93. In the case of the SEZ alphabet, an accuracy of 0.943 was obtained with this filtered 1,2,3-gram representation, as compared to 0.926 with the 2-gram representation. 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 0.92.

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 selection algorithm to these subsets. 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. Forward selection

Table 4: t-test subset selection

	AA		SEZ		DAV	
SIGNIF	**N**	**ACC**	**N**	**ACC**	**N**	**ACC**
20	1	0.37	2	0.5	0	-
19	15	0.88	8	0.77	10	0.83
18	49	0.931	39	0.9	23	0.88
17	105	0.933	79	0.922	58	0.91
16	212	0.937	149	0.93	99	0.92
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	0.93
12	1284	0.942	633	0.94	429	0.927

was quite successful at reducing the number of attributes while retaining an accuracy of approximately 0.94 in all three cases.

Table 5: Forward selection on 12- t-test subsets

AA			SEZ			DAV		
FEAT	**N**	**ACC**	**FEAT**	**N**	**ACC**	**FEAT**	**N**	**ACC**
1284	49	0.939	633	59	0.939	429	60	0.94

4.5 Discussion

N-grams and Feature Selection The experimental results have shown the interest of using feature selection: data dimensionality can be notably reduced without compromising classification quality. Forward selection has been shown to be an effective method, although is computationally too costly when the size of the feature set increases. In this situation, a fast univariate t-test filtering method becomes an appropriate solution to reduce the feature candidate set as a preprocessing step of the forward selection algorithm.

Analysis of t-test values An analysis of the t-test values (hypothesis value and *p*-value) allows measuring to what degree a 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 are also discriminative.

The analysis of the test values of the reduced feature set of the AA alphabet (See Table 5: 49 features: 33 3-grams, 13 2-grams, 3 1-grams) shows that the 3-grams CSL, ITF and FSM are the most significantly discriminative. In particular, CSL is the most significant one according to the t-test values of 20 two-sample tests. This feature was found not to be significant only for the mG-Ph discrimination.

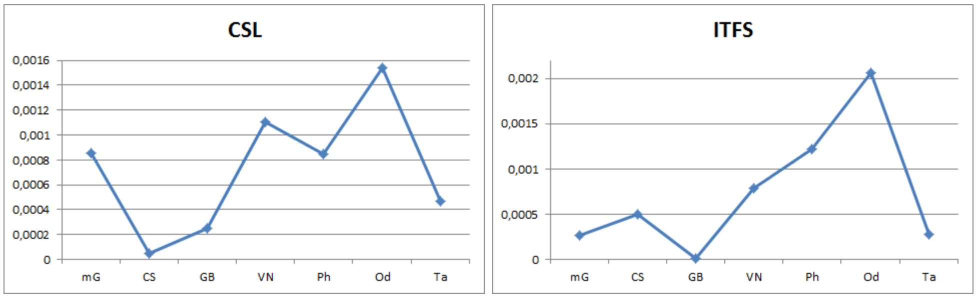

Fig. 1: Mean values of the CSL and ITFS n-gram features for the 7 class C GPCR subtypes.

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-Ta and CS-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-GB, mG-Ta and GB-Ta). The FSMLI 5-gram was also found to be significant for 15 tests. Figure 1 shows the mean values of n-gram features CSL and ITFS for the 7 class C GPCR subtypes.

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 amino acid alphabets. This is followed by dimensionality reduction through combination of a two-sample t-test and forward feature selection, as a preprocessing step prior to classification with SVMs. Reduced sets of n-grams that yield similar classification accuracies have been found for each of the three transformation alphabets.

The analysis of the features of the AA alphabet using the values obtained in the t-tests has provided insight about the n-grams that are best at discriminating between the GPCR subtypes. This might be considered as preliminary evidence of the existence of subtype-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.

Acknowledgments

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

References

1. C. Caragea, A. Silvescu, P. Mitra, P. (2011), Protein Sequence Classification Using Feature Hashing., In *Bioinformatics and Biomedicine (BIBM),2011 IEEE International Conference on*, pp. 538–543, IEEE, 2011
2. C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.
3. 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.
4. M. Can Cobanoglu, Y.l Saygin, and U. Sezerman. Classification of GPCRs Using Family Specific Motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1495–1508, 2011.
5. M. N. Davies, A. Secker, A. Freitas, E. Clark, J. Timmis, and D. R. Flower, Optimizing amino acid groupings for GPCR classification. *Bioinformatics 2008*, 24(18), pp. 1980–1986, 2008
6. 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.
7. J. Kittler. Feature Set Search Algorithms. *Pattern Recognition and Signal Processing* , C.H. Chen, ed., pp. 41–60. Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1978.
8. 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*, Springer, vol. 8158, pp. 336–343, 2013.
9. F. Mhamdi, M. Elloumi, R. Rakotomalala: Textmining, features selection and datamining for proteins classification. In *Information and Comunication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on.*, pp.457 – 458, IEEE, 2004
10. Y. Saeys, I. Inza, and P. Larrañaga, A review of feature selection techniques in bioinformatics. *Bioinformatics 2007*, 23(19), pp. 2507–2517, 2007
11. B. Trzaskowski,D. Latek,S. Yuan,U. Ghoshdastider,A. Debinski, et al. (2012) Action of molecular switches in GPCRs– theoretical and experimental studies. *Current medicinal chemistry* 19(8): 1090–1109.
12. 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.