

# Comparative Analysis of Alignment-based and Alignment-free Protein Classifiers

Pooja K. Strobe <sup>1</sup>, Etsuko N. Moriyama <sup>2</sup>

**Keywords:** protein classification, SVM, amino acid composition

## 1 Introduction

A large amount of new protein sequences is being accumulated rapidly in various databases. It is therefore important to develop fast and efficient methods to classify these proteins into their functional families. Many existing protein classification methods rely on multiple alignments. Generating reliable multiple alignments becomes problematic when dealing with extremely diverged protein sequences as the G-protein coupled receptors (GPCRs). GPCRs are a superfamily of cell membrane proteins that have seven transmembrane regions. Their classification and functional annotation is important in today's medical and pharmaceutical research because GPCRs play key roles in many human diseases. However, due to the high level of sequence divergence found among the GPCR family members, identifying and classifying these membrane proteins turns out to be a difficult task.

While the classification methods specific to the GPCR superfamily was the focus of several recent studies, there has been only few comparative performance analysis. In this study, we compared methods that use multiple alignments with those that do not, and methods that use both negative and positive data for learning with those that only use positive data. Furthermore, there has been little study for using simply the amino acid composition for protein classification. Therefore, we examined the use of amino acid frequencies with various pattern recognition methods and compared their classification performance for the GPCR superfamily.

## 2 Materials and Methods

The protein classification methods (classifiers) included for the performance analysis are: the profile-HMM (using SAM), support vector machines (SVMs) using amino acid compositions, SVM-pairwise developed by Liao et al. [2], SVM-Fisher developed by Karchin et al. [1], and decision trees (DT). Separate classifiers were built using sequences from Class A and those from non-Class A of the GPCR superfamily. The classifiers were tested to identify sequences belonging to the same class (Class A or non-Class A; within-class test) they were trained on (training and test sets were independent to each other). They were also tested to identify sequences belonging to the different class than the training set (between-class test). This second test could show if the methods were able to identify novel GPCRs that were very different from what the classifier has been trained with. Finally, the classifiers were tested to identify short subsequences to see how well the methods work to identify GPCRs when full sequences are not available. The classification performance was analyzed based on the

---

<sup>1</sup>School of Biological Sciences, University of Nebraska, Lincoln, NE 68588-0660, E-mail: [pkhati@cse.unl.edu](mailto:pkhati@cse.unl.edu)

<sup>2</sup>School of Biological Sciences & Plant Science Initiative, University of Nebraska, Lincoln, NE 68588-0660, E-mail: [emoriyama2@unlnotes.unl.edu](mailto:emoriyama2@unlnotes.unl.edu)

accuracy rate, cross validation test, minimum error point calculation, and maximum and median rates of false positives.

### 3 Results

Our results showed that SVM classifiers using amino acid composition have strong advantages over alignment-based classifiers. Using the SVM with amino acid composition (SVM\_AA), as shown in Table 2 the accuracy was better than all the other four methods for between-class tests. For within-class tests, the accuracy was still high for SVM\_AA (Table 1). Using SVMs with amino acid composition is simple yet efficient. It has a potential for discovering diverged novel GPCR candidates even from fragments. This simple classification strategy can be easily applied for general protein families.

Methods	FP	FN	Accuracy	Sensitivity	Specificity
SAM	0	2	1	0.99	1
SVM_Fisher	0	2	1	0.99	1
SVM_pairwise	1	2	0.99	0.99	1
SVM_AA	9	3	0.97	0.99	0.96
DT	14	12	0.94	0.94	0.93

Table 1: Classifier performance for within-class tests. The within-class tests used two independent Class A datasets, one for training and one for tests. FP is the number of false positives. FN is the number of false negatives.

Methods	FP	FN	Accuracy	Sensitivity	Specificity
SAM	0	150	0.60	0.07	1
SVM_Fisher	0	114	0.69	0.30	1
SVM_pairwise	1	107	0.71	0.34	1
SVM_AA	9	49	0.84	0.70	0.96
DT	14	75	0.76	0.54	0.93

Table 2: Classifier performance for between-class tests. The between-class tests used the Class A dataset for training and non-Class A dataset for tests. FP is the number of false positives. FN is the number of false negatives.

## References

- [1] Karchin, R., Karplus, K. and Haussler D. 2002. Karchin, R., Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18(1):147-159.
- [2] Liao, L. and Noble, W. S. 2003. Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships. *Journal of Computational Biology* 13:559-560.