Functional Classification of G-Protein Coupled Receptors, Based on Their Specific Ligand Coupling Patterns

Burcu Bakir¹ and Osman Ugur Sezerman²

¹ School of Biology, Georgia Institute of Technology, Atlanta, USA
² Sabanci University, Istanbul, Turkey

Abstract. Functional identification of G-Protein Coupled Receptors (GPCRs) is one of the current focus areas of pharmaceutical research. Although thousands of GPCR sequences are known, many of them remain as orphan sequences (the activating ligand is unknown). Therefore, classification methods for automated characterization of orphan GPCRs are imperative. In this study, for predicting Level 2 subfamilies of Amine GPCRs, a novel method for obtaining fixed-length feature vectors, based on the existence of activating ligand specific patterns, has been developed and utilized for a Support Vector Machine (SVM)-based classification. Exploiting the fact that there is a non-promiscuous relationship between the specific binding of GPCRs into their ligands and their functional classification, our method classifies Level 2 subfamilies of Amine GPCRs with a high predictive accuracy of 97.02% in a ten-fold cross validation test. The presented machine learning approach, bridges the gulf between the excess amount of GPCR sequence data and their poor functional characterization.

1 Introduction

G-Protein Coupled Receptors (GPCRs) are vital protein bundles with their key role in cellular signaling and regulation of various basic physiological processes. With their versatile functions in a wide range of physiological cellular conditions, they constitute one of the vastest families of eukaryotic transmembrane proteins [29]. In addition to the biological importance of their functional roles, their interaction with more than 50% of prescription drugs have lead GPCRs to be an excellent potential therapeutic target class for drug design and current pharmaceutical research. Over the last 20 years, several hundred new drugs have been registered which are directed towards modulating more than 20 different GPCRs, and approximately 40% of the top 200 synthetic drugs act on GPCRs [6]. Therefore, many pharmaceutical companies are involved in carrying out research aimed towards understanding the structure and function of these GPCR proteins. Even though thousands of GPCR sequences are known as a result of ongoing genomics projects [10], the crystal structure has been solved only for one GPCR sequence using electron diffraction at medium resolution (2.8
A) to date [15] and for many of the GPCRs the activating ligand is unknown, which are called orphan GPCRs [25]. Hence, based on sequence information, a functional classification method of those orphan GPCRs and new upcoming GPCR sequences is of great practical use in facilitating the identification and characterization of novel GPCRs.

Albeit laboratory experiments are the most reliable methods, they are not cost and labour effective. To automate the process, computational methods such as decision trees, discriminant analysis, neural networks and support vector machines (SVMs), have been extensively used in the fields of classification of biological data [21]. Among these methods, SVMs give best prediction performance, when applied to many real-life classification problems, including biological issues [30]. One of the most critical issues in classification is the minimization of the probability of error on test data using the trained classifier, which is also known as structural risk minimization. It has been demonstrated that SVMs are able to minimize the structural risk through finding a unique hyper-plane with maximum margin to separate data from two classes [27]. Therefore, compared with the other classification methods, SVM classifiers supply the best generalization ability on unseen data [30].

In the current literature, to classify GPCRs in different levels of families, there exist different attempts, such as using primary database search tools, e.g., BLAST [1], FASTA [20]. However, these methods require the query protein to be significantly similar to the database sequences in order to work properly. In addition to these database search tools, the same problem is addressed by using secondary database methods (profiles and patterns for classification), e.g., Attwood et al. have worked in particular on GPCRs in the PRINTS database [2] (whose data appeared in INTERPRO database [17]). Hidden Markov Models [24], bagging classification trees [32] and SVMs [13], [31] are other methods that have been used to classify GPCRs in different levels of families. Karchin et al. conducted the most comprehensive controlled experiments for sequence based prediction of GPCRs in [13] and showed that SVMs gave the highest accuracy in recognizing GPCR families. Whereas, in SVMs, an initial step to transform each protein sequence into a fixed-length vector is required and the predictive accuracy of SVMs significantly depends on this particular fixed-length vector. In [13], it is also pointed out that the SVM performance could be further increased by using feature vectors that encode only the most relevant features, since SVMs do not identify the features most responsible for class discrimination. Therefore, for an accurate SVM classification, feature vectors should reflect the unique biological information contained in sequences, which is specific to the type of classification problem.

In this paper, we address Level 2 subfamily classification of Amine GPCRs problem by applying Support Vector Machine (SVM) technique, using a novel fixed-length feature vector, based on the existence of activating ligand specific patterns. We obtain discriminative feature vectors by utilizing biological knowledge of the Level 2 subfamilies’ transmembrane topology and identifying specific patterns for each Level 2 subfamily. Since these specific patterns carry ligand