A LINGUISTIC APPROACH TO THE IDENTIFICATION OF MOTIFS AND PHARMACEUTICAL CLASSIFICATION OF GPCRs

ABSTRACT:

Background: The superfamily of G-protein coupled receptors (GPCR) is the target of approximately 60% of current drugs in the market. Disruption in their regulation can cause diseases such as cancer, cardiovascular disease, Alzheimer's and Parkinson's diseases, stroke, diabetes, and inflammatory and respiratory diseases. GPCRs are classified into subfamilies by their pharmaceutical properties. Due to their low sequence similarity, traditional alignment-based approaches to classification have had limited success on GPCRs at the subfamily levels. A recent study tested BLAST, k-nearest neighbors, hidden Markov models (HMM), and support vector machines (SVM) with alignment-based features on subfamily classification of GPCRs, and concluded that the highly complex support vector machines performed the best.

Technology: In this study, we applied a popular approach to document classification in language technologies research to GPCR subfamily classification. Here, we viewed each protein as a "document" where the "words" are all contiguous sequences of 1 to 4 amino acids. As in document classification, we employed a feature selection algorithm to select the most important "words" for our task and applied a classifier on the counts of those selected "words".

Design: For our task, we have chosen two very simple classifiers, the decision tree and the Näive Bayes classifier, because they allow easy interpretation of the reasons behind their predictions. Moreover, to the best of our knowledge, these classifiers are simpler than any that have been attempted on protein classification. We employed chi-square feature selection which has been shown to be the most successful feature selection method in document classification.

Results: Using the same dataset and training and testing protocol as in the study on SVM, we found our Näive Bayes classifier surpassing the reported accuracy of the SVM by 4.8% in level I subfamily classification and by 6.1% in level II subfamily classification. Our decision tree classifier, while inferior to the SVM, still outperforms the reported accuracy of HMM in both level I and II subfamily classification. More importantly, the "words" chosen by our feature selection method correlate with motifs that have previously been identified in wet lab experiments.

Conclusion: Using a language technologies approach, we have developed a classifier that is much simpler but more accurate than the traditional protein classifiers in the pharmacology-based GPCR subfamily classification. In addition, our method identifies motifs that have been conserved at the subfamily level and may be potential target sites for future drugs.