Abstract

IDENTIFYING IMPORTANT "WORDS" IN THE LANGUAGE OF PROTEINS. Betty Cheng[1], Judith Klein-Seetharaman[2], Jaime Carbonell[1]; [1]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA; [2]University of Pittsburgh, Biomedical Science Tower E1355, Pittsburgh, PA

DNA and protein sequences are often referred to as “biological language” and linguistic metaphors are used frequently to describe biological phenomena. In particular, the typical bioinformatics problems of classifying proteins into a hierarchy of superfamilies, families and subfamilies and detecting motifs are analogous to the problem of document classification in language technology research. Using the language analogy, we developed a novel method of motif detection based on family hierarchy classification. A popular approach in document classification is the application of a machine learning classifier on the counts of words in a document. The choice of the classifier depends on the nature of the dataset. However, no classifier can handle the large number of features the count of each unique word in the document set will generate, nor is it necessary. Instead, a range of feature selection methods have been developed to select the most important keywords for the specific task, of which chi-square has been determined to be the most successful for the classification task (1). In protein classification, each protein sequence is viewed as a “document” where a “word” is an n-gram, a short sequence of n amino acids. Previously, a range of classifiers of varying complexity from k nearest neighbours to support vector machines (SVM) have been tested on the classification of G-protein coupled receptors (GPCR) at the superfamily and subfamily levels (2). The GPCR superfamily is an important dataset for protein classification because they are the target of approximately 60% of current drugs (3). Its classification is difficult due to the extreme diversity among its members (4). It was shown that the more complex classifier SVM performed better than the other classifiers at the subfamily level classification (2). Here, we show that the much simpler Naïve Bayes classifier outperforms the SVM, given the counts of the important “words” as determined by chi-square. Moreover, these important “keywords” correlate with motifs previously identified through wet-lab experiments; thus, providing us with a method to detect motifs conserved at each level in the classification hierarchy. References: 1. Yang and Pedersen (1997), International Conference on Machine Learning, p. 412-420 2. Karchin et al. (2002), Bioinformatics 18(1):147-159 3. Muller (2000), Current Medical Chemistry, 7(9):861-888 4. Moriyama and Kim (2003), Stadler Genetics Symposium