COMS4761: FALL 2007

ASSIGNMENT 5.3: Classifying Proteins With SVM

Problem 1: Classifying Proteins

Consider the ZPCR transmembrane eZproteins of assignment A4.4, depicted below.

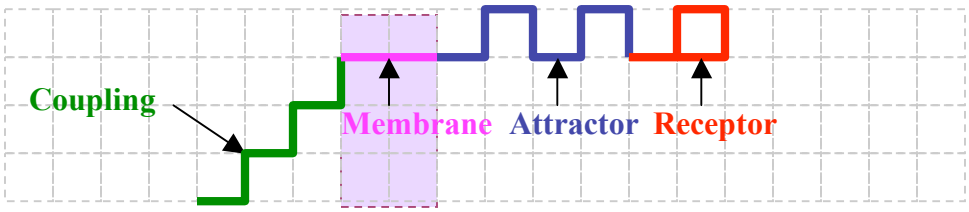

Figure 1: Architecture of ZPCR

It was discovered that ZPCR proteins are the only ezProteins to have the following tail receptor motif: a short Train followed by a single Globulin. The goal of this problem is to construct an SVM that exploits this motif to classify proteins as ZPCR or not ZCPR.

The following tail sequences samples of ezProteins were classified manually:
ZPCR ={bbaabbabbb, cbaabbcbcb, cbaabbacbb, bbaaccabbb}
Not ZPCR={bbaabbaabb, ababcbccac, bbaaccaccc}

(i) Convert the sequences above to a vector space representation by assigning to ezAA numerical values as follows: a=1, b=-1, c=0, d=2. For example, the sequence abbc is uniquely represented by the vector (1,-1,-1,0). Construct an SVM classifier for ZPCR as follows:
A. Reduce dimensionality: Determine dimensions (sequence positions) that do not contribute to the variance between the classes and eliminate them.
B. Train an SVM: Use the sample above to train an SVM with a linear kernel to classify ZPCR.

(ii) Is the following ezProtein a ZPCR: **abcbdddabbaabcabcb**? [Apply the SVM to the tail sequence (10 last ezAA)]

(iii) Now consider a different numerical representation of the sequences in a 4 dimensional space; for a sequence S define the vector representation (A,B,C,D) as follows: A= # of occurrences of a in S, B=#of occurrences of b in S, C=# of occurrences of c in S, D=#of occurrences of d in S. For example, for S=ababcb the vector representation is (2,3,1,0). Can the sequences above be linearly classified using this vectorial representation? Can you find examples of sequences that could not be correctly classified?