Screening genomes for transmembrane barrel proteins

Andrew Garrow* and David R. Westhead

University of Leeds

Transmembrane barrel (TMB) proteins are a functionally important and diverse group of proteins found in the outer membranes of Gram negative bacteria, mitochondria and chloroplasts. Unlike with alpha-helical transmembrane proteins, development of computational strategies to screen sequence sets for TMB protein coding sequences has proved difficult owing to a short and cryptic transmembrane strand motif (Garrow, Agnew, and Westhead, 2005). We have developed a consensus strategy that classifies according to composition (both 20-D amino acid and 400-D pair-ordered amino acid), using both instance based k-nearest neighbor and scoring matrices based algorithms. In total 12 separate algorithms have been developed, of which results are fed into a support vector machine (SVM) for a consensus prediction. Individual and consensus program outputs are calibrated with log-likelihood ratio (LLR) scores giving the probabilities of queries representing TMB or non-TMB proteins, with positive scores for TMB proteins and vice versa.

The consensus algorithm is $>$94% accurate as tested with a rigorous cross-validation procedure. Unfortunately problems will still occur with genome screening, with chance occurrence meaning that many non-TMB proteins will get positive LLR scores owing to the large numbers of sequences tested (e.g. the Escherichia coli contains $>$5000 sequences), of which TMB proteins only constitute a small fraction (2.5% of proteins in E.coli are predicted to be TMB proteins). For this reason expectancy (E) value statistics are also being developed to account for the numbers of queries made. We present our prediction algorithm, show how it is calibrated and demonstrate scoring statistics.

Reference

Garrow, A.G., Agnew, A., Westhead, D.R. (2005) TMB-Hunt: An amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins, *BMC Bioinformatics*, **6(56)**.