BMC Bioinformatics

Poster presentation

Distinguishing between enzyme sequences & non-enzymes without using alignments

Ajanthah Sangaralingam* and Andrew J Doig

Address: Department of Biomolecular Sciences, Manchester University, UK.

Email: Ajanthah Sangaralingam* - A.Sangaralingam@postgrad.manchester.ac.uk

* Corresponding author

from BioSysBio: Bioinformatics and Systems Biology Conference
Edinburgh, UK, 14–15 July 2005

Published: 21 September 2005

BMC Bioinformatics 2005, **6**(Suppl 3):P24

Many protein similarity searching methods rely upon finding sequence homology to a previously annotated protein within a protein database. If a protein has no sequence similarity or only weak sequence similarity to an annotated protein, however, then the task of predicting protein function is often not possible. In this study, support vector machines are used to develop a method not based on sequence alignment to predict whether a protein sequence is an enzyme or a non-enzyme, using features that are calculated from sequences.

A large non-redundant dataset was constructed from the SWISSPROT database. Only sequences that have an EC number and entries in the ENZYME database were included in the enzyme dataset. Sequence annotation was filtered carefully, when selecting the non-enzyme dataset. Sequences for which annotation was sparse, annotated as unknown, probable, homologue, hypothetical and enzyme were excluded.

Features used to describe each protein sequence were: hits to the INTERPRO sequence motif database, sequence length, amino acid frequencies and minimum distances between amino acid pairs. Each protein sequence was represented as an 11429 feature vector. Using all of the features gave a prediction accuracy of 89%. Feature selection was performed to analyse which features were most informative in discriminating between the two classes. 3177 INTERPRO motifs were found in the sequences used in this study. Of these, 1326 were found in only a single sequence. These motifs were therefore not used as features in further experiments. Leaving out these motifs had little effect on the prediction accuracy of the model. Overall, accuracy was improved when only amino acid frequencies and INTERPRO motifs were used as features for model building. This model was built using 1871 features. We plan to apply this model to predict the function of sequences that are annotated as unknown in the SWISS-PROT database.