Poster H-22

Protein solubility assessment using support vector machine

Authors:
Pawel Smialowski (*Technical University Muenchen*)
Antonio Martin-Galiano (*Technical University Muenchen*)
Tobias Girschick (*Technical University Muenchen*)
Dmitrij Frishman (*Technical University Muenchen*)

Short Abstract: We propose a machine-learning approach to sequence-based prediction of protein solubility in which we exploit subtle differences between proteins reported to be soluble and insoluble in TargetDB and PDB databases as well as in literature accounts. An overall prediction accuracy of 73% was achieved in a 10-fold cross-validation experiment.

Long Abstract:
We propose a machine-learning approach to sequence-based prediction of protein solubility in which we exploit subtle differences between proteins reported to be soluble and insoluble in TargetDB and PDB databases as well as in literature accounts. The length distribution of the soluble and insoluble datasets was adjusted to avoid predictions biased by protein size. As feature space for classification, we used frequencies of mono-, di-, and tri-peptides represented by the original 20-letter amino acid alphabet as well as by several reduced alphabets in which amino acids were grouped by their physicochemical and structural properties. The classification algorithm was constructed as a two-layered structure in which the output of primary support vector machine classifiers operating on peptide frequencies was combined by a second-level Naive Bayes classifier. An overall prediction accuracy of 73% (61% on the positive (soluble) and 84% on the negative (insoluble) class) was achieved in a 10-fold cross-validation experiment over the sequence set which was made non-redundant at the 50% identity threshold. This indicates that the proposed algorithm may be a valuable tool for more efficient target selection in structural genomics. A Web server for protein solubility prediction is available at http://webclu.bio.wzw.tum.de:8080/proso