Protein Sequences Classification Based on String Weighting Scheme

N. M. Zaki [1], **S. Deris** [1], and **R. M. Illias** [2]

[1] Department of Software Engineering,
Faculty of Computer Science & Information Technology,
[2] Department of Bioprocess Engineering,
Faculty of Chemical and Natural Recourses Engineering,

University Technology Malaysia,
Skudai 81300, Johor, Malaysia
Emails: nazar@siswa.utm.my; safaai@fsksm.utm.my; i.rosli@fkkksa.utm.my

Abstract

We present a new technique to recognize remote protein homologies that rely on combining probabilistic modeling and supervised learning in high-dimensional feature spaces. The main novelty of our technique is the method of constructing feature vectors using Hidden Markov Model and the combination of this representation with a classifier capable of learning in very sparse high-dimensional spaces. Each feature vector records the sensitivity of each protein domain to a previously learned set of sub-sequences (strings). Unlike other previous methods, our method takes in consideration the conserved and non-conserved regions. The system subsequently utilizes Support Vector Machines (SVM) classifiers to learn the boundaries between structural protein classes. Experiments show that this method, which we call the String Weighting Scheme-SVM (SWS-SVM) method, significantly improves on previous methods for the classification of protein domains based on remote homologies. Our method is then compared to five existing homology detection methods.