PROTEIN FOLD RECOGNITION USING THE GRADIENT BOOST ALGORITHM

Feng Jiao
School of Computer Science, University of Waterloo, Canada

Jinbo Xuy
Toyota Technological Institute at Chicago, USA

Libo Yu
Bioinformatics Solutions Inc., Waterloo, Canada

Dale Schuurmans
Department of Computing Science, University of Alberta, Canada

Protein structure prediction is one of the most important and difficult problems in computational molecular biology. Protein threading represents one of the most promising techniques for this problem. One of the critical steps in protein threading, called fold recognition, is to choose the best-fit template for the query protein with the structure to be predicted. The standard method for template selection is to rank candidates according to the z-score of the sequence-template alignment. However, the z-score calculation is time-consuming, which greatly hinders structure prediction at a genome scale. In this paper, we present a machine learning approach that treats the fold recognition problem as a regression task and uses a least-squares boosting algorithm (LS Boost) to solve it efficiently. We test our method on Lindahl's benchmark and compare it with other methods. According to our experimental results we can draw the conclusions that: (1) Machine learning techniques offer an effective way to solve the fold recognition problem. (2) Formulating protein fold recognition as a regression rather than a classification problem leads to a more effective outcome. (3) Importantly, the LS Boost algorithm does not require the calculation of the z-score as an input, and therefore can obtain significant computational savings over standard approaches. (4) The LS Boost algorithm obtains superior accuracy, with less computation for both training and testing, than alternative machine learning approaches such as SVMs and neural networks, which also need not calculate the z-score. Finally, by using the LS Boost algorithm, one can identify important features in the fold recognition protocol, something that cannot be done using a straightforward SVM approach.