Single Sequence Secondary Structure Prediction for Globular Proteins

Ashwin Subramani* and Christodoulos A. Floudas
Department of Chemical Engineering, Princeton University

Poster (2:20 PM)

Secondary structure prediction of a protein is an important intermediate step in the three dimensional structure prediction of a protein. The importance is particularly brought into focus in first principles based approaches to protein structure prediction, which do not use any database information in the form of homology information [1]. The most common techniques for secondary structure prediction feature a 3 – class secondary structure state for each residue of a protein [2,3].

We have developed an optimization – based method to predict the secondary structure of a target protein without the use of profile information. The model combines two models, an α – helix prediction model HELIOS (HELical prediction using Integer Optimization approacheS) and a β – strand prediction prediction model BEST-PRED (BEta STrand PREDiction). For α – helix prediction, a two – stage infeasibility minimization problem has been introduced. The first stage is a linear programming (LP) model for parameter estimation, while the second stage is an integer programming (ILP) model for helix prediction. The residues of a target protein are divided into 4 regions depending on their putative proximity to the helix termini, and propensity to be in helices is compared to a pre – evaluated residue – dependent threshold propensity, using overlapping nonapeptides surrounding the central residue. BEST-PRED for β – strand prediction has been introduced as an integer programming (ILP) model, which maximizes a residue's propensity to be in a β – strand. The protein is divided into overlapping pentapeptides. The β – strand propensity weight for the central residue is evaluated by implementing a novel combination of Naïve – Bayesian and first order Markov models, which represent the physical nature of a β – strand. In both models, important mathematical constraints are introduced to ensure that biologically meaningful results are output as results. These constraints model the physical nature of the residues, along with the minimum and maximum secondary structure content [4]. Further, the formulation allows the user to add any form of prior knowledge about the secondary structure easily. This method was tested on a set of α, β and mixed α- β proteins. A Q_α accuracy of 82% was seen for purely α – helical proteins using HELIOS, while a Q_β accuracy of 78.9% was seen for purely β – proteins using BEST-PRED. These results compare very favorably with some of the standard secondary structure prediction servers.

[1] Klepeis JL and Floudas CA, *BioPhysical J.*, 85, 2119 – 2146 (2003).
[2] Altschul SF *et al.*, *Nuc. Acids Res.*, 25, 3389 – 3402 (1997)
[3] McGuffin LJ *et al.*, *BioInformatics*, 16, 404 – 405 (2000)
[4] Homaeian L *et al.*, *Proteins*, 69, 486 – 498 (2007)