

Prediction of Disordered regions in proteins using Amino-acid properties, Secondary Structure and Relative Solvent Accessibility

Karthikeyan Swaminathan¹ and Jarosław Meller^{1,2}

¹Division of Biomedical Informatics, Children's Hospital Research Foundation,
3333 Burnet Avenue, Cincinnati, OH 45242, USA

²Department of Informatics, Nicolaus Copernicus University,
Grudziądzka 5, 87100 Toruń, Poland.

Abstract

Many proteins or protein domains do not adopt stable three-dimensional structures. These disordered regions in proteins have been studied in the past and have been implicated in many protein-protein and protein-DNA interactions, in cell signaling and regulation in many signal-transduction pathways and in a number of diseases including Huntington's, Alzheimer's and other neurological conditions. Experimental methods to determine disordered regions such as Nuclear Magnetic Resonance (NMR) and Circular Dichroism (CD) are time consuming and have other limitations. This provides a great opportunity for computational studies on disorder in proteins.

Many prediction methods have been developed for this problem. Examples include Uversky et al who made predictions using hydrophobicity and net charge profiles of the amino-acid sequences. Jones et al devised the DISOPRED server that uses the position specific scoring matrices (PSSMs) of sequences as features and employs neural networks method of classification. They also used secondary structure predictions to filter false-positives resulting from the disorder predictions. Linding et al designed a neural network predictor called DisEMBL that is based on the hypothesis that disordered regions include segments in the protein with have a coiled secondary structure and have high temperature factors. Dunker et al have developed a number of predictors (latest called VSL1), which employ neural networks and use features like hydrophobicity, charge, frequency of amino acids, sequence complexity, entropy and secondary structure predictions.

In this work, we present a novel method for prediction of disordered regions in proteins. The new method is based on the hypothesis that enhanced secondary structure and relative solvent accessibility (SS & RSA) predictions can significantly improve the current state of the art. SS & RSA predictions are obtained from the recently developed SABLE server that was shown to achieve state of the art accuracies in independent tests. Our predictor makes use of Support Vector Machines (SVMs) for the classification of protein regions into ordered or disordered. Results show that SS & RSA play a significant role in predicting disordered regions, supporting our hypothesis. We intend to make this predictor available for free to academic researchers and also intend to submit it to the upcoming CASP7 experiment (April 2006).

Keywords: protein disorder, protein structure prediction, solvent accessibility, SABLE, machine learning.