

Predicting Protein Disorder using a Neural Network and Amino Acid Hydropathy Values

Deborah Stoffer

Protein structure clearly plays an integral role in protein function yet many proteins do not adopt the 3D structures that can be ascertained through crystallography. Disordered proteins or intrinsically unstructured proteins are just two of the terms typically used to refer to this class of proteins. This research focuses on the creation of computational tools for prediction and analysis of disordered proteins. The experimental techniques X-ray crystallography, NMR spectroscopy, circular dichroism spectroscopy, protease digestion, and Stoke's radius determination can be used to identify disorder in proteins; however, these methods are quite costly and time consuming. Computational prediction tools have been developed to accurately predict protein disorder using a variety of computational intelligence and machine learning techniques such as artificial neural networks (ANNs), support vector machines (SVMs), logistic regression, and discriminant analysis. These computational tools effectively help to speed up the process of identifying disordered proteins. The use of such prediction tools can help to identify potential drug targets and aid the protein chemists in first line analysis of protein chemistry.

This paper describes a novel disorder prediction tool that utilizes only hydropathy information compiled from amino acid sequence data. The hydropathy information obtained from the proteins is used as attributes for training a feed-forward artificial neural network (ANN). The novel feature of this technique is that the hydropathy information is compiled using different sized windows surrounding each amino acid in the sequence in an effort to capture both local and long-range hydropathy characteristics. The predictor was trained and tested using datasets containing both disordered and ordered proteins. The disordered proteins were obtained from the Database of Protein Disorder (DisProt), a curated database developed and maintained jointly between the Indiana University School of Medicine and Temple University. Ordered proteins were obtained from the RCSB Protein Data Bank (PDB) from entries for which structures had been determined using X-ray diffraction. Five subsets were created for use in a five-fold cross-validation experimental design. The attributes used as input into the predictor were calculated by averaging amino acid hydropathy values over five different window sizes where two separate windowing techniques were used, overlapping windows included the previous windows in the next larger window, and gapped windows excluded the previous windows from the next larger window.

A feed-forward ANN was trained and tested using hydropathy attributes calculated with overlapping windows and then with gapped windows. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were determined and used in the standard way to calculate sensitivity and specificity values. Results were obtained for disorder prediction as measured at the individual residue level and at the level of entire protein domains. The observed accuracy of our hydropathy attributes trained predictor are competitive with other disorder predictors that used more complex combinations of attributes for training and prediction. Other analysis conducted on the data included Matthew's Correlation Coefficient (MCC), the confidence values of the ANN for the predictions, and a T-test on the per-residue and per-protein predictions.