Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of protein mutants

Michael Fernández, Julio Caballero, Leyden Fernández, José Ignacio Abreu

a. Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba.
b. Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba.

Abstract

Development of novel computational approaches for modeling protein properties from their primary structure is a main goal in applied Proteomics. 1D Autocorrelation vectors are widely applied to encoded chemical structural information in structure-property/activity studies. We reported the extension of the autocorrelation vector formalism to amino acid sequences for encoding protein structural information with modeling purposes. Amino Acid Sequence Autocorrelation (AASA) vectors are calculated by measure the autocorrelations of 48 amino acid/residue properties selected from the AAindex data base at sequence lags ranging from 1 to 15 on the protein primary structure. A total of 720 AASA descriptors were tested for building predictive models of the thermal unfolding Gibbs free energy change of protein mutants. Ensembles of Regularized Genetic Neural Networks (BRGNNs) were used for obtaining optimum nonlinear models for the conformational stability. In the context of neural network modeling of biological interactions we introduced BRGNNs as a robust nonlinear modeling technique that combines Genetic Algorithm and Bayesian regularization for neural network input selection and supervised network training, respectively. This approach attempts to solve the main weaknesses of neural network modeling: the selection of optimum input variables and the adjustment of network weights and biases to optimum values for yielding regularized neural network predictors. The ensemble predictors described about 88% and 68% variance of the data in training and test sets respectively. Models exhibited a high contribution on the conformational stability of the distributions in the protein sequence of entropy, hydrophobicity and secondary structure-related properties. Furthermore, optimum AASA vector subsets not only showed to successfully model unfolding thermal stability but well distribute wild-type and mutant proteins on a stability Self-organized Map (SOM) when used for unsupervised training of competitive neurons.

References