Prediction of Protein Solubility in *Escherichia Coli* Using Discriminant Analysis, Logistic Regression, and Artificial Neural Network Models

Reese Lennarson, Rex Richard, Miguel Bagajewicz and Roger Harrison

Abstract

Recombinant DNA technology is important in the mass production of proteins for academic, medical, and industrial use, and the prediction of the solubility of proteins is a significant part of it. However, the protein solubility when overexpressed in a host organism is difficult to predict. Thus, a model capable of accurately estimating the likelihood of proteins to form insoluble inclusion bodies would be highly useful in many applications, indicating whether proteins necessitate chaperones to remain soluble under the conditions within the host organism. To this end, solubility data for proteins when overexpressed in *Escherichia coli* was compiled, and properties of the proteins likely affecting solubility were identified as parameters for building solubility prediction models. In this paper, three models were constructed using discriminant analysis, logistic regression, and neural networks. Significant parameters were determined, and the efficiencies of solubility prediction for the three procedures were compared. Among the properties investigated, α-helix propensity and asparagine fraction were the most important parameters in the discriminant analysis model; for logistic regression, molecular weight, total number of hydrophobic residues, hydrophilicity index, approximate charge average, asparagine fraction, and tyrosine fraction were found to be the greatest contributors to protein solubility. For the neural network, the most important parameters included the asparagine fraction, total number of hydrophobic residues, and tyrosine fraction. The asparagine fraction was of great importance, as it was the only parameter found to be among the five most significant parameters in all three models. *Post hoc* evaluations of the models indicated that the discriminant analysis model was 66.5% accurate, the logistic regression model was 73.9% accurate, and the neural network model was 91.0% accurate. For the logistic regression model, *post hoc* accuracies were shown to increase as predictions of solubility or insolubility neared high probabilities. *A priori* evaluations were used to determine how well logistic regression and the neural network would predict solubility of new proteins. The discriminant analysis was excluded from this study because its *post hoc* accuracy was exceedingly low. These studies showed that the logistic regression models tended to give higher prediction accuracies than neural networks for proteins not previously used in creating the respective models, but logistic regression predictions were highly skewed toward insolubility, while neural network predictions were more balanced overall.