MACHINE LEARNING APPROACHES ON PREDICTION OF PROTEIN-PROTEIN INTERACTIONS AND FEATURE SELECTION WITH AN EYE TO MUTATIONAL ANALYSES

Angshuman Bagchi, Corey Powell, Sean D. Mooney*

Buck Institute for Age Research, 8001 Redwood Blvd
Novato, CA 94945 USA

*Email: smooney@buckinstitute.org
Email: abagchi@buckinstitute.org

Matthew Mort, David N. Cooper

Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

Eunseog Youn

Department of Computer Science,Texas Tech. University, Lubbock, Texas

Fuxiao Xin, Predrag Radivojac

School of Informatics and Computing, Indiana University, Bloomington, Indiana 47408

Protein-protein interactions are required in many important biological processes, for example, hormone–receptor binding, antigen-antibody interactions and so on. Evidences also suggest that disease-causing mutations disrupt protein interactions. In the present work, we used machine learning tools to discriminate between interface and non-interface residues of proteins using structure and sequence information. We generated features from protein sequence and structure using a non-redundant set of protein hetero-complexes from the protein data bank. The training dataset comprised of (A) interface residues and non-interface surface residues for structure based prediction and (B) interface residues and all other residues (viz., non-interface surface and core residues) for sequence based prediction. The datasets were used to build classification algorithms using random forest (RF) and support vector machine (SVM) coupled with 10 fold cross-validation for evaluation. Overall, RF outperformed SVM in most cases. The best performing sequence-based classification tool achieved an accuracy of 73% and when protein structure was included the accuracy was 70%. Protein structural flexibilities and sequence conservation were among the best classification features ranked on the basis of their area under the receiver operating characteristics (AUC) curves (ROC). We used our predictor to analyze disease causing mutations and observed enrichment of protein-protein interaction sites in disease situations. Overall, our results indicate that machine learning may be used to analyze disease mutations.

Acknowledgments

We would like to thank Prof. Philip E. Bourne of UCSD for providing us the dataset used for training. This research is supported by K22LM009135 (PI: Mooney), R01LM009722 (PI: Mooney), Grant from IU Biomedical Research Council, Indiana University, the Showalter Trust and the Indiana Genomics Initiative. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment.