Machine Learning Methods for Computational Proteomics and Beyond

Pierre Baldi

Institute for Genomics and Bioinformatics
University of California, Irvine, CA, USA

Abstract. Predicting protein structure is a fundamental problem in biology, especially in the genomic era where over one third of newly discovered genes have unknown structure and function. Because sequence and structure data (hence training sets) continue to grow exponentially, this area is ideally suited for machine learning approaches. Neural networks, in particular, have had remarkable success and have led, for instance, to the construction of the best secondary structure predictors. We will provide an overview of our own work and the state-of-the-art for several structure prediction problem including:

- (1) prediction of protein secondary structures;
- (2) prediction of relative solvent accessibility;
- (3) prediction of contacts;
- (4) prediction of three-dimensional protein structures;
- (5) prediction of interchain beta-sheet quaternary structures; using machine learning methods.

The methods we have developed are based on the theory of graphical models but use deterministic recursive neural networks to speed up learning. We will discuss their applicability to other problems and the lessons learnt for the design of complex neural architectures.