Generalized IOHMMs and Recurrent Neural Network Architectures

Pierre Baldi and Gianluca Pollastri
Department of Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 92697-3425
pfbaldi,gpollast@ics.uci.edu
(949) 824-5809
(949) 824-4056 (FAX)

Abstract

Accurate prediction of protein contact maps is an important step in computational structural proteomics. Because contact maps provide a translation and rotation invariant topological representation of a protein, they can be used as a fundamental intermediary step in protein structure prediction. We develop a new set of flexible machine learning architectures for the prediction of contact maps, as well as other information processing and pattern recognition tasks. The architectures can be viewed as recurrent neural network parameterizations of a class of Bayesian networks we call generalized input-output HMMs. As Bayesian networks, these architectures can be viewed as generalizations of input-output hidden Markov models (GIOHMMS), involving an input layer, an output layer, and a hidden layer supported by one or several directed acyclic graphs. Under stationarity assumptions, when the hidden graphs have a regular structure, these architectures can be approximated by recurrent neural network architectures that can be trained with the proper algorithms. This methodology is widely applicable and illustrated here on the problem of predicting contact maps in proteins yielding a state-of-the-art predictor. For the specific case of contact maps, contextual information is propagated laterally through four hidden planes, one for each cardinal corner. While several extensions and improvements are in progress, the current version can accurately predict 54.5% of contacts at a distance cutoff of 8$\mathring{A}$. The general methodology also suggests an important role for lateral processing of information in the hidden layers of learning systems. *Availability:* The contact map predictor will be made available through http://promoter.ics.uci.edu/BRNN-PRED/ as part of an existing suite of proteomics predictors.

Keywords: graphical models, bayesian networks, recurrent neural networks, protein structure prediction.