Statistical Machine Learning for Structural Bioinformatics

David L. Wild
University of Warwick, Coventry, UK
Email: D.L.Wild@warwick.ac.uk

[Joint work with Karsten Borgwardt, Alexei Podtelezhnikov, Chu Wei and Zoubin Ghahramani]

Abstract

Protein structure prediction from sequence is one of the central problems of structural bioinformatics. There are three main paradigms that are employed to address this problem. The first approach is homology or comparative modelling. The second is fold recognition, or threading, using sequence-structure compatibility between the sequence of interest and proteins with known three-dimensional structures. The third approach is ab initio molecular simulation guided by physical forces. Humans readily learn new concepts after observing a few examples and show extremely good generalization to new instances. In contrast, search tools on the internet exhibit little or no learning and generalization. Recent work on Bayesian Sets by Ghahramani and Heller (Advances in Neural Information Processing Systems (NIPS 2005), 2006) shows that infor- mation retrieval can be firmly grounded in a Bayesian statistical model of human learning and generalization. Given a set of items, the algorithm finds other items that belong to the same concept. For example, given Monday, Wednesday, it should return the days of the week; given three Jim Carrey movies, it should return other Jim Carrey movies; given a couple of proteins with some three dimensional fold, it should return other proteins with a similar fold. In the first part of my talk, I will describe an application of this approach to the problem of protein fold recognition, (recognizing proteins that have similar tertiary structures based on a database of known protein structures), which was recently cast in the framework of in- formation retrieval by Cheng and Baldi (Bioinformatics, 22, s1456s1463, 2006). On several fold recognition tasks, including benchmark datasets and recent CASP targets, Bayesian Sets allow us to accurately predict whether proteins belong to the same fold. Furthermore, our method shows a runtime that is vastly superior to that of existing state-of-the-art approaches. Interactions between amino acids define how proteins fold and function. In the second part of the talk I will focus on the search for adequate potentials that can distinguish the native fold from misfolded states in ab initio protein folding. Since direct measurements of these interactions are impossible, known native structures themselves have become the best experimental evidence. Traditionally, empirical knowledge-based statistical potentials were proposed to describe such interactions from an observed ensemble of known structures. This approach is based on the hypothesis of a Boltzmann distribution of distances and angles between the interacting atoms. I will describe an alternative approach, which uses a novel statistical machine learning methodology, called contrastive divergence, to learn the parameters of statistical potentials from data, thus inferring force constants, geometrical cut-offs and other structural parameters from known structures. Contrastive divergence is intertwined with an efficient Metropolis Monte Carlo procedure for sampling protein backbone conformations. Applications of this approach have included a study of protein backbone hydrogen bonding, which yields results which are in quantitative agreement with experimental characteristics of hydrogen bonds. From a consideration of the requirements for efficient and accurate reconstruction of secondary structural elements in the context of protein structure prediction, I will also demonstrate the applicability of the framework to the problem of reconstructing the overall protein fold for a number of commonly studied small proteins, based on only predicted secondary structure and contact map.