Poster I-96

Ab Initio Modelling of Protein C-alpha Traces Through Structural Constraints Predicted by Machine Learning

Authors:
Davide Bau (*School of Computer Science and Informatics, UCD Dublin*)
Alessandro Vullo (*School of Computer Science and Informatics, UCD Dublin*)
Gianluca Pollastri (*School of Computer Science and Informatics, UCD Dublin*)

Short Abstract: We describe Distill, a fully automated computational system for ab initio prediction of protein C-alpha traces composed of: a set of predictors of protein features; a simple optimisation algorithm. On a small PCs Distill can solve protein coordinates on a genomic scale in the order of days.

Long Abstract:
Of the nearly 2 million protein sequences currently known, only about 10% are human-annotated, while for fewer than 2% has the three-dimensional structure been experimentally determined. Attempts to predict protein structure from primary sequence have been carried out for decades by an increasingly large number-of research groups. Experiments of blind prediction such as the CASP series demonstrate that the goal is far from being achieved, especially for those proteins for which no resemblance exists, or can be found, to any structure in the PDB - the field known as ab initio prediction.

We describe Distill, a fully automated computational system for ab initio prediction of protein C-alpha traces. Distill is composed of: a set of predictors of protein features based on machine learning techniques and trained on large, non-redundant subsets of the PDB; a simple optimisation algorithm that searches the space of protein C-alpha trace configurations under the guidance of a potential based on these predicted features. Structural constraints (secondary structure, relative solvent accessibility, residue contact maps, contact density, contact maps between secondary structure elements) are predicted from primary sequence at the first stage of the reconstruction pipeline. The underlying predictive systems are implemented using the DAG-RNNs learning technique and proved to be state-of-the-art.

The optimisation algorithm we used for the reconstruction of the coordinates of protein C-alpha traces is organised in two sequential phases, bootstrap and search. The function of the first phase is to bootstrap an initial physically realisable configuration with a self-avoiding random walk and explicit modelling of predicted helices. A random structure is generated by adding C-alpha positions one after the other until a draft of the whole backbone is produced. If the i-th residue is predicted at the beginning of an helix all the following residues in the same segment are modelled as an ideal helix with random orientation.

In the search step, the algorithm refines the initial bootstrapped structure by global optimisation of a pseudo-potential modelled as a function of geometric constraints inferred from the underlying set of 1D and 2D predictions using local moves and a simulated annealing protocol. The conformational space search is carried out by displacing a randomly

chosen C-alpha atom at position r(i) to the new position r(j) by a crankshaft move, while leaving all the others C-alpha atoms of the protein in their original position. Secondary structure elements are displaced as a whole, without modifying their geometry. A new set of coordinates is accepted as the best next candidate with probability defined by the annealing protocol.

The protein data set used in reconstruction simulations consists of a non redundant set of 258 protein structures showing no homology to the sequences employed to train the underlying predictive systems. This set includes proteins of moderate size (51 to 200 amino acids) and diverse topology as classified by SCOP (all-alpha, all-beta, alpha/beta, alpha+beta, surface, coiled-coil and small). For each protein in the test set, we run 10 folding simulations and average the distance measures obtained over all 258 proteins.

An analysis of the results according to the SCOP assigned structural class and sequence length, highlights that a significant fraction of alpha-helical proteins and those lacking significant structural patterns are correctly modelled. Reliable identification of strands and the corresponding patterns of connection is a major source of difficulty. Nevertheless, the reconstruction pipeline identifies almost correct folds for about a third of the cases in which a short protein contains a significant fraction of beta-paired residues.

A preliminary implementation of Distill showed encouraging results at CASP6, with model 1 in the top 20 predictors out of 181 for GDT TS on NovelFold hard targets, and for Z-score for all NovelFold and NearNovelFold targets. Although Distill's predictions are often still rather crude, they are nonetheless useful. For instance they can be satisfactorily used to refine secondary structure and contact map predictions, and may provide a valuable source of information to identify protein functions more accurately than it would be possible by sequence alone. Distill's modelling scheme is fast: on a small cluster of state-of-the-art PCs it can solve protein coordinates on a genomic scale in the order of days.