What Can You Prove About Proteins?

Suchetan Dontha, Matthew Coudron

July, 2024

Abstract

Biologically realistic mathematical models of protein molecules are complex and seemingly messy objects which have long eluded predictability via the techniques of proof-based mathematics, despite concerted efforts. In contrast empirical and heuristic approaches to problems such as protein structure prediction have made enormous progress in recent years. In this work we re-examine, from a new angle motivated by recent progress, the old question: "What can one *prove*, mathematically, about proteins?".

We wish to emphasize the following features and contributions of our approach:

- In contrast to prior theoretical work on proteins, we abandon all "toy models" of proteins in favor of working exclusively with the industry standard mathematical representation of proteins used ubiquitously by biologists, the so-called PDB (protein database) file format. This makes our proven results directly compatible with the model most familiar to biologists, at the expense of making each result much more difficult to prove. We believe this tradeoff gives our research the best chance of providing value in the long term.
- We exhibit a Linear Program which, given a folded protein described in PDB format, can efficiently and *provably* certify whether the given fold is an energy local minimizer under a given, industry-standard, potential energy function.
- To demonstrate the relevance of our Linear Program, we use it to show that some industry-standard Monte Carlo heuristics for protein folding, which are implicitly assumed to converge to an local-energy-minimizing fold, in fact converge to a fold which is not a local minimum at all. To address this, we contribute a simple alternative gradient descent algorithm which, we demonstrate, finds true local-energy-minimizing folds as provably certified by our Linear Program. We provide examples of these phenomena for proteins consisting of a handful of amino acids, and show that our algorithms can scale efficiently to larger proteins, given more compute.
- We demonstrate that our Linear Program can be extended from certifying local minima to certifying the a given protein fold, in PDB format, lies in a "rough funnel" in the energy landscape. This is a concept considered by biologists but never previously formally proven about any protein fold to our knowledge.
- Since heuristic Monte Carlo energy minimization algorithms form a subroutine in the protocol used to calibrate tuneable parameters in industry-standard potential energy functions for protein folding, we propose a completely new way of calibrating the same tuneable parameters based on a provable algorithm instead. We describe a method to compute the optimal values of these tuneable paramaters directly from real protein imaging data based on a linear program with provable guarantees, thereby eliminating the need to use a heuristic Monte Carlo algorithm to define a centrally important object in the field.