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. However, while proof-based approaches may be slower and more difficult to design, they may also be easier to concretely evaluate for correctness, particularly when minimal experimental data is available. In a context where the collection of high quality experimental data remains the true bottleneck, this may yet be an advantageous trade-off. In this work exhibit new results and a new framework to address the old question: "What can one *prove*, mathematically, about proteins?".

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

- **Framework Compatibility 1:** In contrast to prior theoretical work on proteins, we abandon all "toy models" of proteins in favor of working directly with the industry standard mathematical representation of proteins used ubiquitously by biologists, the so-called PDB (protein data bank) 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 trade-off gives our research the best chance of providing value in the long term.
- **Framework Compatibility 2:** In our proofs we take, as mathematical "ground truth", widely accepted potential energy functions defined and used by the most popular protein folding software (Rosetta, CHARMM, etc). We show that a number of interesting results can be proven based on a fixed potential energy function alone, which, could yield new proof-based tools for the process of selecting, calibrating, and evaluating these functions which are used ubiquitously in the field.
- **Provable Stability Checking:** We exhibit a Linear Program which, given a folded protein described in PDB format, can efficiently and *provably* certify whether the given fold is a local energy minimizer under a given, industry-standard, potential energy function.
- **On the need for Provable Stability Checking:** We use our linear program to show that industry-standard Monte Carlo heuristics for protein folding do not always converge to a fold which is a local energy minimizer. This provides an explicit counterexample to the implicit intuition in the field, namely, that Monte Carlo heuristics converge precisely because they are arriving at a fold which is a local energy minimizer.
- **New Algorithms to Generate Explicit Proofs (Certificates):** We propose a simple gradient descent algorithm which, we show, features two advantages over other popular heuristics for protein folding under a potential energy function:
  - **Advantage on some Proteins:** For some amino acid sequences our algorithm converges to a protein fold which is a local energy minimizer, even though standard Monte Carlo approaches, acting on the same sequence, do not.
  - **Certificate of Provable Correctness:** When our algorithm completely converges to a particular fold it outputs a computer generated proof that the obtained fold is a local energy minimizer, ensuring that no mistake has been made.

We list several concrete examples in which we apply our folding algorithm to proteins consisting of a handful of amino acids, with an explicitly specified potential energy function, and including explicit computer-generated proofs of the stability of each obtained fold under that potential energy function. Our stability checking algorithm has a runtime which is polynomial (quadratic) in the size of protein molecules, and can therefore be scaled to handle larger proteins. We make our code freely available on github.