Poster J-43

On the Creation of Provably-Correct Models of Biological Structure and Process

Authors:
Randall L. Owen (*University of Oxford*)
Charlotte Deane (*University of Oxford*)

Short Abstract: We present an application of formal software engineering methods to an established bioinformatics model: the HP lattice model for protein folding. Our results demonstrate the rigorous creation and implementation of provably correct biological models, demonstrating how we can formally verify that complex models is performing as intended.

Long Abstract:
The success of the various genome sequencing projects, coupled with new high-throughput experimental biological methods, are generating large genomic and proteomic data sets that too complex for manual analysis. For those researchers in structural bioinformatics and genomics working at the atomic and subcellular spatial scales, modeling is becoming an important tool for analyzing, storing, retrieving, and displaying information. Yet models present their own set of issues which, if not carefully addressed, can generate misleading or incorrect results.

One of the key modeling issues is correctness: we need a way to formally verify that the model is performing as intended. Because structural data is nonlinear and the search space continuous, most models include multiple approximations and simplifications. In addition, since models are abstractions of physical objects the relation between model predictions and the underlying physics must be maintained. Thus, a rigorous framework for specifying model performance, coupled with theorems and rules for transforming the specification to a correct software implementation, is required.

In this work, we present one of the first applications of formal software engineering methods to an established bioinformatics model: the HP lattice model of protein folding. A formal specification framework for the HP lattice model, based on the Z notation and semantics, is introduced. The basic specification abstracts away unnecessary details and focuses on the key characteristics of the protein folding problem. In addition, an extension of the basic specification to include real numbers, three dimensions, and more complex energy interactions is discussed. The overall goal is to demonstrate that formal software engineering methods can be applied to create more realistic models of biological structure and processes.

One advantage of the formal specification and its development process is that many implementation details can be ignored in order to focus on the most important aspects of the problem. In this first phase of the process, the focus was on defining a state schema for the HP lattice model and a basic set of operation schemas for changing the state (conformation). With respect to finding the native (minimum energy) conformation, the actual minimization algorithm, is neither specified nor modeled.

A second advantage of building a formal specification is that it provides two very important proof opportunities. These proof opportunities are relatively
simple mathematical tasks that demonstrate that the data type requirements are consistent and that the operations are applied only within their domains. The first task is accomplished by proving the initialization theorem to show that an initial state exists, while the second task investigates the operation preconditions using the one-point rule. The challenge with both of these tasks depends on the complexity of the underlying problem.

While the present application is a simple and well known model from bioinformatics research, from an theoretical point of view our method provides a improved way of formulating and proving the characteristics of complex models, which may be extended to more complex biological systems.