

Research Statement: Structural and Thermodynamic Properties of Proteins

Overview & Aims

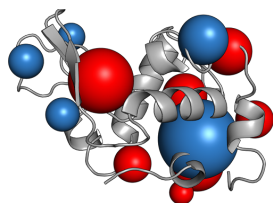
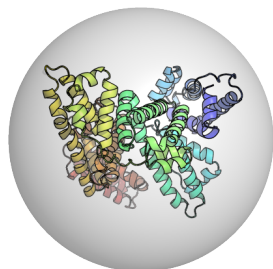
My research has been primarily concerned with the structural properties of proteins and protein-protein interactions. Current biomolecular *in silico* approaches rely on a hierarchy of approximations with the shortest timescale dictating the level of detail required. While all-atom approaches used in state-of-the-art molecular dynamics simulations can integrate approximately 10 μ s per day, general purpose machines are limited to a hundredth of that time. Biological processes such as protein folding, crystallization, or phase separations can occur on timescales that vary from seconds to days. This leaves the all-atom approach computationally intractable without significant simplification. Especially when multiscale approaches are required or if there are significant interactions with the solvent, the challenge is to develop a framework that captures the salient features of the system at hand. Much of my research involves reducing the complexity of such systems.

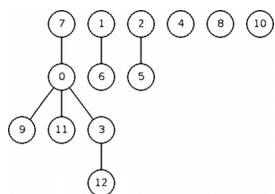
My time at the National Institutes of Health has afforded me a strong interdisciplinary network of collaborators. This network will help expand the scope and depth of my research. I plan to prioritize three research projects: the solvation of protein solutions, protein topology and enhanced sampling methods.

Research Proposals

Implicit Solvation & Electrostatic Protein Interactions

Accurate prediction of phase separation in a protein solution holds both great pharmacological and biochemical benefits. Liquid-liquid phase separation is present in the cytoplasm, in the formation of mammalian cataracts and sickle-cell disease, and often a prerequisite for protein crystallization. Phase separation is interesting as a thermodynamically driven event and is well-studied in both a theoretical and computational context. The simulation of explicitly solvated systems however, is currently limited to small system sizes since each solvent molecule must be individually modeled. If the water and any ionic species is modeled implicitly, the scope of modeling can be extended to scales where the calculation of phase separations becomes feasible. Current theoretical treatments cover only a coarse depiction of a solvated protein; liquid state theories are usually expressed as an expansion of a simpler isotropic potential. In contrast, the potential interaction surface of a real protein can be highly anisotropic and is often driven by charge-charge interactions of multiple solvated proteins. Development of an implicit solvation model that extends beyond a simple Debye-Hückel potential would build off my previous work at the NIH. The primary goal of this project would be the prediction of phase separations and radial distribution profiles over a wide variety of solvent conditions of varying pH and salt concentrations.





Topological Description of Protein Structure

Advances in crystallography and NMR have led to a wealth of information about the folded state of well-ordered proteins. In general however, the prediction of secondary and tertiary structure from a primary sequence remains an open problem. This is especially true for proteins that are intrinsically disordered or have multiple metastable states. The topological structure of a protein is often represented as a contact map of adjacent amino acids. This provides a topological representation of protein structure and naturally leads to a weighted graph. This graph can be analyzed to study the interactions of residues and their relationship to the final geometrical folded state of the protein.

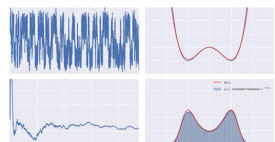
Interestingly these topological protein graphs can also be used as a dynamical variable. For example, the contact graph as a function of time could follow a folding trajectory along a reaction coordinate. My research has previously explored this idea by mapping the conformational states of a short peptide onto an Ising model. The model was sufficient to explain a primitive model of folding, but a more comprehensive description is needed for systems of larger complexity.

I intend to analyze experimentally resolved protein structures in the context of their graph properties. As a first step towards this goal, I have developed a catalog of simple graphs and their invariant properties. This database has already produced a number of new and interesting invariant integer sequences and has suggested relationships between the invariants. I plan to extend this database to biologically relevant macromolecules. The preliminary work on this project has already yielded insights into the structure of RNA molecules. Additionally, combining the invariant database with biological function can help elucidate the complex protein structure-function relationship. This project is well-suited for graduate and undergraduate levels of student work in both the mathematical and biophysical communities.

Development of Enhanced Sampling Methods

In biomolecular simulations, the objective is often an accurate sample of the thermodynamically relevant regions. A perfect unbiased simulation would sample all the regions of phase space with the appropriate Boltzmann weighting. In practice, the computational effort involved in a typical simulation requires a bias to enhance sampling around free-energy barriers. As an example, the sampling in a molecular dynamics simulation can be enhanced via modifications of the potential through Umbrella Sampling, Metadynamics, or Hamiltonian Replica Exchange. Other techniques such as the Wang-Landau (WL) method modify the acceptance ratio of Metropolis-Hastings Monte-Carlo. In flat-histogram methods such as WL, the invariant measure is modified from the Boltzmann ensemble to the density of states. For glassy systems, such as the canonical Ising model, this approach has been extremely successful in the prediction of critical values. It is less clear what the optimal sampling method is for a generic system with a rough funneled energy landscape. I plan to develop enhanced sampling methods for biological systems, especially those that sample the density of states.

Since the sampling method is the foundation for many computational models, their development is always an ongoing area of research. In addition to developing new algorithmic techniques, I have been in collaboration with the sampling community to develop a "Benchmark of



Sampling Algorithms.” These benchmarks would measure the accuracy and convergence times of a set of standardized problems in biological modeling. While a common set of benchmarks exists in other disciplines, especially computational biology, none yet exist for the problems faced in the protein folding community. This makes the comparison of competing algorithms difficult and often leads to the adoption of only a small subset of the available algorithms. I will bring this project to full maturity. I plan to develop a full suite of sampling test problems, the results of each algorithm applied to the problem, and an open, downloadable implementation.