My principal research interests lie in planning the motions of movable objects such as proteins and robots. This work [5], unified under the theme of intelligent motion planning, addresses important questions of analysis and feasibility in both robotic [1,2,9] and biological domains [3,4,6–8,10,12]. While these sound like divergent domains, we have shown that the same underlying algorithmic framework can be applied to solve these high-dimensional, complex problems. We even use the same code base for all projects! Our method is derived from probabilistic roadmap methods (PRMs) originally developed for robotic motion planning. This method builds a graph, or roadmap, where conformations are represented as vertices and transitions between conformations are edges. I am particularly interested in expanding my research program in the area of biological problems where my research can have direct and critical impact on the study of diseases. This research is particularly accessible to undergraduate and master’s level students. I have supervised one masters and twelve undergraduate students on protein folding and robot motion planning. For complete descriptions, paper availability, and animations please visit my web page (url above).

Figure 1: (a) The folding energy landscape is the set of all protein conformations and their associated energy. Building an approximate map of the energy landscape consists of two steps: (b) conformation sampling and (c) connecting samples together with feasible transitions.

1 Computational Biology

Molecular motions play an essential role in many biochemical processes. For example, as proteins fold to their native, functional state, they sometimes undergo critical conformational changes that affect their functionality, e.g., diseases such as Mad Cow or Alzheimer’s are associated with protein misfolding and aggregation. Knowledge of the stability, kinetics and detailed mechanics of the folding process may provide insight into how and why the protein misfolds. Also, it has recently been found that some RNA functions are determined by the folding process itself and not just by the sequence and the resulting native state.

Since it is difficult to experimentally observe molecular motions, computational methods for studying such issues are essential. Traditional computational approaches for generating folding trajectories such as molecular dynamics (MD) and Monte Carlo simulation are so expensive that they can only be applied to relatively small structures even when they use massive computational resources, such as tens of thousands of PCs, XBoxes, and PlayStations in the Folding@Home project or large supercomputers. In a recent study, IBM’s massive Blue Gene Server ran a protein of record size, just less than 130 amino acids. In comparison, biochemists are delving into the prion protein that misfolds and causes diseases such as Mad Cow and human Creutzfeldt-Jakob. Prion proteins have been found to be larger, e.g., 209 amino acids for human PrP and 467 amino acids for yeast Sup35. Another computational method, statistical mechanical models, has been applied to compute statistics related to the global folding process for protein and RNA molecules. While
computationally more efficient than molecular dynamics or Monte Carlo simulation, these methods do not produce individual folding trajectories and are limited to studying global averages of the folding process.

In order to computationally study interesting, large, and biologically-relevant molecules, I have explored a novel and efficient computational technique for studying molecular motions. In a matter of a few hours on a desktop PC, both microscopic folding pathways and global folding properties for protein and RNA molecules of hundreds of residues can be studied with our PRM-based method. The roadmap we construct approximates a molecule’s energy landscape. As shown in Figure 1, the energy landscape relates conformations to energy. While each molecule has its own unique landscape, the global minimum of each landscape is the lowest energy point, the native state. The unique physical features of a folding landscape, e.g., the hills and valleys, determine the folding behavior for that molecule. Our approximate map of the landscape quickly and efficiently captures the principal features of the landscape through both global views of the folding process and microscopic views of many (typically thousands) folding pathways.

We have developed new techniques based on Monte Carlo solution, master equation calculation, and non-linear dimensionality reduction to run simulations and analysis on the approximate map. The first method, Map-based master equation calculation (MME), extracts global properties of the folding landscape such as global folding rates. On the other hand, another method, Map-based Monte Carlo solution (MMC), can be used to extract microscopic features of the folding process. Also, the application of dimensionality reduction returns a lower-dimensional representation that still retains the principal features while facilitating both modeling and analysis of motion landscapes. The key advantage of our methods is the efficiency in which biologically relevant folding behaviors can be studied.

**Protein Folding.** In our preliminary work, we studied protein folding by building approximate maps, or roadmaps, for several proteins of varied length and structure. We obtained promising results that were validated by comparing secondary structure formation order with known experimental results [10, 11]. Subsequently, we were able to extend these results through the introduction of the MMC and MME techniques for roadmap-based analysis [6]. Both MMC and MME use the roadmap as a framework for computation and encode the edges as Boltzmann probabilities. These new methods have allowed us to compare time-ordered structural events extracted from our roadmaps to lab experimental methods that show folding over time, folding kinetics. For example, we have explored rate of conformational change from the unfolded state to the native state (folding rate), the times at which the different conformations are populated (population kinetics), and structural measurements that relate to experimental techniques such as fluorescence, CD spectra, and hydrogen exchange [6, 12].

Many computational techniques struggle when simulating the motions of large proteins because the space of possible conformations grows exponentially with protein size. For this reason, we have explored an analysis method that can be applied to landscape models, called dimensionality reduction [8]. This computational technique finds the principal features of a high-dimensional space, represented by our motion landscapes, and returns a lower-dimensional representation that still captures the principal features. Dimensionality reduction enables quick and useful global analysis of our landscapes. Through a new use of it as an analysis tool, it can reduce our original model size by almost half, thus facilitating the study of larger proteins.

Our results are quite promising. Our new techniques have been able to capture structural events that have been shown in lab experiments, such as those found for protein G and its mutants, NuG1 and NuG2 [6]. We also demonstrated in [6] that kinetic measurements based on lab experimental techniques give greater detail into the folding process and provide new ways to validate our methods. In [12] we show that these kinetic methods are critical to detailed insight into the folding process, e.g., identifying the folding core. Also, the application of dimensionality reduction to our roadmaps produced maps that were up to 53% smaller for all proteins studied, yet were still able to capture the experimentally determined folding orders including those for Protein G and its mutants [8].

**Proposed Research Protein Folding.** I plan to extend this research in two complementary directions: computational theory and experimental exploration. First, since our underlying models are represented by graphs, there are many graph-based methods and analysis techniques to be explored. Graph-based algorithms have proven useful in many computational biology problems such as molecular structure comparison and analysis of protein interaction networks. One such method, isomorphism, could be applied to two models to compare their similarity. It also could be applied on subgraphs, e.g., paths, to find paths with similar features. Second, experimental exploration would complement the computational tools by identifying interesting and
biologically relevant problems to explore. For example, there has been a long standing experimental debate on the folding behavior of the protein Lysozyme. Some theorize that it folds in multiple distinct ways. Others believe there is a single folding behavior. New computational analysis would likely provide helpful insight into this debate and explain why Lysozyme can misfold and cause disease.

**RNA Folding.** Ribonucleic acid (RNA) motions are responsible for many biological processes including synthesizing proteins, catalyzing reactions, splicing introns, and regulating cellular activities. For example, it has recently been found that RNA folding velocity may regulate the number of copies of DNA strands that are present in a cell (plasmid copy number).

Due to the exponential costs, enumerating all secondary structure conformations is only possible for small RNA (less than 20 nucleotides). We have explored the use of a probabilistic Boltzmann sampling method for larger RNA. Kinetic analysis of our approximate RNA folding landscapes through the application of the MMC and MME techniques has produced results that we can validate against experimental methods [3, 4]. For example, we were able to replicate the kinetic functional rates of MS2 phage RNA and three mutants that were seen in experiment.

Despite the fact that RNA conformations can be represented with a secondary structure model, the configuration space represented by all possible RNA conformations is not simple. We have shown that non-linear dimensionality reduction techniques are well-suited to find the representative features of the RNA landscape [8]. With these reduced models, we have demonstrated that important landscape features such as coverage can be better explored.

**Proposed Research RNA Folding.** Similar to protein folding, the extension of RNA research will involve both computational techniques and experimental exploration. Due to its conformational representation, RNA provides us with many interesting experimental results to explore. For example, many folding intermediates have been identified for RNA. Through a new combination of dimensionality reduction and our existing analysis techniques, we should be able to identify important intermediates. Also, since both our protein and RNA folding models are graphs, techniques developed in one domain should translate to the other.

2 **Machine Learning Applied to Robotic Motion Planning.**

While there have been many different algorithmic methods developed for motion planning, no one method works well in all planning spaces. For example, some spaces might have narrow passageways that are difficult to plan in or open regions that are easier. These space characteristics can exist in any planning domain (such as proteins and RNA), but they have been best explored in the area of robotic motion planning. In this domain, there are many individual planning methods developed whose strengths are known by domain experts, e.g., the original PRM method for open regions and the Obstacle-based PRM in constrained regions.

In order to take advantage of this existing library of methods, we have explored using the features of the planning space to help decide where and when to apply particular planners. In preliminary work, a supervised learning method classified features of the space and selected a sampler to apply in a certain region of the space [1, 2]. In recent work, we have used spatially and temporally identified features in order to better decompose the problem and selectively apply planners that adapt over time [9]. This new strategy takes advantage of unsupervised learning methods at all stages of the planning process and produces solutions in complex spaces with little cost and less manual intervention compared to other adaptive methods.

An example is shown in Figure 2 for a maze environment with a movable object. First, features from a small sampling of the space are identified and used to cluster the samples. Each cluster relates to a region of the space (Figure 2(b)). In order to define the optimal number of clusters (n), the elbow criterion is calculated from the variance in the clusters (Figure 2(d)). Intuitively, this criterion selects n such that adding additional clusters does not add sufficient information. Subsequently, an appropriate planner can be selected from a library and applied in each region.

**Proposed Research Robotics.** There are many directions in which to continue work in adaptive planning methods. For example, new features and new data mining methods could be applied to better represent the problem space. Also, there are many known planning methods including specialized heuristic-based roadmap methods, trees, and even some complete methods. It would be helpful to continue to explore
Figure 2: Automatic region identification in a maze environment. (a) Environment shown with movable body shown above and enlarged. Notice there are three different regions which the robot must traverse: open, constrained, and open. (b) Clustering identifies 3 regions (circled) corresponding to the features of the space. (c) Continued clustering can unnecessarily split the regions further. (d) An automated method, the elbow criterion, determines the best number of regions (red star).

how these methods can be used in concert in order to simplify and automate planning while taking advantage of each method’s strengths.

References


