Abstract

Protein folding is a central problem in bioinformatics. The protein folding problem asks how a protein's amino acid sequence dictates its three-dimensional atomic structure. This problem has wide applications and a long history dating back to the 1960s. From the viewpoint of theoretical computer science, there is ongoing research aiming at revealing insights into reality by working on simplified abstract models. A protein is modeled by an abstract linkage in a lattice space, which has been proven to be an extremely useful tool for reasoning about the complexity of protein structure prediction and design. Since protein functionality is controlled by the native state structure, these models provide the ideas of theoretical investigations for future use in studies on real proteins.

In this thesis, we aim to explore the computational tractability or intractability of protein structure prediction and design using techniques from computational complexity theory and algorithm design and analysis. Inspired by the popular hydrophobic-polar (HP) model, we have developed some new graph-theoretic models and problems that may be seen as variations or extensions of the standard HP one. We then studied the tractability of these problems, by finding polynomial-time algorithms or by proving that they are NP-hard. In our model, we combine the basic ideas of protein folding with the complementary problem of *protein design*, where the goal is to synthesize a protein of a given shape (and function) from an amino acid sequence. We modeled proteins and amino acid chains in two different settings. In each setting, we studied structure prediction and design of proteins by folding amino acid chains.

The first setting is protein folding prediction and design in grid graphs. It is inspired by the famous hydrophobic-polar (HP) model for protein folding. A protein in the HP model is represented as an abstract open chain, where each link has unit length and each joint is marked either hydrophobic or polar. Grid graphs are graphs that form a regular tiling of the 2D plane or 3D space; these graphs are the standard setting for the traditional HP model. Here a protein is a (connected) subgraph G of the grid graph, possibly with colors assigned to nodes. An amino acid chain is a path graph P with colored nodes. We thus propose the bicolored path embedding problem. A graph is said *bicolored* if each vertex is assigned a label in the set {red, blue}. For a given bicolored path P and a given bicolored graph G, our problem asks whether we can embed P into G in such a way as to match the colors of the vertices. In our model, G represents a protein's "blueprint," and P is an amino acid sequence that has to be folded to form (part of) G.

In this setting, we first proved that the bicolored path embedding problem is NP-complete even if P is monochromatic (e.g., all its vertices have the same color). Then we proved that the problem is NP-complete even if G and P are bicolored and have the same number of vertices. The latter result is a fairly technical reduction from the Hamiltonian path problem. The importance of this result lies in the fact that, when G has the same number of vertices as P, it represents an exact "blueprint" of the protein we want to obtain, as opposed to just an "ambient space" for it.

Next, we contrasted these hardness results with a polynomial-time algorithm for the case where G is a grid of fixed height: thus, the bicolored path embedding problem, parameterized according to the height of G, is in XP. The technique we used is dynamic programming, and the significance of this result is that it offers an efficient algorithm for a generic grid G, although the algorithm's performance deteriorates with the height of G (but not as much with its width).

We further showed that the classical problem of maximizing H-H contacts is also NP-hard in the context of the bicolored path embedding problem. Note that, in previous work, it has been established that the problem of maximizing H-H contacts is NP-hard when G is not given, and P can be embedded in any way on a grid.

In the second setting, we studied the protein folding prediction and design in general graphs. These are graphs with either colored vertices or edges of a given length. Here a protein is a graph: the idea is that a protein has a "high-level" shape that can be represented by some graph G, even if at "low level" the protein is just a chain. An amino acid chain is a path P, possibly with colored nodes or fixed edge lengths. The prediction problem asks, for a given amino acid chain, what graphs it can fold into (i.e., what graphs it can be mapped onto), satisfying some local constraints. The design problem asks, for a given graph (i.e., a protein), whether

we can design an amino acid chain that can be mapped onto it satisfying some local constraints.

In this setting, we first studied graphs with colored vertices. We proved that the bicolored path embedding problem is NP-complete even if G is a dense graph with the same number of vertices as P. Here, the previous remark about G being an exact "blueprint" of a protein holds, since it has the same number of vertices as P. Additionally, the fact that G is dense (i.e., it has a quadratic number of edges) makes this result surprising for another reason: intuitively, a blueprint with many edges should allow greater leeway in the construction of an embedding of P. As it turns out, a greater amount of freedom does not necessarily translate into our ability to easily find embeddings. We also prove a complementary result: the problem of constructing a path P that embeds in a given bicolored graph G maximizing H-H contacts is Poly-APX-hard. In particular, it has no polynomial-time approximation algorithm with a sub-polynomial approximation ratio, unless P = NP.

We also considered a different model, where G is a (non-colored) graph with given edge lengths, and P is a linkage (a path with edge lengths). Our goal is to find an embedding of P in G that matches the lengths of edges. The problem asks if there is an edge-weighted Eulerian path of target graph G spanned by the linkage P. We showed that the problem is strongly NP-hard even if edges have only two possible lengths. Together with the fact that the problem is solvable in linear time if edge lengths are all the same, this result gives a precise characterization of the problem's tractability.

We tackled this intractability result by considering two different variants of the problem. In the first variant, we allow the edges in P to be elastic, that is, we can stretch or shrink the edges in the elastic linkage P. The goal is to minimize the elastic ratio of the embedding. Remarkably, we found that when G is a path, there is a polynomial-time algorithm based on dynamic programming. In the second variant, we allow P to cover an edge of G twice or more. We showed that the problem is NP-hard even if G consists of a single edge. Furthermore, with the requirement that each edge of G is covered by P exactly twice, we obtained three hardness results and one polynomial-time algorithm when G and edge lengths are restricted.

This research has applications that go beyond protein folding and design, and has laid the foundations for interesting future developments, as well.

Keywords— protein folding problem, HP (hydrophobic-polar) model, embedding problem, Hamiltonian path problem, edge-weighted Eulerian path problem