Interactive Protein Manipulation

Oliver Kreylos* Bernd Hamann* Nelson L. Max* Silvia N. Crivelli† E. Wes Bethel‡

1 Introduction

One of the grand challenges in computational biology is the prediction of the three-dimensional structure of a protein,which determines its function, from its chemical makeup alone. A protein's *primary structure*, i. e., its amino acid sequence, is directly encoded in its DNA sequence,which is a purely one-dimensional structure that does not directly encode a three-dimensional shape. It is commonly believed that the "native" shape of a protein is the one corresponding to the global minimum of its internal energy; thus, the *protein folding problem* has been treated as an optimization problem in recent years. It is important to start solving any optimization problem from a "good" set of initial configurations that allow the optimization code to, ideally, search the complete optimization space for a global minimum. Our work focuses on providing an interactive, visual tool to rapidly create many initial configurations for a given amino acid sequence, which are then used as input for an optimization algorithm.

2 Protein Structure Hierarchy

Proteins considered by our tool have three levels of structure [1]:

Primary Structure. A protein's primary structure is its amino acid sequence. It is directly encoded in a protein's gene (each triple of bases defines one amino acid). The chemical makeup of a (simple) protein is a single chain of amino acid residues connected by peptide bonds (see Figure 1).

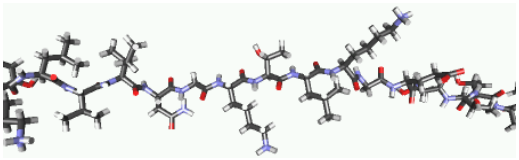

Figure 1: Part of the primary structure of a protein.

Secondary Structure. Adjacent amino acid residues inside a protein can interact with each other to form substructures such as α-helices (see Figure 2) and β-strands. Inside an α-helix, each residue forms hydrogen bonds with two other residues, accounting for their rigidity. For each amino acid type, the probabilities of it forming either one of these substructures are known, and neural networks are used succesfully to predict secondary structure occurrences from amino acid sequences [2].

*Center for Image Processing and Integrated Computing (CIPIC), Department of Computer Science, University of California, Davis; {kreylos,hamann,max}@cs.ucdavis.edu

†Lawrence Berkeley National Laboratory; {sncrivelli,ewbethel}@lbl.gov

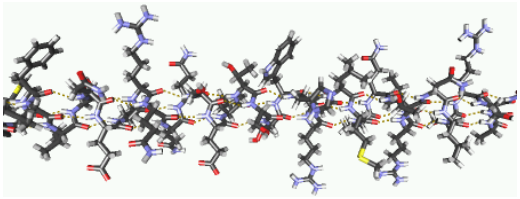

Figure 2: An α-helix. Hydrogen bonds are depicted as dashed yellow lines.

Tertiary Structure. A protein's three-dimensional structure is formed by amino acid residues from distant parts of the chain forming bonds with each other. β-strands, not very rigid by themselves, hydrogen-bond with each other to form stable β-sheets (see Figure 3) whereas α-helices cluster to each other to hide their hydrophobic amino acids from the surrounding watery solution. Prediction of tertiary structure is still an unsolved problem.

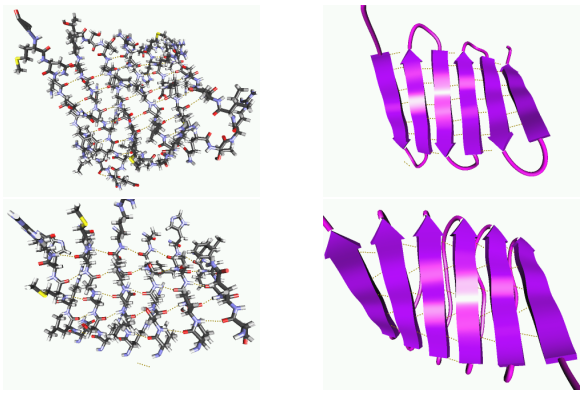

Figure 3: Two β-sheets. Hydrogen bonds are depicted as dashed yellow lines. Top row: anti-parallel sheet (left) and cartoon rendering (right); bottom row: parallel sheet (left) and cartoon rendering (right).

Quaternary Structure. Many proteins, e. g., hemoglobin, contain more than one amino acid chain. For those, quaternary structure describes how separate chains interact with each other to form an overall shape. Our tool, and the used optimization code, do not consider multi-domain proteins.

3 Protein Folding as Optimization Problem

An optimization problem is defined by its configuration space and target function. In the case of protein folding, the configuration space is the space of all possible three-dimensional configurations of a given protein, and the target function is its internal energy. Although simple proteins are single molecules, they are of surprising flexibility. All amino

acid residues except proline have two rotational degrees of freedom, the two *dihedral angles* ϕ and ψ (see Figure 4), and all except glycine have additional degrees of freedom in their side chains. Typical proteins consist of hundreds of residues, making the optimization space high-dimensional. Furthermore, the internal energy function has local minima in abundance. Together, these two facts make protein folding a very difficult optimization problem.

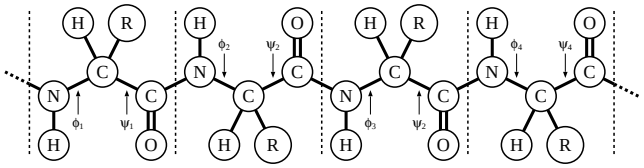

Figure 4: Rotational degrees of freedom along a residue chain. Adjacent residues are separated by dashed lines; side chains are denoted by R.

4 Creating Initial Configurations

In order to exhaustively search the space of all possible configurations of a given protein for its global energy minimum, the optimization algorithm needs to be provided with dozens of initial configurations distributed over the entire space. In the past, initial configurations were designed by the computational biologists, and then created by a separate constrained local optimization system forming them out of an unfolded residue chain. This approach is time-consuming and non-intuitive, and many possible initial configurations can be overlooked in the process. We have created a visual tool to directly manipulate protein structures. The idea is to let biologists assemble proteins as if using plastic stick-and-ball models. Our goal was to keep proteins intact during manipulation by using their intrinsic degrees of freedom to achieve intended movements. If a user selects a secondary structure, say a β-strand, and moves it towards another β-strand to form a β-sheet, then the amorphous *coil regions* between those two structures will bend and twist to allow the motion.

5 Inverse Kinematics for Molecular Modelling

Since proteins are inherently flexible and thus allow a wide range of motion, the main problem is to translate a user's six-degree-of-freedom motions into changes of a chain segment's dihedral angles ϕ_i and ψ_i. This problem, *inverse kinematics* (IK) [3], has been studied in the field of robotics, where it is used to translate intended motion of a robot's hand into changes in joint parameters along a robot's arm. What makes this application of IK more difficult is scale: A robot assembly typically has up to a dozen joints, whereas we encounter linked assemblies of 40–80 joints in medium-sized proteins. Nevertheless, IK has turned out to be the method of choice for natural interaction with large molecules.

6 Using the Modelling Tool

A typical modelling session starts with reading a prediction file created by one of many publicly available secondary structure prediction servers [2]. These files contain the sequence of amino acid residues and, for each residue, an indicator determining whether that residue is part of an α-helix, a β-strand, or an amorphous coil region. With this information, and a set of standard amino acid structure files, our program creates a "pre-configuration" consisting of fully formed secondary structures, but no tertiary structure (see Figure 5).

Figure 5: Pre-configuration for protein 1PGX.

A user proceeds by aligning adjacent β-strands to form initial β-sheets. Later, the order of β-strands is permuted to quickly create several dozen configurations. During protein assembly, users typically ignore the exact aligment of α-helices and coil regions since the subsequent optimization process handles them well. Figure 6 shows two initial configurations for the same protein.

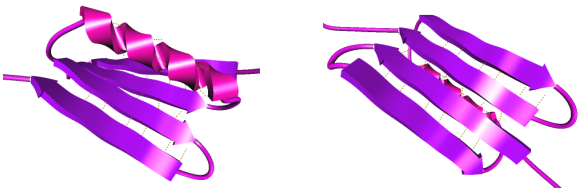

Figure 6: Two initial configurations for 1PGX.

7 Conclusions and Future Work

Our tool has been deployed for use by computational biologists and computer scientists at the Lawrence Berkeley National Laboratory, the University of California at Berkeley, the University of Colorado at Boulder, and the University of California, Davis. Our group is currently competing in the international CASP5 protein structure prediction competition, and our tool is being used to create dozens of initial protein structures every day. With the new tool, our group has been able to attack proteins of sizes that were not manageable before, and the high quality of the created configurations has exposed new behaviour in the existing optimization algorithm. We plan to integrate the tool with the optimization code, to be used as a front-end for monitoring and steering massively parallel optimizations.

References

[1] Lehninger, A. L., et al., *Principles of Biochemistry*, 2nd ed. (1993), Worth, New York

[2] McGuffin, L. J., et al., *PSIPRED: a Protein Structure Prediction Server*, `http://www.psipred.net`

[3] Welman, C., *Inverse Kinematics and Geometric Constraints for Articulated Figure Manipulation*, Master's Thesis, Simon Fraser University, Vancouver, Canada, 1993