ProteinShop and POSE: Bringing Robotics and Intelligent Systems into the Field of Molecular Modeling

Ting-Cheng Lu[1], Nelson L. Max[2], Silvia N. Crivelli[1,*]
[1]*California Institute for Quantitative Biomedical Research*
[2]*Department of Computer Science, University of California, Davis*
Tlu@lbl.gov, max@cs.ucdavis.edu, SNCrivelli@lbl.gov

**Corresponding author*

Abstract

ProteinShop and POSE are graphical infrastructures for the interactive modeling, manipulation, optimization and analysis of molecules. They were designed to bring interactive computer graphics in the field of molecular modeling to a level not attempted by other visualization programs. To achieve that goal, we adapted inverse kinematics algorithms commonly used in robotics to permit interactive manipulation of protein structures in a natural and intuitive way.

1. Introduction

A key component in the prediction of protein structure and docking is the symbiotic interaction between computing simulations and human knowledge. ProteinShop [1] was developed for the specific purpose of manipulating protein structures with pinpoint control guided by an energy function that can be provided by the user. It permits one to select elements of secondary structure and move them to form a desired tertiary structure. As dihedral angles change along the backbone between those elements, bond lengths and angles are maintained. Based on ProteinShop, we have started development of POSE *(Protein Optimization Steering Environment)*, with the goal of incorporating human knowledge and intuition into the protein simulation process itself. We hope that this environment can record and eventually "learn" from human decisions, which may lead us to both gain insight into and automate the simulation process. Currently, POSE supports interactive, user-driven monitoring and steering of a protein structure prediction process but the resulting infrastructure will easily translate to protein-protein interactions and RNA modeling.

2. ProteinShop

ProteinShop simplifies the process of creating a variety of protein structures by manipulating the proteins as if with our own hands, visually guided by a number of dynamically visualized cues such as energy functions, atomic collisions, and hydrogen bonds. Among its main features are: 1) automatic creation of 3-D structures from sequence, 2) alignment of beta strands to form beta-sheets, 3) easy coupling with user-provided energy and scoring functions, 4) real-time hydrogen-bond and atom-collision detection, and 5) local minimizations.

ProteinShop uses inverse kinematics (IK) algorithms [2] commonly used in robotics to implement an interactive manipulation feature that allows the user to select pieces of secondary structure and arrange them to form desired tertiary structures. The IK method computes backbone dihedral angles to achieve the molecular shape created by interactive manipulation. In fact, IK allows ProteinShop to translate 3D motions into bond rotations, which in turn allows the user to manipulate proteins in a very intuitive manner.

BuildBeta is another IK-based feature that increases ProteinShop's ability to quickly create a variety of protein configurations. It attempts to automate the process of creating beta-sheet configurations by generating an ensemble of plausible and likely initial configurations using all the predicted beta strands in the chain. Its input may be the output of a prediction server, which for each residue predicts the secondary structure type as belonging to a coil, alpha helix, or beta strand, and also gives confidence estimates for these predictions. The input may also be an existing PDB structure file, which has been annotated with remarks identifying the regions with these three types of structure. BuildBeta then uses probabilities for beta sheet topologies and matching alignments, generated from known structures, to decide automatically how to call the zipping routines. These routines perform IK to

form a specified pair of backbone hydrogen bonds between two beta strands, aligning them into the sheet, using the flexibility of specified coils.

The probabilities of the different strand topologies in a sheet are taken from [3]. The two components to a topology are the order of the strands in the sheet and the orientations of each strand (up or down), which determine whether adjacent strands are parallel or anti-parallel. The relative probability scores for all possible topologies are computed, depending on whether the structure is alpha/beta (more than 20% residues are alpha helix) or "all beta", the length of the coils between strands: "short" if ten or less residues and "long" otherwise, and “jumps” in the strand order. (See [3] for details.) Those with the best scores are further analyzed for good strand alignments.

The various alignments are generated by sliding a strand along an adjacent strand to change which residue pairs are opposite each other. The relative probabilities for the different alignments of two beta strands are taken from Zhu and Braun [4]. This paper contains three twenty-by-twenty matrices for the potential energy of pairs of residues, one from each strand, either (a) opposite to each in the alignment, (b) separated by one residue (diagonally opposite), or (c) separated by two residues (second diagonal). These matrices were generated from the statistics of actual alignments in a training set of known structures. When these energies are added for all pairs of types (a), (b), or (c) in a proposed alignment they produce a score for the alignment, with lower scores being more probable. The user can select the number of best scores to be considered for zipping attempts.

Sometimes the structure of a core part of a protein is determined by homology with a known structure and the task is to extend a sheet in the core by placing the remaining beta strands. In this case, BuildBeta uses a PDB structure file with the core regions in the chain identified by remarks. It also takes a prediction file with the structure predictions and confidence values for those predictions. Then, when generating and scoring proposed sheet topologies, it only considers those where the core region matches the topology in the given structure. It only proposes alignments for the beta strands outside the core regions. When zipping, it only marks as flexible the coil regions outside the fixed core. Thus, all structures produced will contain the input core region. The current version of BuildBeta can only construct a single beta-sheet, therefore, we can only match cores with at most one sheet. We hope to extend multiple sheets from known cores in the future.

When beta sheets are built with this automatic zipping, they may intersect coils and alpha helices that were poorly placed by the inverse kinematics routine, which makes no use of contact forces or potential energy computations. BuildBeta makes an attempt to move at least the alpha helices away from the beta sheet. For each pair of beta strands, which are consecutive along the chain, any intervening alpha helices are moved to the side of the sheet that would produce a right hand turn according to the definition of Richardson [5]. No attempt is made to move the coils, which may still intersect other structures, but energy minimization should be able to fix this problem.

3. POSE

ProteinShop's visualization capabilities let the user quickly develop a number of candidate configurations for input to a global optimization algorithm that predicts the 3D protein structure. Global optimization algorithms are computationally intensive. We are developing POSE, an application that supports interactive, user-driven monitoring and steering of the protein structure prediction process. It will provide an interface between a protein structure prediction solver and ProteinShop, with the goal of incorporating human knowledge and intuition into protein structure prediction, to accelerate the time to solution. The POSE philosophy is that a knowledgeable researcher who is following a global optimization process can make changes to certain structures and may return them to an energy-decreasing path. Furthermore, the researcher can reduce the time needed to find a solution by eliminating large, unproductive regions of the search space and by focusing the search on the most promising ones. POSE will provide an interface for easily plugging in energy and scoring functions as well as for easily coupling programs running on remote machines. This unique combination of intuitive human knowledge and supercomputing power should produce optimized protein structures more quickly than ever before.

4. References

[1] Crivelli, S., Kreylos, O., Hamann, B., Max, N. & Bethel, W., JCAMD, 2004, 18: 271-285.

[2] Welman C., Master’s Thesis, 1993, Simon Fraser University, Vancouver.

[3] Ruczinski, I., Kooperberg, C., Bonneau, R., and Baker D., Proteins: Structure, Function, and Genetic, 2002, 48, 85-97.

[4] Zhu, H., Braun, W., Protein Science, 1999, 8, 326-342.

[5] Jane Richardson, PNAS Vol. 73, No. 8, August 1976, pp. 2619 - 2623.