Template based modeling using a homology-based algorithm

Xiaokai Qian, Sean Lander, Caiwei Wang, Haipei Fan, Puneet Gaddam, Brett Koonce
University of Missouri - Columbia

Feburary 18, 2013

1 Abstract

We implement basic homology/template modeling of a DNA protein from CASP, T0644. First, we search (BLAST), find the best candidate template, and align it with our target sequence. Next, we build a copy of the the target protein's backbone using the template data wherever possible (our custom python tool). Then, we add sidechains (SCWRL) to produce a candidate PDB file. Finally, we optimize (3D-Refine) our machine-generated PDB file. Ultimately, we visualize (JMOL) our results against the known structure of the protein and discuss how our method performs compared to others via objective functions (TM-Score/RMSD).

2 Introduction

Template modeling is a technique very popular in bioinformatics. First, we begin with a DNA sequence of unknown structure. By searching a database of known sequences, we can find similar proteins to our unknown target. Then we can build a model of our target by using known elements from other proteins.

When the two sequences do not differ greatly, this approach can generate usable models far quicker than traditional methods (real-world x-ray crystal modeling can take years to produce a single structure). However, this approach is not foolproof. Of considerable importance is filling in the gaps (loop modeling) in the differences between the two sequences, as well as making sure the final molecule is still a viable real-world model.

For our project, we implemented basic homology modeling in python, using a number of external tools to produce a pipeline capable of taking initial input sequence and presenting a final visualized model.

2.1 Pipeline

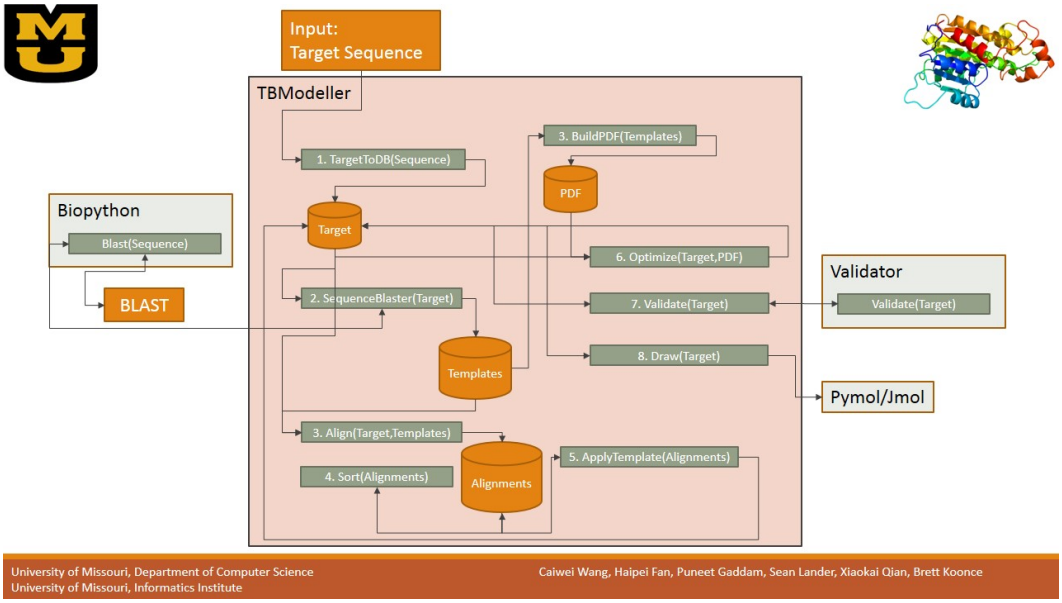

Figure 1: An overview of the TBPy pipeline

3 Template Search and Alignment

3.1 BLAST

The Basic Local Alignment Search Tool (BLAST) is a relatively time-efficient algorithm to compare amino-acid sequences of different proteins or the nucleotides of DNA sequences. BLAST can be used for inferring functional and evolutionary relationships between sequences as well as identifying members of gene families. The BLAST algorithm finds resembled sequences by locating short similar regions between two sequences using a heuristic approach. It compares a query sequence to a sequence database and computes the statistical significance of alignments to find sub-sequence in the database that resemble the query sequence within a certain threshold. The BLAST program is based on an open-source format and can be accessed freely over the internet. This enables anyone to be able to modify the code and implement similarity searches using the up-to-date datasets. The BLAST is available at http://blast.ncbi.nlm.nih.gov/Blast.cgi

3.2 CASP

Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a community-wide, worldwide experiment performed every two years since 1994, aiming to help improve the protein structure prediction techniques. The main goal of CASP is to evaluate the state-of-art in protein three-dimensional structure prediction, identify the advances that have been made, and pay close attention to future work that can be done. The methods

of objective testing of these techniques have been provided through the process of blind prediction. It has been viewed more as a "world championship" in this field of science.

3.3 PDB

The Protein Data Bank (PDB) is a repository for the collection, data processing and dissemination of the three-dimensional data on macromolecular structure, such as proteins and nucleic acids. The PDB is overseen by the Worldwide Protein Data Bank (wwPDB). It is a crucial resource for scientists nowadays who work in the area of structural biology as most major scientific journals require them to submit their data to the PDB. The structural data are commonly collected by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists all over the world. These data can be freely and publically accessed by the global community on the Internet via the websites of its member organizations (PDBe, PDBj, and RCSB).

4 Building a protein

4.1 Algorithm

5 Optimization and sidechains

5.1 SCWRL

SCWRL is a program for adding sidechains to a protein backbone.

SCWRL4 is based on an improved algorithm based on graph theory that solves the combinatorial problem in side-chain prediction more rapidly than many other available program.

SCWRL4 is based on a new potential function that results in improved accuracy at reasonable speed.It will converge on very large proteins or protein complexes or those with very dense interaction graphs.It depends on a backbone-dependent rotamer library. The library provides lists of chi1-chi2-chi3-chi4 values and their relative probabilities for residues at given phi-psi values, and explores these conformations to minimize sidechain-backbone clashes and sidechain-sidechain clashes.

The SCWRL4 executable saves the resolved optimal conformation of the whole protein model into PDB file. The corresponding value of the total energy is printed into the standard output, which can be redirected to a file for further analysis.

5.2 3DRefine

One of the major limitations of computational protein structure prediction is the deviation of predicted models from their experimentally derived true, native structures. We used a two-step refinement protocol, called 3Drefine, to consistently bring the initial model closer to the native structure. The first step is based on optimization of hydrogen bonding (HB) network and the second step applies atomic-level energy minimization on the optimized model using a composite physics and knowledge-based force fields. The approach has been

Figure 2: Target structure

Figure 3: Native structure

evaluated on the CASP benchmark data and it exhibits consistent improvement over the initial structure in both global and local structural quality measures. 3Drefine method is also computationally inexpensive, consuming only few minutes of CPU time to refine a protein of typical length (300 residues).

6 Visualization and results

After we obtain final target pdb file, we use Jmol to visualize it. At the same time, we also visualize the native one to compare it with our prediction.

Jmol is an open-source Javaviewerfor chemical structures in3D, that does not require 3D acceleration plugins. It is cross-platform, running on Windows, Mac OS X, and Linux/Unix systems. The JmolApplet is a web applet that can be integrated into web pages. The Jmol application is a standard application based on Java which can be run on the desktop. The JmolViewer is a development tool kit that can be integrated into other Java application. Besides, Jmol has many interesting features like Animation, Surfaces, Measurements and so on. With these features, we can fulfill many requirements of our visualization task. In addition, Jmol can not only accept PDB file format, but also MOL, XYZ, CIF file formats. Jmol is available here: http://jmol.sourceforge.net/.

The following two figures are two examples of Jmol. The left one is the target structure we generated, while right one is the native structure.

6.1 Visualizations

7 Results

After we generated the target PDB file, we use TM-Score and RMSD to evaluate our prediction. Also we compared our prediction with other servers' results based on same target.

In our project, we choose MuFold and MULTICOM to compare.

TM-score is an algorithm to calculate the structural similarity of two protein models. It is often used to quantitatively assess the accuracy of protein structure predictions relative to the experimental structure. Because TM-score weights the close atom pairs stronger than the distant matches, it is more sensitive to the topology fold than the often-used root-mean-square deviation (RMSD) since a local variation can result in a high RMSD value. TM-score has the value in (0,1). Based on statistics, a TM-score <0.17 corresponds to a random similarity and a TM-score >0.5 generally corresponds to the same fold in SCOP/CATH. The definition of TM-score is independent on the length of the proteins.

RMSD s the measure of the average distance between the atoms (usually the backbone atoms) ofsuperimposedproteins. In the study of globular protein conformations, one customarily measures the similarity in three-dimensional structure by the RMSD of the C atomic coordinates after optimal rigid body superposition.(adopted from Wikipedia)

Our results are shown in the following table. From the table, we can clearly see that our results is better than both MULTICOM and MuFold. But it's just one target we have tried. In the future, we should take more targets into consideration. By this way we are able to evaluate our prediction.

7.1 Scores

	Group1	MULTICOM	MuFold
TM-Score	0.9992	0.9072	0.1985
RMSD	0.118	1.666	14.978

7.2 Citations

Biopython citation
Blast citation
Casp
PDB citation
3DRefine paper
SCWRL paper
jMol/PyMol citations