University of Missouri-Columbia

Department of Electrical Engineering and Computer Science

Computational Modeling of Molecular Structure

Template based protein structure prediction

Prepared by

Tianqi Wu Anes Ouadou

Allen Effu

# Introduction

Protein structure prediction is one of the most important problems that persists despite the tremendous number of researchers and the amount resources dedicated to solving this problem. Many algorithms have been developed to try to predict protein structures. One of the widely used approaches is the template based prediction. This approach relays on the structure of known proteins to predict the structure of the unknown target protein with the condition that these proteins’ sequences bear some similarity to the target’s sequence. These proteins with known structures are referred to as templates, hence the name template based prediction. The pdb files that contains the coordinates of the atoms in the structure of the templates can be found in the known databases such the Protein Database (PDB). See figure1 to see a sample of a pdb file. These templates pdb files are used to create the pdb file for the target protein. Further techniques for gap handling and model refinement are employed to try to bring the final model of the target protein as close as possible to its correct native structure.

In this project we predict the three-dimensional structure of some proteins using template based approach.

![A picture containing text

Description generated with high confidence]()

Figure ‎1 sample format of the pdb file of a protein

# Method

The most important input we have is the protein sequence, without it we cannot predict the protein structure. The process of predicting the protein structure goes through important steps: finding alignments, generating the initial model, refining the model and lastly evaluation and visualization.

## Finding alignments:

The first step is to look for templates that wholly or partially match the sequence of our target protein. There are many softwares that perform this task. One prominent software is called “Blast”. It is available in the form of webservice and it can also be downloaded. when a query protein is run through Blast, a list of templates is obtained, the number of templates can be from few ones to few hundreds depending on how many matches are found.

## Generating the initial model

At this stage we can either use a single template or multiple templates to build the initial model. We opted to use one template with the top coverage score less than 90%. Every atom in this top template that matches an atom in the target protein will have its coordinates mapped to the target protein. An obvious situation arises here, which is gaps. Gaps are parts of the target protein sequence for which no match was found in the top template. For each gap, we search the subsequent templates to find a match. For the gap to be filled with the appropriate residues, we require that the residues found need to have the same residues (one before and one after) before and after the gap. This is illustrated in figure2 below. A superposition of the residue coordinates will be performed if necessary.

![A screenshot of a cell phone

Description generated with high confidence]()

Figure 2 Gap handling when generating the initial model. The target protein sequence

## Refining the model

We can stop after the initial model is generated or we can try to refine it to try to bring it closer to the native structure. There are different ways to perform the refinement. So, since this is template based approach, we relay again on the templates we generated using blast. However, this time we use the coordinates of the atoms in the pdb file to generate distances between residues within the sequence. These distances serve as a restraint. These distances are shown in figure 3.

![A close up of a logo

Description generated with high confidence]()

Figure 3 distance restraints between residues within sequence

# Data

Our approach is tested on three proteins from CASP12. Table1 contains some basic information about these proteins.

Table 1 List of the three proteins tested and some basic information about them

|  |  |  |  |
| --- | --- | --- | --- |
| protein | NO of residues | domains | Range of domain |
| T0860 | 137 | 1 | 1-136 |
| T0889 | 242 | 1 | 4-242 |
| T0921 | 149 | 1 | 5-142 |

# Experiment

# Results and discussion

# conclusion