CONFOLD NEW VERSION: CONTACT-GUIDED AB INITIO PROTEIN FOLDING WITH NEW FEATURES

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In partial Fulfillment

of the Requirements for the Degree

Master of Science

by

XIANGYU LI

Dr. Jianlin Cheng, Thesis Advisor

DECEMBER 2019

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

CONFOLD VERSION 3: CONTACT-GUIDED AB INITIO PROTEIN FOLDING WITH NEW FEATURES

Presented by Xiangyu Li, a candidate for the degree of Master of Science and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Jianlin Cheng

Dr. Yunxin Zhao

Dr. Yuyi Lin

ACKNOWLEDGEMENTS

First and foremost, I would like to show my best regards to my thesis advisor Dr. Jianlin Cheng for his valuable guidance and supervision. He pointed out the direction for me and helped me in every stage of my research study. I would also like to thank my committee members, Dr. Yunxin Zhao and Dr. Yuyi Lin. Both are my most respected teachers.

I also want to thank my senior lab mates, who helped me with this research project. They helped me to get familiar with the CNS solve and provided me with a lot of guidance on biological knowledge.

Finally, I am especially grateful to my family for providing me with continued support and encouragement. I am so thankful for having you all.

TABLE OF CONTENTS

TABLE OF FIGURES

NOMENCLATURE

CONFOLD: ab initio protein folding method.

CASP 12: Protein datasets.

CNS: Crystallography and NMR System.

MSA: Multiple sequence alignments.

APC: Average product correction.

ABSTRACT

CONFOLD is an ab initio protein folding method that can build three-dimensional models using predicted contacts and secondary structures. Under this method, we can translate contact distance map and secondary structure into the distance, dihedral angle, and hydrogen bond restraints according to a set of new conversion rules, and then using this information as input to build structure models.

To improve this method, we added some new features to CONFOLD, such as disulfide bond information, Beta contact prediction, and contacts distance multi-threshold. CONFOLD New Version allows using disulfide bond information and Beta strands prediction as input so that the Crystallography and NMR System can get the information directly, improving the accuracy and efficiency in some specific cases. And it can exclude some low probability residues contact information by setting multi-thresholds. I tested this method based on CASP 12 datasets, and results show that it can improve the efficiency of the program while keeping the TM-score.

Chapter 1 Introduction

CONFOLD is a method that can predict new protein folds using contact-guided protein modeling [1]. It accepts contacts distance map, secondary structure information as input to build three-dimensional models. When the predicted contacts are accurate, the CONFOLD method can generate high-quality tertiary structures. It reconstructs models from predicted contacts based on the Crystallography & NMR System (CNS), which is a method designed for building models from Nuclear Magnetic Resonance (NMR) experimental data.

There are some other tools, such as IMP [4] and Tinker, that can use different kinds of contact distance restraints to build models, but in some specific cases, these tools have some particular limitations. For example, they cannot reduce the low probability contacts from the distance map. Even the Modeller, which is used widely for reconstruction, cannot work on template-free modeling.

CONFOLD designed two stages to overcome these disadvantages. In stage one, it can use contacts distance and secondary structure information to reconstruct protein models, then filter out the information that does not match the conversion rules to ensure high quality.

In stage two, it takes updated distance restraints and secondary structure as input to generate models using CNS suite [6], and select the best model for evaluation.

In this research, we added some new features into CONFOLD, included disulfide bond information, beta contacts prediction, and multi-threshold contacts probability.

Disulfide bonds in protein can be found in both bacteria and eukaryotes. We choose to use DIpro2 to predict disulfide bonds based on a 2D recurrent neural network. And the beta contact prediction can be completed by bbcontacts, which is used for the prediction of β-strand pairing from direct coupling patterns.

Chapter 2 Crystallography & NMR System

2.1 Background.

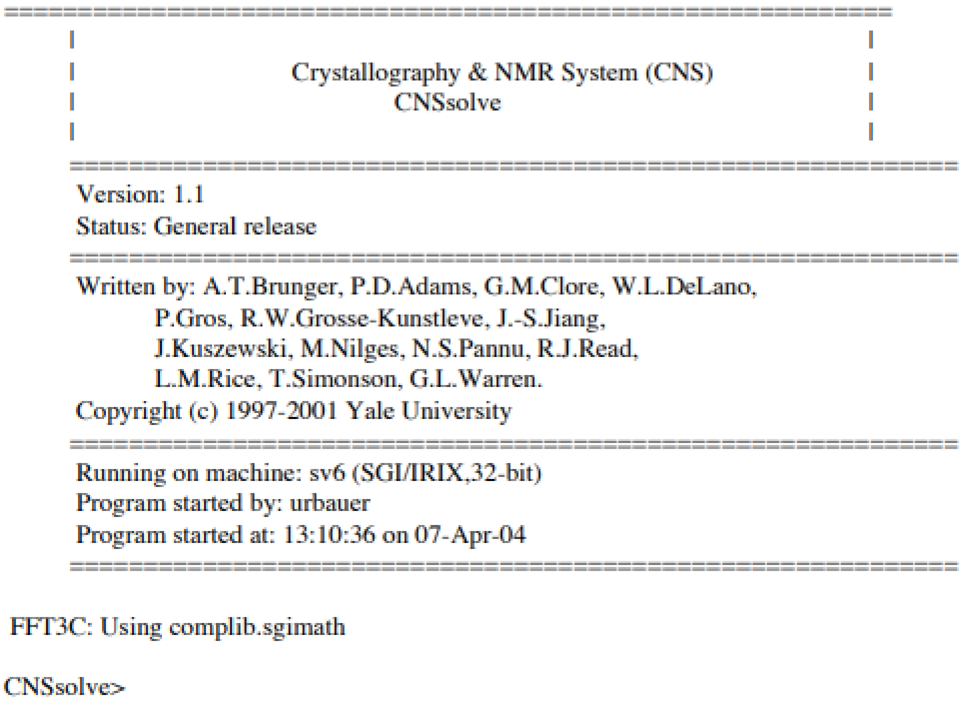
Crystallography & NMR System (CNS) is designed to provide a flexible multi-level hierarchical approach for the most commonly used algorithm in macromolecular structure determination. The CONFOLD is built based on this system.

The CNS can build models from Nuclear Magnetic Resonance experimental data and reconstruct protein models from predicted contacts. In this research, our first step is to get familiar with the CNS, knowing how it gets the distance restraints between atoms — and then adding the desired new features on CONFOLD.

There are three CNS files used in CONFOLD to reconstruct the protein models: “gesq.inp”, “extn.inp”, “dgsa.inp”.

- Gesq.inp: Generate structure file for protein from sequence information only.
- Extn.inp: Generate an extended strand with ideal geometry for rach connected polymer.
- Dgsa.inp: Distance geometry with simulated annealing regularization starting from extended strand.

And in the “extn.inp”, the molecular structure cannot include any closed loops except disulfide bonds. Because disulfide bonds can be automatically excluded from the generation of the strand conformation. This file is a CNS macro for generating extended polypeptide chains as starting structures for our calculations.



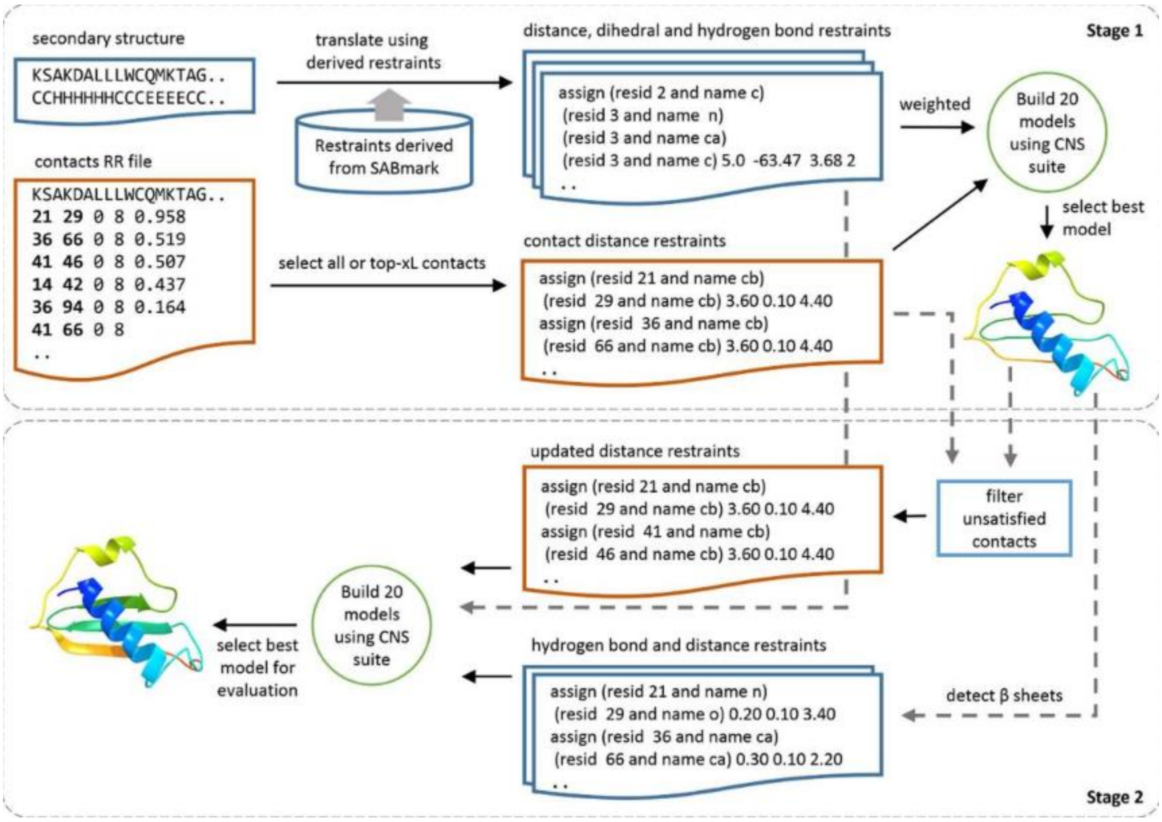
```
============================================================
    |                                                    |
    |          Crystallography & NMR System (CNS)        |
    |                      CNSsolve                      |
    |                                                    |
    ============================================================
     Version: 1.1
     Status: General release
    ============================================================
     Written by: A.T.Brunger, P.D.Adams, G.M.Clore, W.L.DeLano,
                 P.Gros, R.W.Grosse-Kunstleve, J.-S.Jiang,
                 J.Kuszewski, M.Nilges, N.S.Pannu, R.J.Read,
                 L.M.Rice, T.Simonson, G.L.Warren.
     Copyright (c) 1997-2001 Yale University
    ============================================================
     Running on machine: sv6 (SGI/IRIX,32-bit)
     Program started by: urbauer
     Program started at: 13:10:36 on 07-Apr-04
    ============================================================

 FFT3C: Using complib.sgimath

CNSsolve>
```

Figure2.1: The interactive mode of CNS.

CNS can run in two modes: interactive mode or non-interactive mode. Figure 2.1 shows the interactive way, and in this mode, we can see all the output of the program, and you can exit the system by typing “stop” or “return” at the CNS solve prompt.


Stage 1
secondary structure
translate using derived restraints
Restraints derived from SABmark
distance, dihedral and hydrogen bond restraints
weighted
Build 20 models using CNS suite
select best model
contacts RR file
select all or top-xL contacts
contact distance restraints
updated distance restraints
filter unsatisfied contacts
Build 20 models using CNS suite
select best model for evaluation
hydrogen bond and distance restraints
detect β sheets
Stage 2

```
KSAKDALLLWCQMKTAG..
CCHHHHHHCCCEEEECC..
```

```
KSAKDALLLWCQMKTAG..
21 29 0 8 0.958
36 66 0 8 0.519
41 46 0 8 0.507
14 42 0 8 0.437
36 94 0 8 0.164
41 66 0 8
..
```

```
assign (resid 2 and name c)
 (resid 3 and name n)
 (resid 3 and name ca)
 (resid 3 and name c) 5.0 -63.47 3.68 2
..
```

```
assign (resid 21 and name cb)
 (resid 29 and name cb) 3.60 0.10 4.40
assign (resid 36 and name cb)
 (resid 66 and name cb) 3.60 0.10 4.40
..
```

```
assign (resid 21 and name cb)
 (resid 29 and name cb) 3.60 0.10 4.40
assign (resid 41 and name cb)
 (resid 46 and name cb) 3.60 0.10 4.40
..
```

```
assign (resid 21 and name n)
 (resid 29 and name o) 0.20 0.10 3.40
assign (resid 36 and name ca)
 (resid 66 and name ca) 0.30 0.10 2.20
..
```

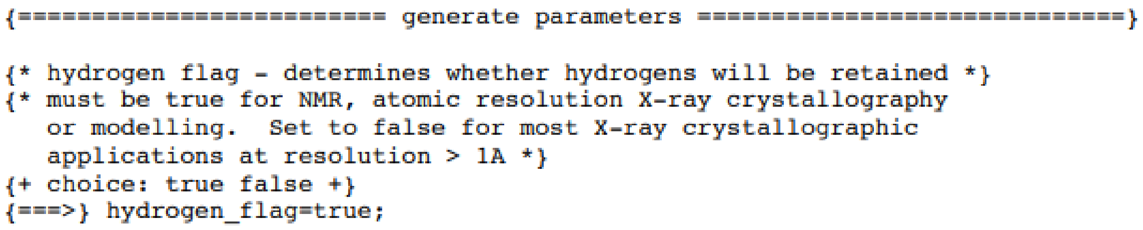
Figure2.2: Process of CONFOLD.

The CONFOLD is built based on the CNS system. The input files are secondary structure files and contact RR files. It can translate the secondary structure file using derived restraints and generate the dihedral and hydrogen bond restraints information, and then select top-xL contacts from contacts RR file and create the contact distance restraints. Using those two restraints, it can build 20 models using the CNS suite. And selecting best models, filtering unsatisfied contacts, and detecting beta-strands. Finally, we can get the best model for evaluation.

2.2 Relationship between new features and CNS.

In this research, one of our goals is to enable CONFOLD to identify the disulfide bonds prediction and β-sheet contacts prediction. From the introduction, we know that the “gesq.inp” file is used to generate structure file for protein from sequence information only, so it is a CNS macro for creating a molecular topology file for our molecules.

In this file we can define a disulfide bond between cysteine residues in protein or between protein segments, and it includes the hydrogen flag, which determines whether the hydrogens will be retained.

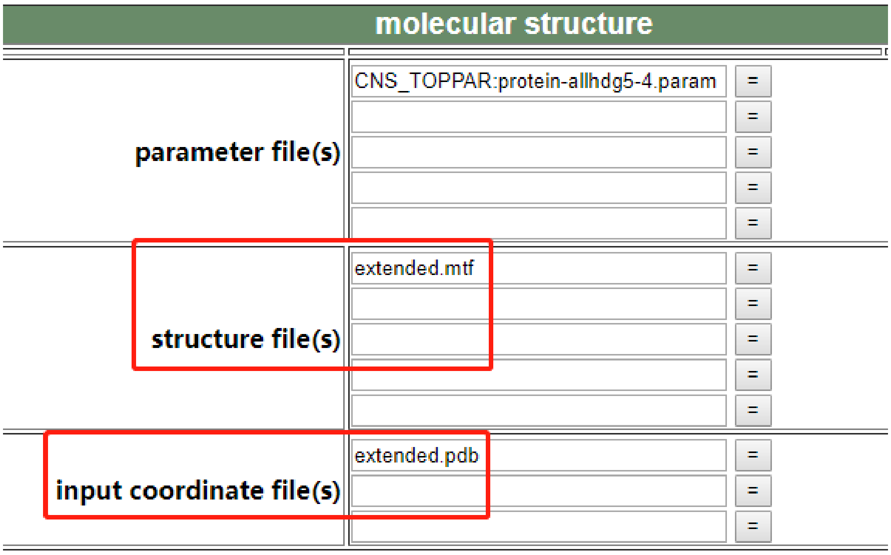


```
{========================= generate parameters =============================}

{* hydrogen flag - determines whether hydrogens will be retained *}
{* must be true for NMR, atomic resolution X-ray crystallography
   or modelling.  Set to false for most X-ray crystallographic
   applications at resolution > 1A *}
{+ choice: true false +}
{===>} hydrogen_flag=true;
```

Figure2.3: The hydrogen flag in CNS.

In the “dgsa.inp” file, there are also some crucial parameters.



molecular structure

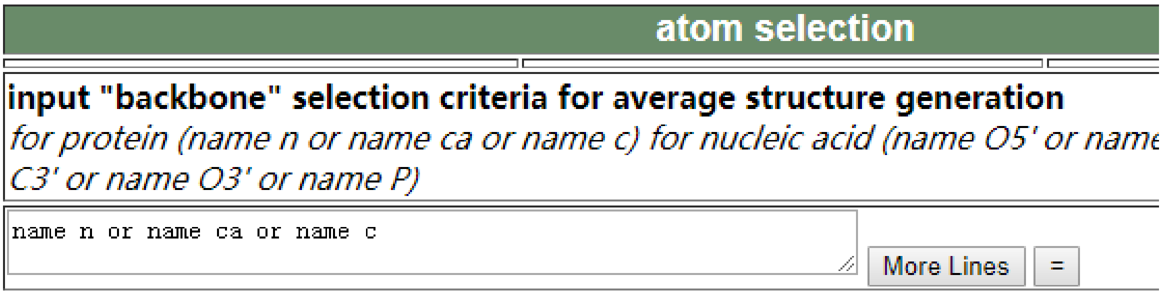
parameter file(s): CNS_TOPPAR:protein-allhdg5-4.param

structure file(s): extended.mtf

input coordinate file(s): extended.pdb

Figure2.4: Molecular structure file in CNS.

Figure 2.4 shows the structure file and input coordinate file required by CNS. The structure file is “extended.mtf” which contains the information describing the topology of the molecule. And the molecular topology file cannot be edited manually. The coordinate input file is “extended.pdb” which contains the atomic coordinates in PDB type format.



atom selection

input "backbone" selection criteria for average structure generation

for protein (name n or name ca or name c) for nucleic acid (name O5' or name C3' or name O3' or name P)

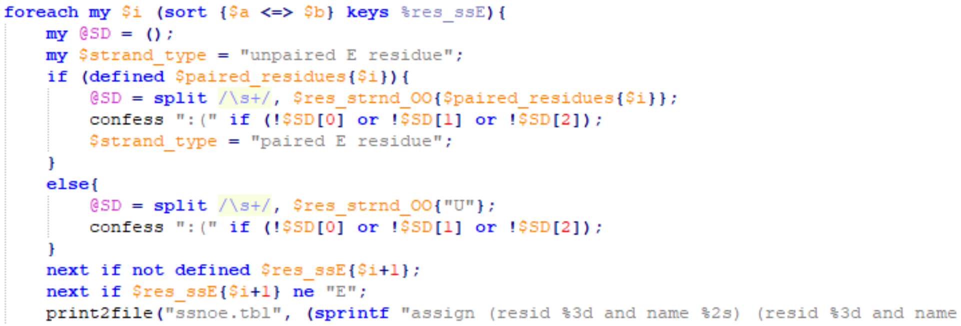
```
name n or name ca or name c
```

Figure2.5: The atom selection in CNS.

Figure 2.5 shows how to define the atom selection in CNS. Atom selection identifying the "backbone" atoms for average structure generation. For the protein molecules, the format is:

$$(name \quad n \quad or \quad name \quad ca \quad or \quad name \quad c)$$

After the atom selection, CONFOLD needs to read the information from the contacts restraints RR file and exclude the unsatisfied pairs. Then generating the generic restraints which are required by the next step.



```
foreach my $i (sort {$a <=> $b} keys %res_ssE){
    my @SD = ();
    my $strand_type = "unpaired E residue";
    if (defined $paired_residues{$i}){
        @SD = split /\s+/, $res_strnd_OO{$paired_residues{$i}};
        confess ":(" if (!$SD[0] or !$SD[1] or !$SD[2]);
        $strand_type = "paired E residue";
    }
    else{
        @SD = split /\s+/, $res_strnd_OO{"U"};
        confess ":(" if (!$SD[0] or !$SD[1] or !$SD[2]);
    }
    next if not defined $res_ssE{$i+1};
    next if $res_ssE{$i+1} ne "E";
    print2file("ssnoe.tbl", (sprintf "assign (resid %3d and name %2s) (resid %3d and name
```

Figure2.6: CONFOLD generates generic restraints.

In the code, it will identify the strands that are not used for pairing and generate generic restraints for them.

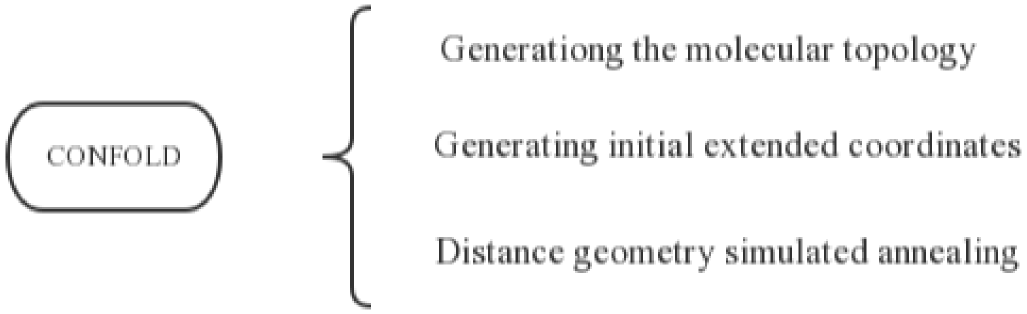
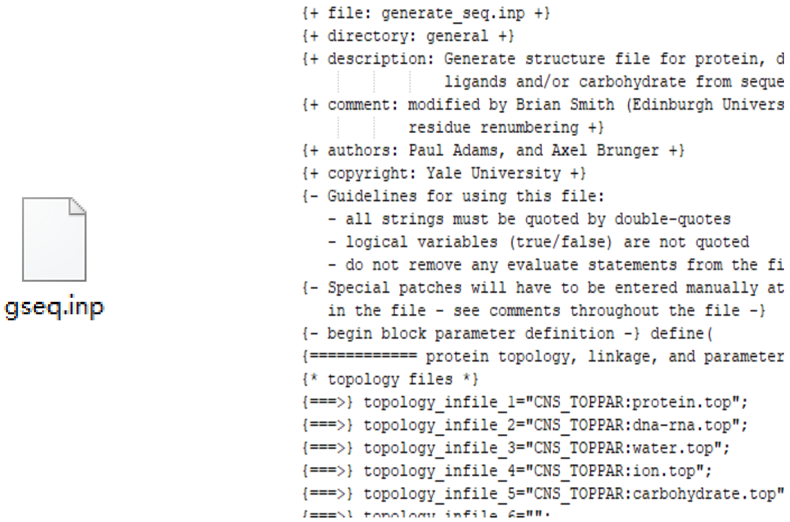

CONFOLD
Generationg the molecular topology
Generating initial extended coordinates
Distance geometry simulated annealing

Figure2.7: CONFOLD in the CNS system.

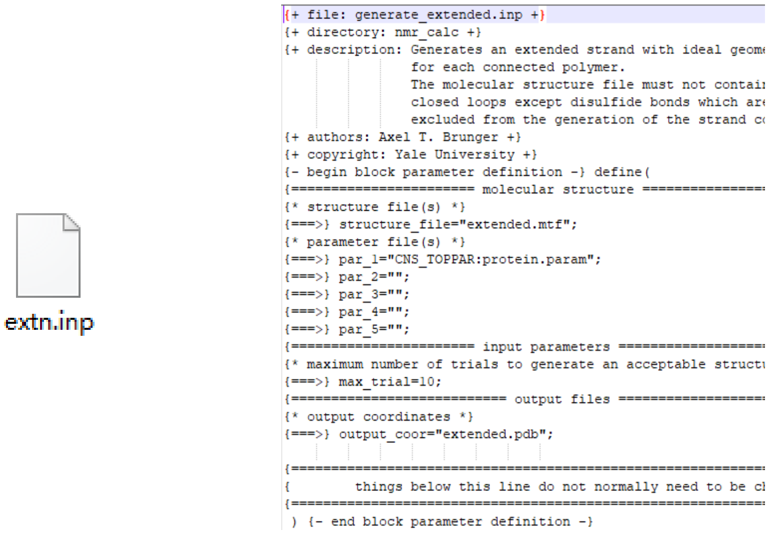
The CONFOLD using CNS solve to reconstruct the model based on three functions. Generation of the molecular topology, generation of the initial extended coordinates, and distance geometry simulated annealing.



```
{+ file: generate_seq.inp +}
{+ directory: general +}
{+ description: Generate structure file for protein, d
                ligands and/or carbohydrate from seque
{+ comment: modified by Brian Smith (Edinburgh Univers
            residue renumbering +}
{+ authors: Paul Adams, and Axel Brunger +}
{+ copyright: Yale University +}
{- Guidelines for using this file:
   - all strings must be quoted by double-quotes
   - logical variables (true/false) are not quoted
   - do not remove any evaluate statements from the fi
{- Special patches will have to be entered manually at
   in the file - see comments throughout the file -}
{- begin block parameter definition -} define(
{============ protein topology, linkage, and parameter
{* topology files *}
{===>} topology_infile_1="CNS_TOPPAR:protein.top";
{===>} topology_infile_2="CNS_TOPPAR:dna-rna.top";
{===>} topology_infile_3="CNS_TOPPAR:water.top";
{===>} topology_infile_4="CNS_TOPPAR:ion.top";
{===>} topology_infile_5="CNS_TOPPAR:carbohydrate.top"
{===>} topology_infile_6="";
```

Figure2.8: Generation of the molecular topology.

The molecular topology information [11] must be first generated for the structure - this contains the information about molecular connectivity. This information is then be used in the next step to create extended conformation.

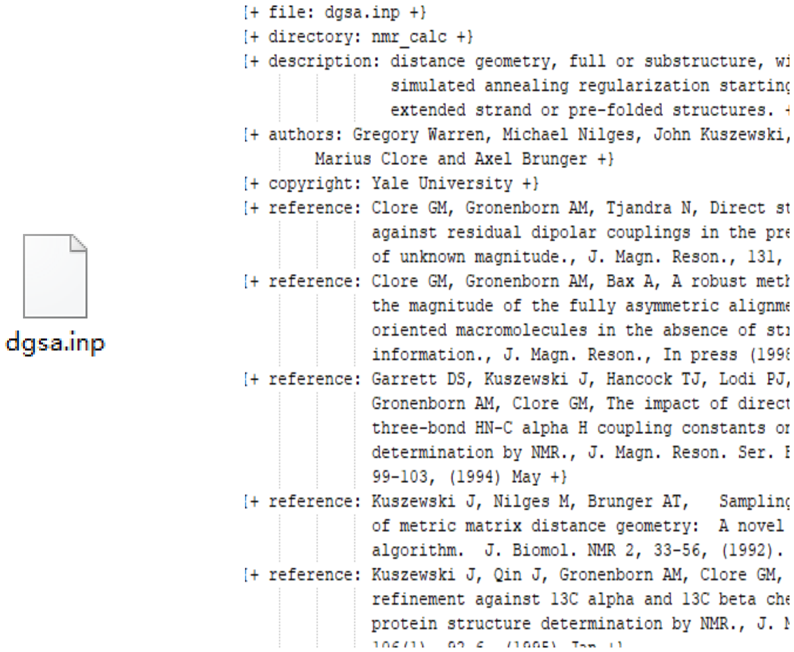


```
{+ file: generate_extended.inp +}
{+ directory: nmr_calc +}
{+ description: Generates an extended strand with ideal geom
                for each connected polymer.
                The molecular structure file must not contai
                closed loops except disulfide bonds which ar
                excluded from the generation of the strand c
{+ authors: Axel T. Brunger +}
{+ copyright: Yale University +}
{- begin block parameter definition -} define(
{====================== molecular structure ===============
{* structure file(s) *}
{===>} structure_file="extended.mtf";
{* parameter file(s) *}
{===>} par_1="CNS_TOPPAR:protein.param";
{===>} par_2="";
{===>} par_3="";
{===>} par_4="";
{===>} par_5="";
{====================== input parameters ==================
{* maximum number of trials to generate an acceptable struct
{===>} max_trial=10;
{========================== output files ==================
{* output coordinates *}
{===>} output_coor="extended.pdb";

{===========================================================
{        things below this line do not normally need to be c
{===========================================================
 ) {- end block parameter definition -}
```

Figure2.9: Generation of the initially extended coordinates.

Because the structure calculation needs a starting model, so the next step is for the starting model. It provides proper local geometry but contains no information about the fold of the structure.



```
[+ file: dgsa.inp +}
[+ directory: nmr_calc +}
[+ description: distance geometry, full or substructure, wi
                 simulated annealing regularization startin
                 extended strand or pre-folded structures. +
[+ authors: Gregory Warren, Michael Nilges, John Kuszewski,
        Marius Clore and Axel Brunger +}
[+ copyright: Yale University +}
[+ reference: Clore GM, Gronenborn AM, Tjandra N, Direct st
               against residual dipolar couplings in the pre
               of unknown magnitude., J. Magn. Reson., 131,
[+ reference: Clore GM, Gronenborn AM, Bax A, A robust meth
               the magnitude of the fully asymmetric alignme
               oriented macromolecules in the absence of str
               information., J. Magn. Reson., In press (199
[+ reference: Garrett DS, Kuszewski J, Hancock TJ, Lodi PJ,
               Gronenborn AM, Clore GM, The impact of direct
               three-bond HN-C alpha H coupling constants on
               determination by NMR., J. Magn. Reson. Ser. B
               99-103, (1994) May +}
[+ reference: Kuszewski J, Nilges M, Brunger AT,   Sampling
               of metric matrix distance geometry:  A novel
               algorithm.  J. Biomol. NMR 2, 33-56, (1992).
[+ reference: Kuszewski J, Qin J, Gronenborn AM, Clore GM,
               refinement against 13C alpha and 13C beta che
               protein structure determination by NMR., J. M
               106(1), 93-6, (1995) Jan +}
```

Figure2.10: Distance geometry simulated annealing.

And the last one is for distance geometry simulated annealing. Here a structure is calculated using experimentally measured interproton distance estimates, hydrogen bonds, and coupling-constant-derived dihedral angle restraints.

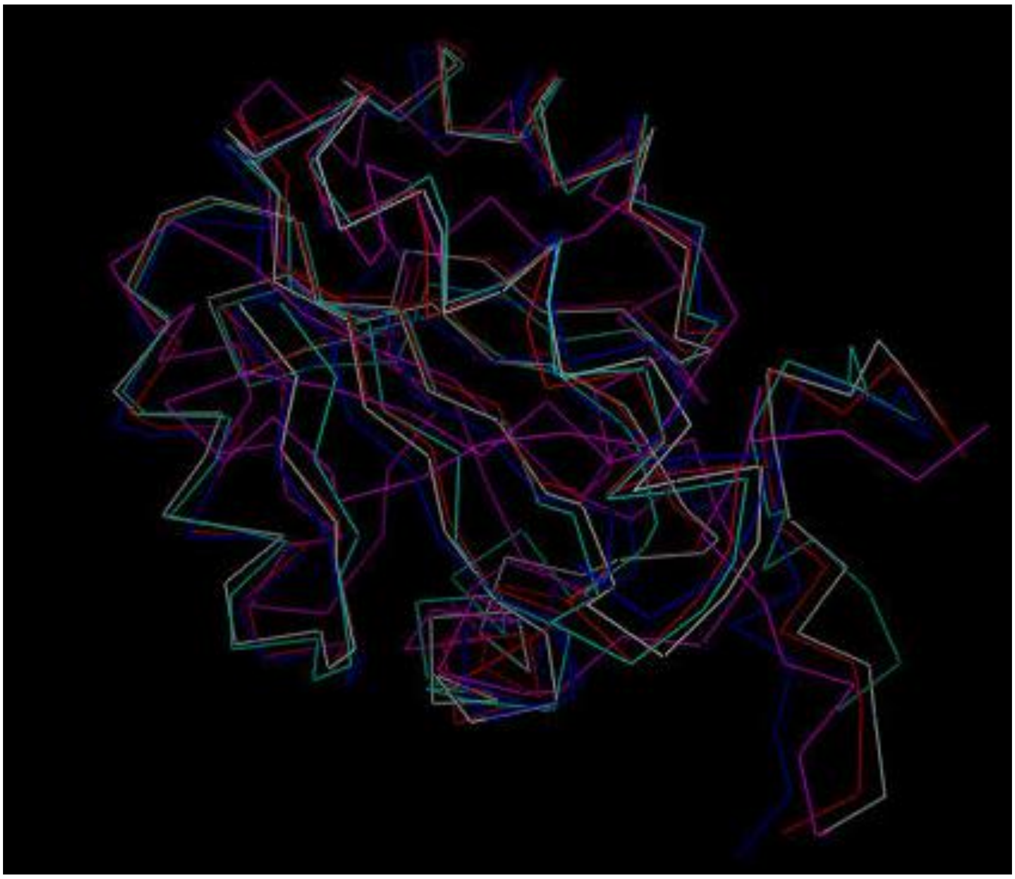


Figure2.11: Five structures after simulated annealing refinement.

After simulated annealing refinement, a summary of the structure calculation is written at the top of each output PDB file. The information about violations can be used to select acceptable structures.

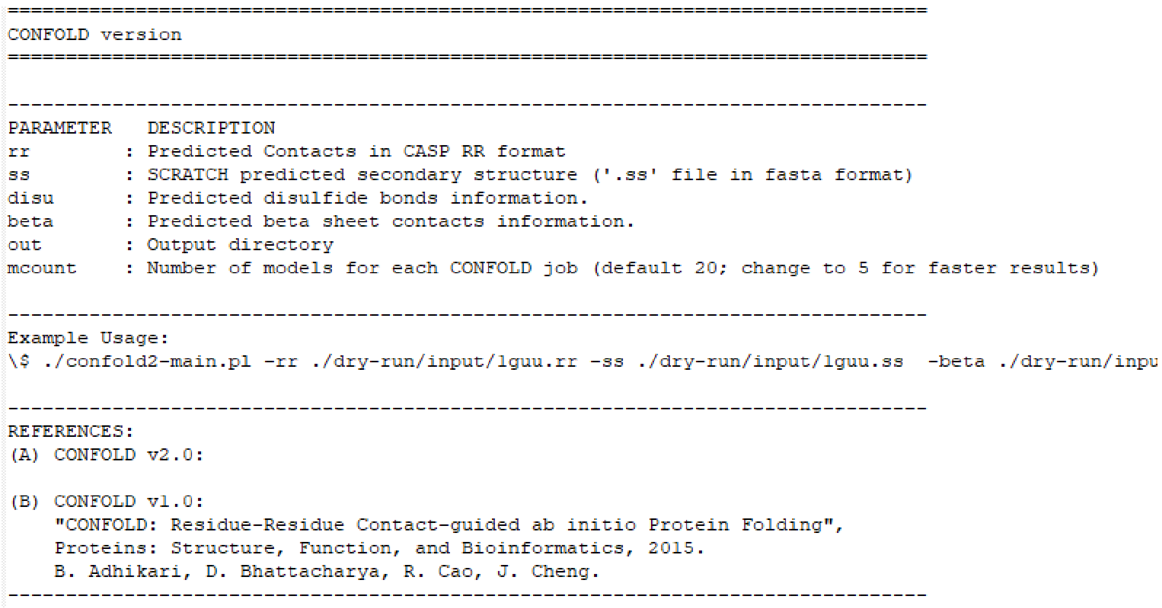
2.2 Preparing work.

Before we start our work, we need to promise that the dssp-2.0.4 linux kernel [7] exists. The DSSP algorithm is the standard method for assigning secondary structure to the amino acids of a protein, given the atomic-resolution coordinates of the protein.

It can identify the intra-backbone hydrogen bonds of the protein using a purely electrostatic definition. A hydrogen bond is identified as:

$$\mathrm{E} = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} * 332\ kcal/mol$$

After installing the dssp kernel, we need to determine the usage of our system. In this version, we required four different inputs: predicted contacts in CASP RR format [14], predicted secondary structure file, predicted beta-sheet contacts file, and predicted disulfide bond information file.



```
================================================================================
CONFOLD version
================================================================================

--------------------------------------------------------------------------------
PARAMETER   DESCRIPTION
rr         : Predicted Contacts in CASP RR format
ss         : SCRATCH predicted secondary structure ('.ss' file in fasta format)
disu       : Predicted disulfide bonds information.
beta       : Predicted beta sheet contacts information.
out        : Output directory
mcount     : Number of models for each CONFOLD job (default 20; change to 5 for faster results)

--------------------------------------------------------------------------------
Example Usage:
\$ ./confold2-main.pl -rr ./dry-run/input/1guu.rr -ss ./dry-run/input/1guu.ss  -beta ./dry-run/inpu

--------------------------------------------------------------------------------
REFERENCES:
(A) CONFOLD v2.0:

(B) CONFOLD v1.0:
    "CONFOLD: Residue-Residue Contact-guided ab initio Protein Folding",
    Proteins: Structure, Function, and Bioinformatics, 2015.
    B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng.
--------------------------------------------------------------------------------
```

Figure2.12: CONFOLD new version usage.

Under this usage, we can run the system and test our results based on the CASP 12 dataset.

Chapter 3　Disulfide Bonds

3.1 Background

Disulfide bonds are relatively stable covalent bonds and are usually responsible for stabilizing tertiary structures of proteins. In biochemistry, the disulfide bond is used to describe the terminology R-S-S-R connectivity [15]. The most common way of creating this bond is by oxidation of sulfhydryl groups. The length of the disulfide bond is 2.05 Å, and the dissociation energy of a disulfide bond is 60 kcal/mole. We choose to use DIpro2 [2], the disulfide bond prediction tools to get the information. And processing this result as input, to modify the Generate Structure file of CNS solve.

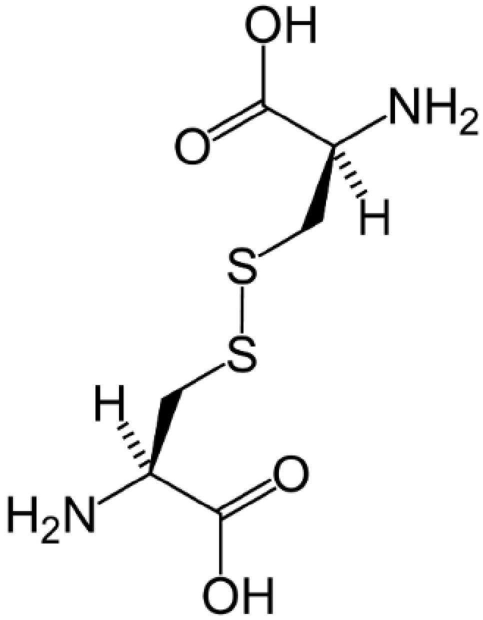


Figure3.1: Two cysteine residues linked by a disulfide bond to form cystine.

3.2 Disulfide bonds information.

CONFOLD using CNS solve to reconstruct protein models, and the CNS system required the molecular topology information must be first generated. It means that the disulfide bond information must be prepared in the first stage of CONFOLD.

According to the CNS system, it can automatically detect the disulfide bonds based on the distance between the sulfur atoms, which are less than 3Å. But if we want to pursue more reliable results, the molecular topology file should be modified as input changes.

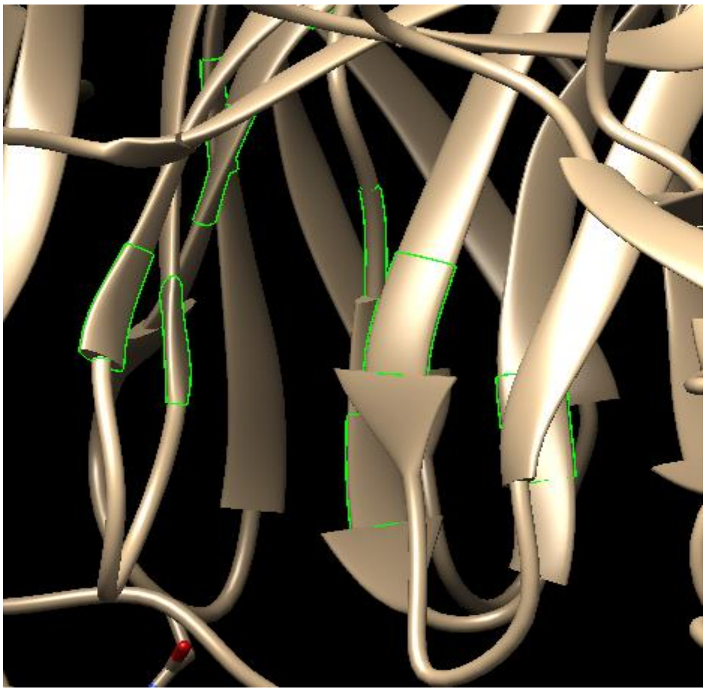
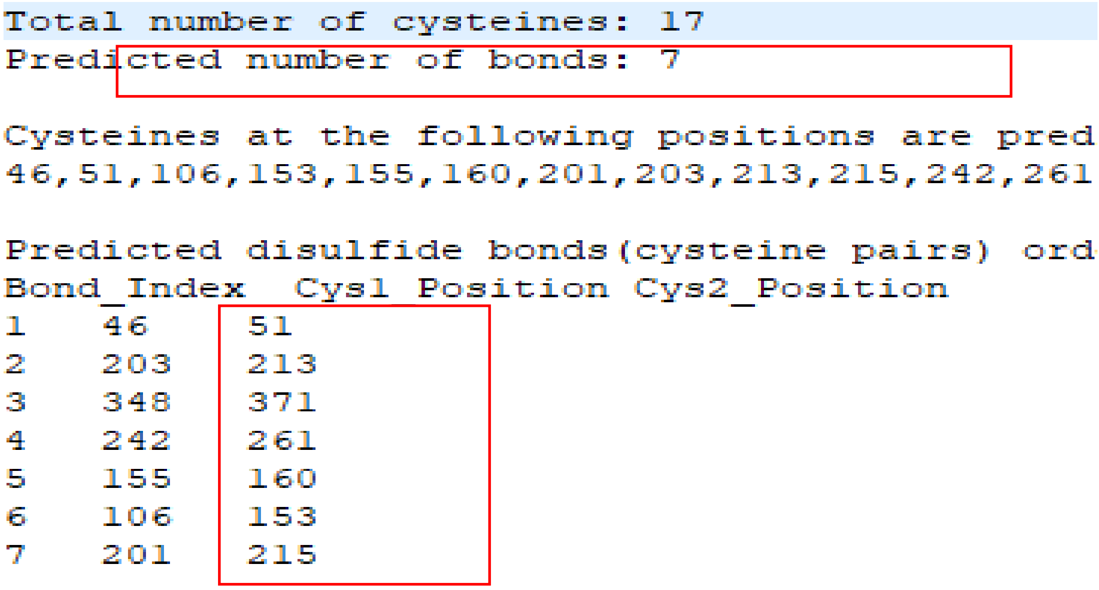


Figure3.2: The disulfide bonds in protein 1a4g.

DIpro2 is a cysteine disulfide bond predictor, and it can predict if the sequence has disulfide bonds or not and predict the bonding state of each cysteine and the bonded pairs.

What we need from the prediction is the total number of cysteines in sequence and the positions of cysteines, which are predicted to form disulfide bonds.



```
Total number of cysteines: 17
Predicted number of bonds: 7

Cysteines at the following positions are pred
46,51,106,153,155,160,201,203,213,215,242,261

Predicted disulfide bonds(cysteine pairs) ord
Bond_Index  Cys1 Position Cys2_Position
1    46     51
2    203    213
3    348    371
4    242    261
5    155    160
6    106    153
7    201    215
```

Figure3.3: Prediction of the disulfide bonds based on DIpro2.

Figure 3.3 shows the format of DIpro2 Prediction. In this file, there two crucial pieces of information that can be used in the next step. The first one is the predicted number of bonds. Based on this number, we can determine how many inputs we need to write into CNS. And the second one is the cysteines' position, which is required by the CNS system to build the disulfide bonds.

3.3 Processing of disulfide bonds information.

The prediction of disulfide bond information cannot be used directly, and we developed two subfunctions to process the result file. First, recognizing if the sequence contains disulfide bonds. If the number of disulfide pairs is more than one, we can read the position of the cysteine into hash. Then we need to check the distance between the two cysteines. If there are no disulfide bonds in the sequence, we will generate a list of which flag is “false”, the position of cysteines is 0, and the confirmed cysteines will be written into the list. After generating a list of disulfide bonds information, we need to modify the “gesq.inp” file for creating a molecular topology. The CNS system is divided into two segments, with segment identifiers “A” and “B”.

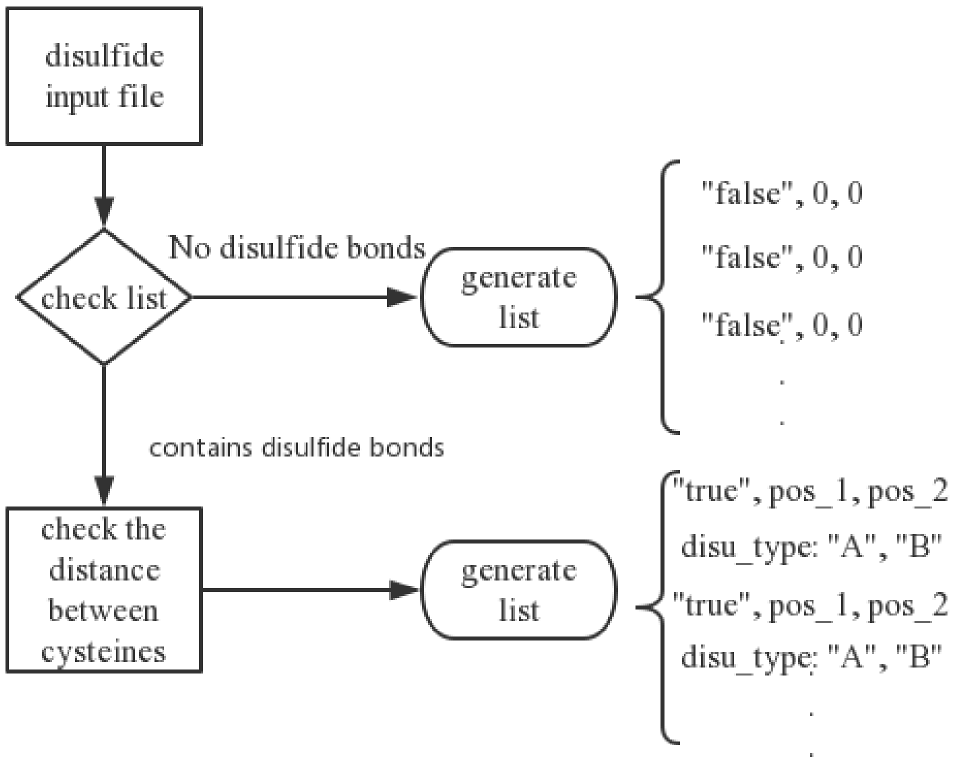
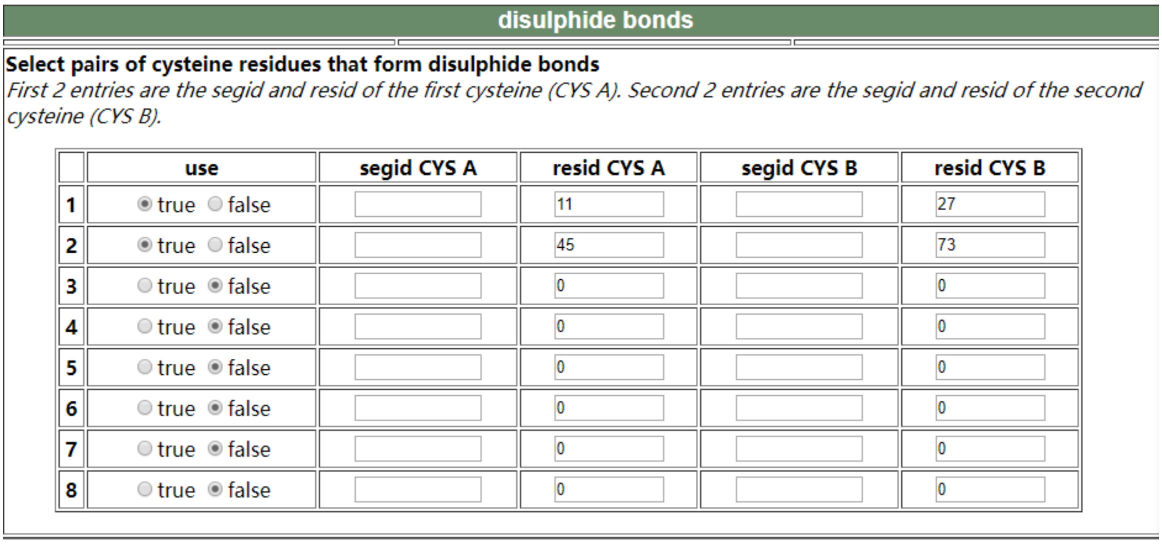


Figure3.4: Processing of disulfide bond information.

Running with the “gesq.inp” file, CNS can generate a molecular topology file named “trx.mtf”. It can record the two protein molecules connected by a disulfide bond. In this file, the first information we can see is information concerning the identity of each atom and atomic charge and mass. Next, we can still see in this system how each atom is connected to other atoms.



disulphide bonds

Select pairs of cysteine residues that form disulphide bonds
First 2 entries are the segid and resid of the first cysteine (CYS A). Second 2 entries are the segid and resid of the second cysteine (CYS B).

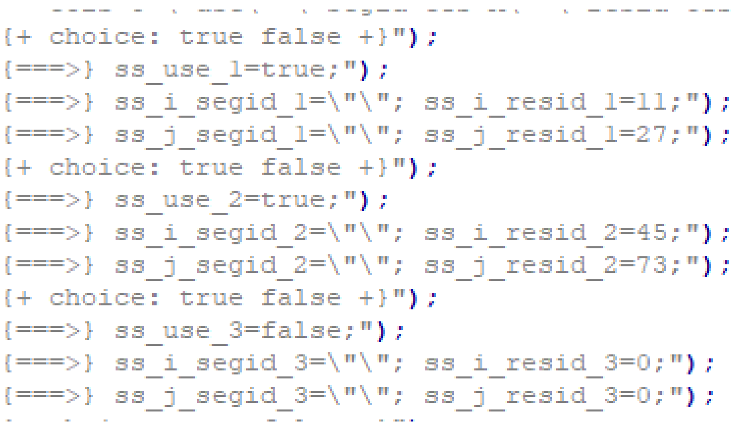
	use	segid CYS A	resid CYS A	segid CYS B	resid CYS B
1	● true ○ false		11		27
2	● true ○ false		45		73
3	○ true ● false		0		0
4	○ true ● false		0		0
5	○ true ● false		0		0
6	○ true ● false		0		0
7	○ true ● false		0		0
8	○ true ● false		0		0

Figure3.5: The disulfide bonds part in the CNS system.

Figure 3.5 shows the interface of the disulfide bonds part in the CNS system. In the first column, there is a flag specifying whether a disulfide bond should be created between the specified residues. And we can set the flag to true or false. If true, we need to fill in the columns “resid CYS” and “segid CYS”. The “resid CYS” is the number specifying the residue for cysteine in a disulfide bond, and the “segid CYS” is the string specifying the segment identifier for cysteine in a disulfide bond.

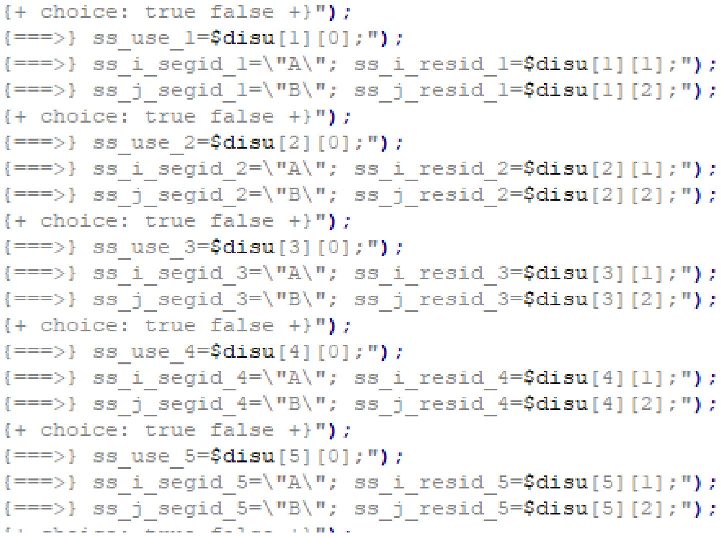
3.4 Comparing with original methods.

In CNS solve, the molecular topology information must be first generated for the structure. Because this information is then be used in the next step to create starting coordinates (extended PDB). CONFOLD Version 2 cannot get the information of cysteine residues from disulfide bonds. It selects two pairs of cysteines and never changes. In the new version, we can improve our accuracy with DIpro2's prediction and make full use of the functions provided by CNS solve.



```
{+ choice: true false +}");
{===>} ss_use_1=true;");
{===>} ss_i_segid_1=\"\"; ss_i_resid_1=11;");
{===>} ss_j_segid_1=\"\"; ss_j_resid_1=27;");
{+ choice: true false +}");
{===>} ss_use_2=true;");
{===>} ss_i_segid_2=\"\"; ss_i_resid_2=45;");
{===>} ss_j_segid_2=\"\"; ss_j_resid_2=73;");
{+ choice: true false +}");
{===>} ss_use_3=false;");
{===>} ss_i_segid_3=\"\"; ss_i_resid_3=0;");
{===>} ss_j_segid_3=\"\"; ss_j_resid_3=0;");
```

Figure3.6: Molecular topology file cannot use prediction.

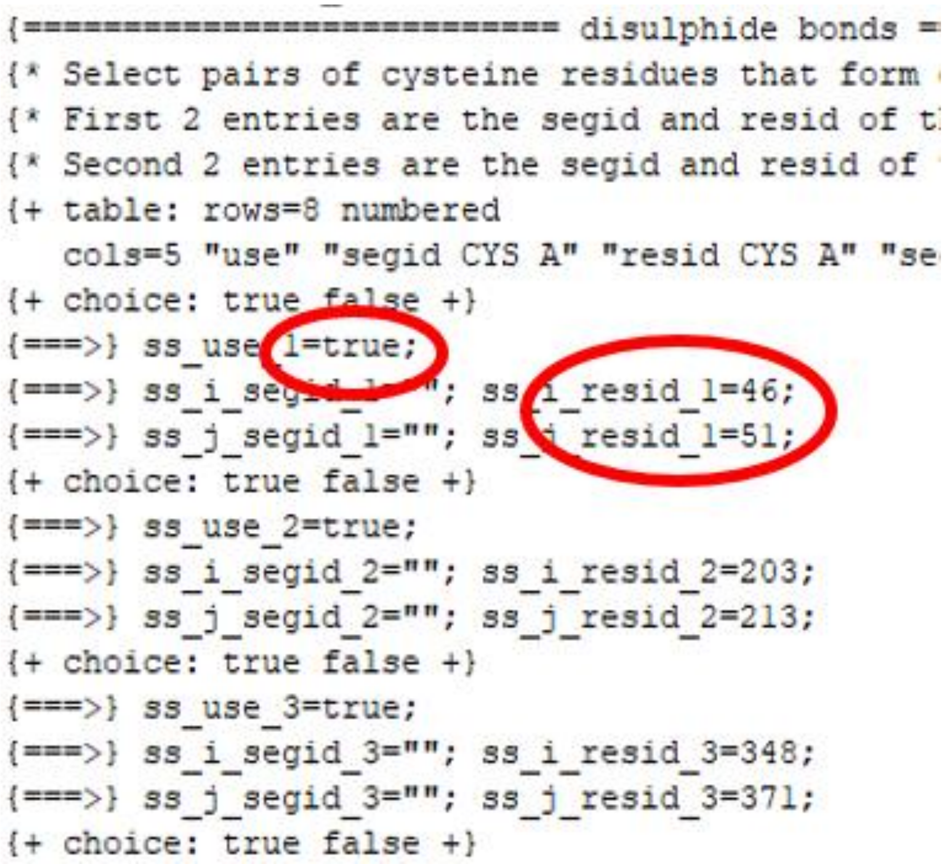


```
{+ choice: true false +}");
{===>} ss_use_1=$disu[1][0];");
{===>} ss_i_segid_1=\"A\"; ss_i_resid_1=$disu[1][1];");
{===>} ss_j_segid_1=\"B\"; ss_j_resid_1=$disu[1][2];");
{+ choice: true false +}");
{===>} ss_use_2=$disu[2][0];");
{===>} ss_i_segid_2=\"A\"; ss_i_resid_2=$disu[2][1];");
{===>} ss_j_segid_2=\"B\"; ss_j_resid_2=$disu[2][2];");
{+ choice: true false +}");
{===>} ss_use_3=$disu[3][0];");
{===>} ss_i_segid_3=\"A\"; ss_i_resid_3=$disu[3][1];");
{===>} ss_j_segid_3=\"B\"; ss_j_resid_3=$disu[3][2];");
{+ choice: true false +}");
{===>} ss_use_4=$disu[4][0];");
{===>} ss_i_segid_4=\"A\"; ss_i_resid_4=$disu[4][1];");
{===>} ss_j_segid_4=\"B\"; ss_j_resid_4=$disu[4][2];");
{+ choice: true false +}");
{===>} ss_use_5=$disu[5][0];");
{===>} ss_i_segid_5=\"A\"; ss_i_resid_5=$disu[5][1];");
{===>} ss_j_segid_5=\"B\"; ss_j_resid_5=$disu[5][2];");
```

Figure3.7: Molecular topology file modified based on DIpro2 prediction.

In figure 3.6 and figure 3.7, you can see that the cysteine residues information in the original version is hard to code. Users cannot modify the flag and the residue position.

After adding the new feature, we can read the prediction information from the input file and then modify the parameters of CNS.



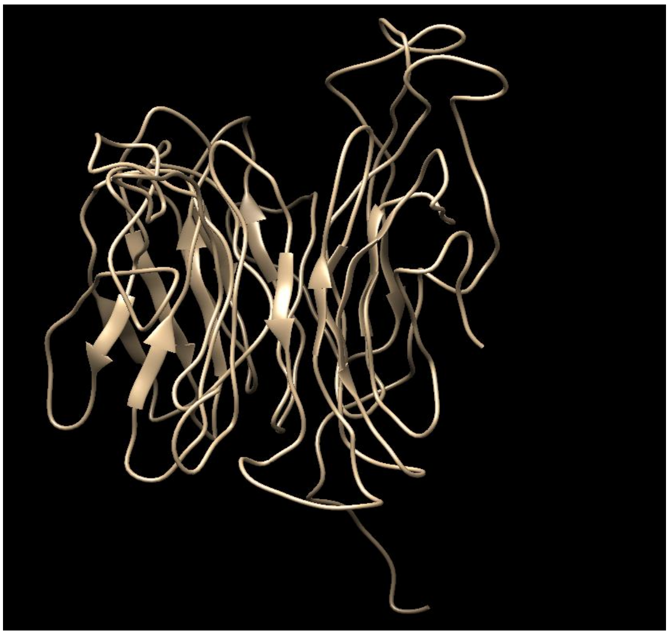
```
{============================ disulphide bonds =
{* Select pairs of cysteine residues that form
{* First 2 entries are the segid and resid of t
{* Second 2 entries are the segid and resid of
{+ table: rows=8 numbered
   cols=5 "use" "segid CYS A" "resid CYS A" "se
{+ choice: true false +}
{===>} ss_use_1=true;
{===>} ss_i_segid_1=""; ss_i_resid_1=46;
{===>} ss_j_segid_1=""; ss_j_resid_1=51;
{+ choice: true false +}
{===>} ss_use_2=true;
{===>} ss_i_segid_2=""; ss_i_resid_2=203;
{===>} ss_j_segid_2=""; ss_j_resid_2=213;
{+ choice: true false +}
{===>} ss_use_3=true;
{===>} ss_i_segid_3=""; ss_i_resid_3=348;
{===>} ss_j_segid_3=""; ss_j_resid_3=371;
{+ choice: true false +}
```

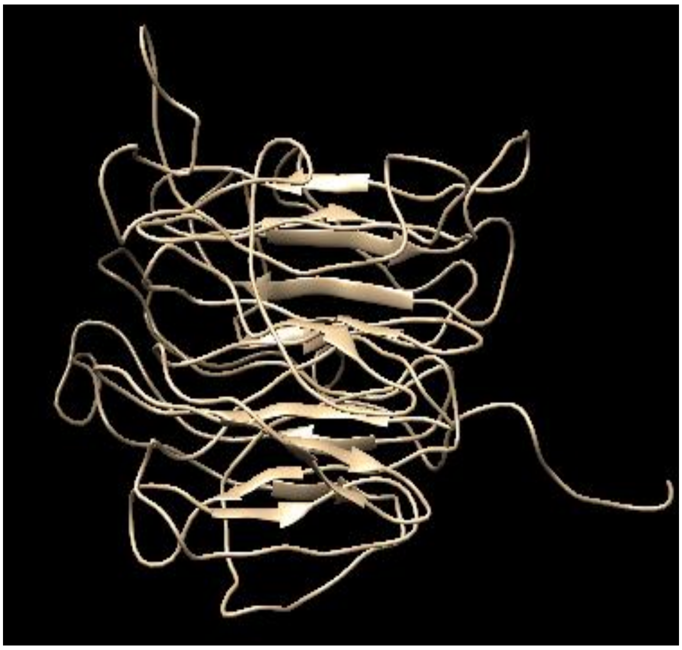
Figure3.8: The gesq.inp file after modified.

Figure 3.8 shows gesq.inp file after modified. In this file, you can see the “true” or “false” flag, and the position of cysteine is the prediction from DIpro2. And then, CNS can use that information to build the molecular topology file.

3.5 Results

Adding disulfide bonds information is essential at the first stage because the CNS solve the addition of bond information to the molecular topology, which describes the covalent topology of the molecule. It means that we can improve the accuracy of reconstruction.





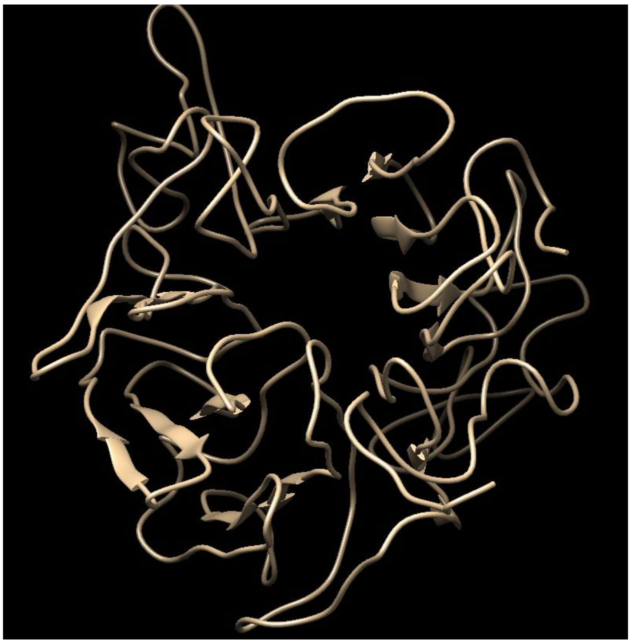
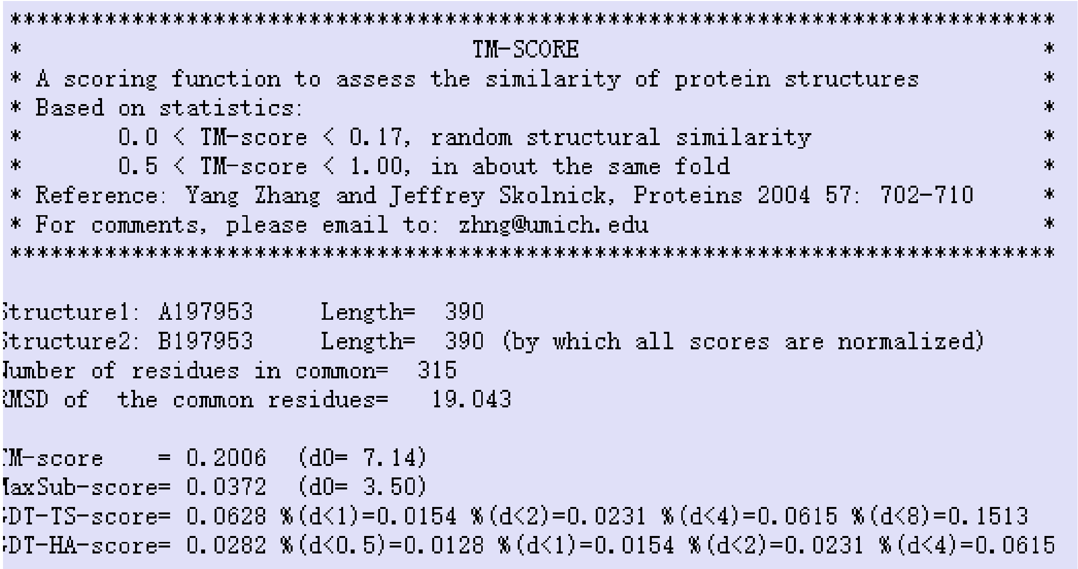


Figure3.9: Reconstructed model of protein 1a4g.

To compare the performance after adding new features, we selected one sequence to reconstruct the protein models and observing the final TM-score. The sequence we decided is 1a4g. The length of 1a4g is 390, contains seven predicted disulfide bonds. We will test this protein sequence separately in two versions and see if the version with new features will improve the test results.

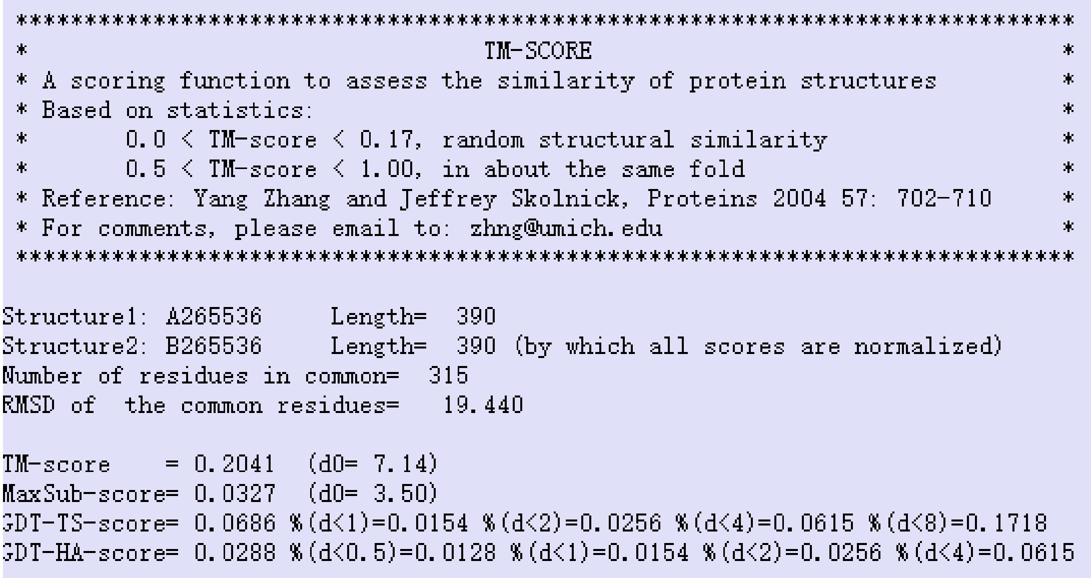


```
*****************************************************************************
*                                 TM-SCORE                                  *
* A scoring function to assess the similarity of protein structures         *
* Based on statistics:                                                      *
*       0.0 < TM-score < 0.17, random structural similarity                 *
*       0.5 < TM-score < 1.00, in about the same fold                       *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710     *
* For comments, please email to: zhng@umich.edu                             *
*****************************************************************************

Structure1: A197953     Length=  390
Structure2: B197953     Length=  390 (by which all scores are normalized)
Number of residues in common=  315
RMSD of  the common residues=   19.043

TM-score    = 0.2006  (d0= 7.14)
MaxSub-score= 0.0372  (d0= 3.50)
GDT-TS-score= 0.0628 %(d<1)=0.0154 %(d<2)=0.0231 %(d<4)=0.0615 %(d<8)=0.1513
GDT-HA-score= 0.0282 %(d<0.5)=0.0128 %(d<1)=0.0154 %(d<2)=0.0231 %(d<4)=0.0615
```

(1) The best TM-score before adding disulfide bond prediction



```
*****************************************************************************
*                                 TM-SCORE                                  *
* A scoring function to assess the similarity of protein structures         *
* Based on statistics:                                                      *
*       0.0 < TM-score < 0.17, random structural similarity                 *
*       0.5 < TM-score < 1.00, in about the same fold                       *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710     *
* For comments, please email to: zhng@umich.edu                             *
*****************************************************************************

Structure1: A265536     Length=  390
Structure2: B265536     Length=  390 (by which all scores are normalized)
Number of residues in common=  315
RMSD of  the common residues=   19.440

TM-score    = 0.2041  (d0= 7.14)
MaxSub-score= 0.0327  (d0= 3.50)
GDT-TS-score= 0.0686 %(d<1)=0.0154 %(d<2)=0.0256 %(d<4)=0.0615 %(d<8)=0.1718
GDT-HA-score= 0.0288 %(d<0.5)=0.0128 %(d<1)=0.0154 %(d<2)=0.0256 %(d<4)=0.0615
```

(2) The best TM-score after adding disulfide bond prediction

Figure3.10: The TM-score comparison of protein 1a4g.

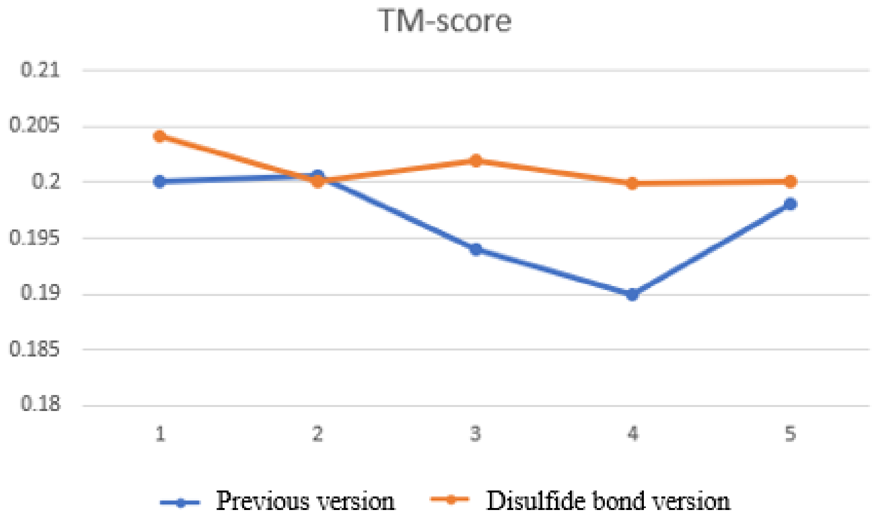


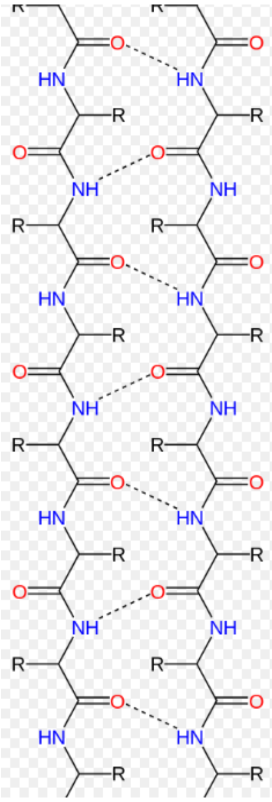
Figure3.11: TM-score line chart.

Figure 3.10 and Figure 3.11 shows that under the same protein sequence, the performance of protein model reconstruction can be improved slightly after adding disulfide bond prediction.

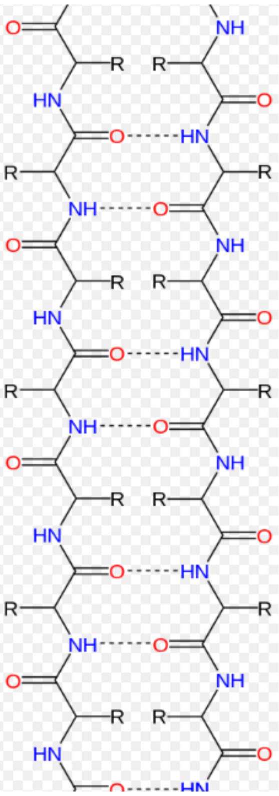
Chapter 4 Beta sheet contacts

4.1 Background.

The β-sheet is a common motif of regular secondary structure in proteins [17]. β-sheets are formed by at least two or three backbone hydrogen bonds, and one β-strand is a stretch of polypeptide chain typically 3 to 10 amino acids long with a backbone in an extended conformation. There are three ways that adjacent β-strands form hydrogen bonds: parallel, antiparallel and mixed arrangements.



Parallel β-sheet hydrogen bond



Antiparallel β-sheet hydrogen bond

Figure4.1: Parallel and Antiparallel β-sheet hydrogen bond.

We choose to use the bbcontacts method to predict the β-strand pairing because it is different from other methods. Most of the existing techniques use true secondary structure as input, but in CONFOLD, we take predicted secondary structure as input, so bbcontacts is the best choice. Before using bbcontacts, we are required to use HHblits, CCMpred and Psipred to generate the input files.

The NOE distance restraints required by CNS solve are specified with the following syntax:

$$assign \quad (atom - selection) \quad (atom - selection) \quad d \quad dmines \quad dplus$$

$$\text{e.g.}, assign \;\; (resid\ 74\ and\ name\ O) \;\; (resid\ 112\ and\ name\ H) \quad 2.8 \quad 0.4 \quad 0.9$$

This kind of selection defines the atoms between which the distance restraint will be applied. In the CNS system, building pseudo atoms can be completed by the "assign" statement. According to the restraining functions, CNS can calculate the R-6 averaged distance or the distance between the geometric centers of selected atoms. We only need to change the format of the prediction of the beta-sheet contact to the "assign" statement, and then CNS can start the NMR structure calculation automatically.

4.2 Using HHblits to generate multiple sequence alignments

HHblits is a part of the HH-suit that can build high-quality multiple sequence alignment, and the input file of HHblits is a single query sequence. It can speed up the slow HMM-HMM comparison process by the fast prefilter because the fast prefilter reduces the tens of millions of HMMs to match against to a few thousands of them.

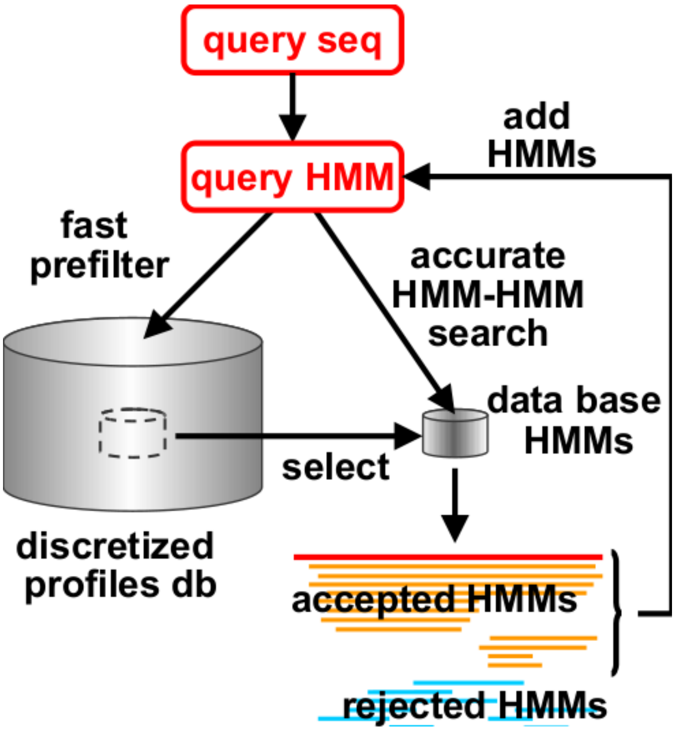


Figure4.2: Process of using HMMs search.

Hidden Markov Model (HMM) is a statistical Markov model that can be represented as the dynamic Bayesian network [19]. The definition is:

$$P\,(Y_n \in A \mid X_1 = x_1, \ldots, X_n = x_n) = P\,(Y_n \in A \mid X_n = x_n)$$

The Markov process itself cannot be observed, only the sequence of labeled clusters, thus this arrangement is called a “hidden Markov process”.

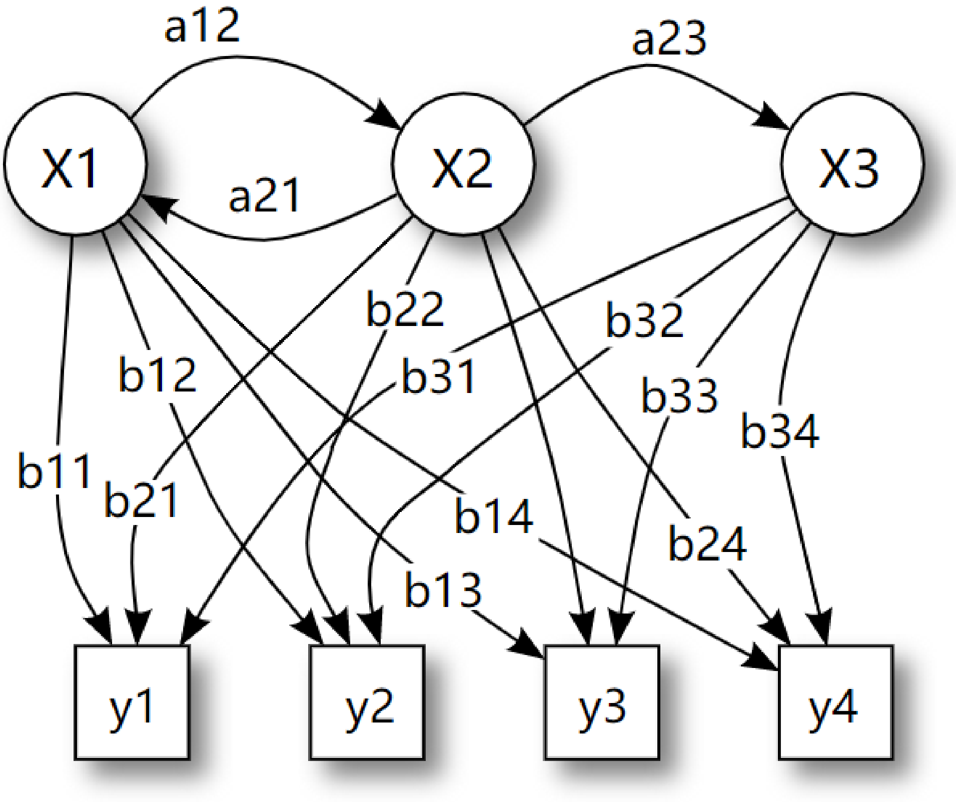
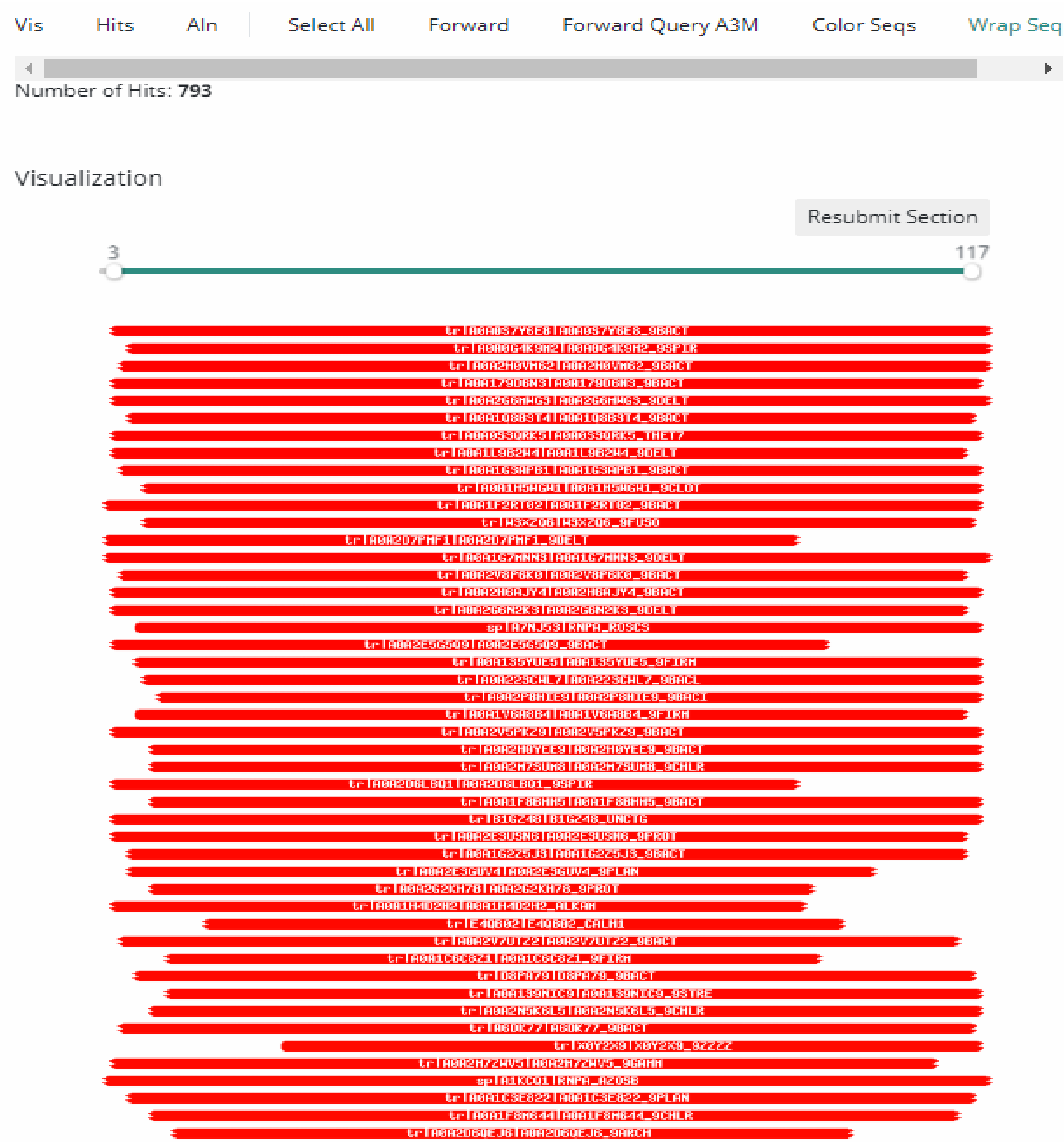


Figure4.3: Probabilistic parameters of a hidden Markov model.

X represents the states, y represents the possible observations, represents the state transition probabilities, and b represents the output probabilities.

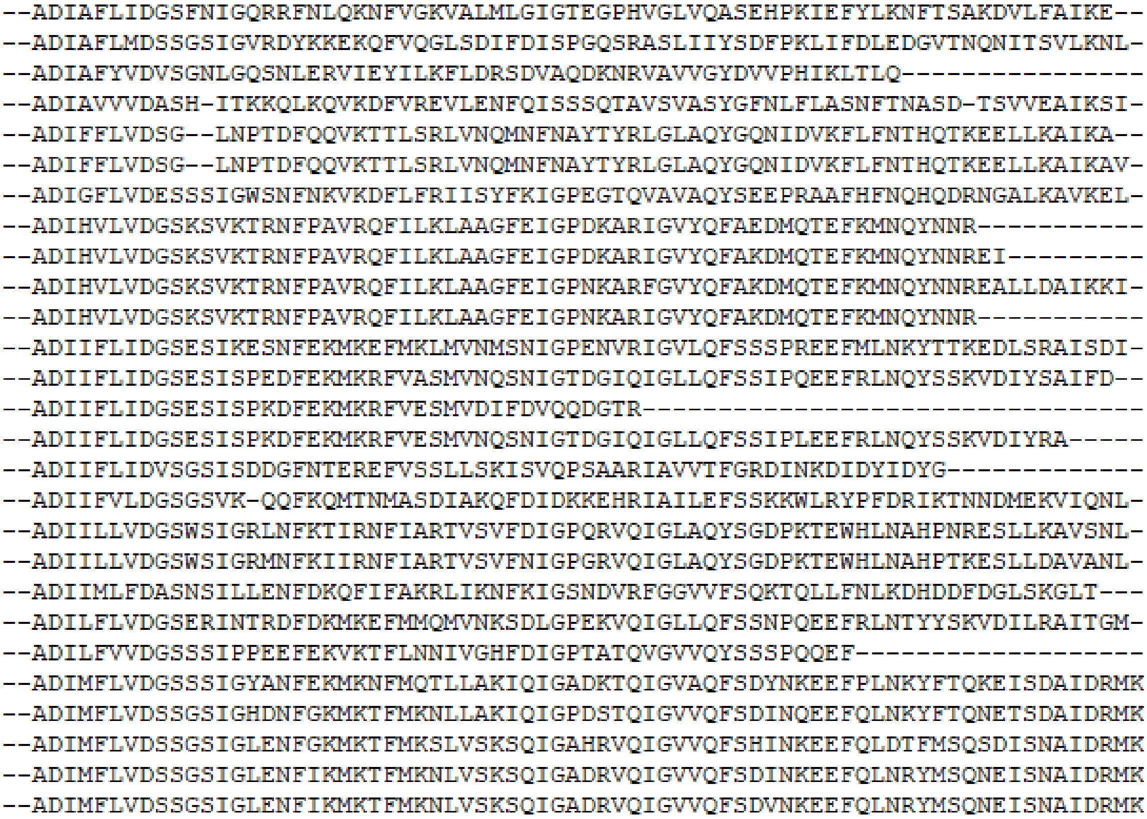
First, we run HHblits against the uniprot20 database, avoiding any filtering in order to retrieves as many homologous sequences as possible. We set the number of target sequences up to 10000, and the minimum probability in the hit list is 20%. Figure 3.2 shows the result of running the HHblits.


Vis
Hits
Aln
Select All
Forward
Forward Query A3M
Color Seqs
Wrap Seq
Visualization
Resubmit Section
3
117
tr|A0A0S7Y6E8|A0A0S7Y6E8_9BACT
tr|A0A0G4K9M2|A0A0G4K9M2_9SPIR
tr|A0A2H0VM62|A0A2H0VM62_9BACT
tr|A0A179D6N3|A0A179D6N3_9BACT
tr|A0A2G6MWG3|A0A2G6MWG3_9DELT
tr|A0A1Q8B3T4|A0A1Q8B3T4_9BACT
tr|A0A0S3QRK5|A0A0S3QRK5_THET7
tr|A0A1L9B2W4|A0A1L9B2W4_9DELT
tr|A0A1G3APB1|A0A1G3APB1_9BACT
tr|A0A1H5WGW1|A0A1H5WGW1_9CLOT
tr|A0A1F2RT02|A0A1F2RT02_9BACT
tr|H5X2U6|H5X2U6_9FUSO
tr|A0A2D7PMF1|A0A2D7PMF1_9DELT
tr|A0A1G7MNN3|A0A1G7MNN3_9DELT
tr|A0A2V8P6K0|A0A2V8P6K0_9BACT
tr|A0A2H6AJY4|A0A2H6AJY4_9BACT
tr|A0A2G6N2K3|A0A2G6N2K3_9DELT
sp|A7NJ53|RNPA_ROSCS
tr|A0A2E5G509|A0A2E5G509_9BACT
tr|A0A135YUE5|A0A135YUE5_9FIRM
tr|A0A223CHL7|A0A223CHL7_9BACL
tr|A0A2P8H1E9|A0A2P8H1E9_9BACI
tr|A0A1V6A8B4|A0A1V6A8B4_9FIRM
tr|A0A2V5PKZ9|A0A2V5PKZ9_9BACT
tr|A0A2H0YEE9|A0A2H0YEE9_9BACT
tr|A0A2H7SUH8|A0A2H7SUH8_9CHLR
tr|A0A2D6LBQ1|A0A2D6LBQ1_9SPIR
tr|A0A1F8BHH5|A0A1F8BHH5_9BACT
tr|B1G248|B1G248_UNCTG
tr|A0A2E3U5N6|A0A2E3U5N6_9PROT
tr|A0A1G2Z5J3|A0A1G2Z5J3_9BACT
tr|A0A2E3GUV4|A0A2E3GUV4_9PLAN
tr|A0A2G2KH78|A0A2G2KH78_9PROT
tr|A0A1H4D2H2|A0A1H4D2H2_ALKAM
tr|E4Q6Q2|E4Q6Q2_CALH1
tr|A0A2V7UT22|A0A2V7UT22_9BACT
tr|A0A1C6C8Z1|A0A1C6C8Z1_9FIRM
tr|D8PA79|D8PA79_9BACT
tr|A0A139N1C9|A0A139N1C9_9STRE
tr|A0A2N5K6L5|A0A2N5K6L5_9CHLR
tr|A6DK77|A6DK77_9BACT
tr|X0Y2X9|X0Y2X9_9ZZZZ
tr|A0A2H7ZHV5|A0A2H7ZHV5_9GAMM
sp|A1KC01|RNPA_AZOSB
tr|A0A1C3E822|A0A1C3E822_9PLAN
tr|A0A1F8M644|A0A1F8M644_9CHLR
tr|A0A2D6QEJ8|A0A2D6QEJ8_9ARCH

Number of Hits: **793**

Figure4.4: Visualization of the multiple sequence alignment.

After running the HHblits, we get the query template multiple sequence alignments. But this a3m file cannot be used directly; the length of every alignment is different and contains much useless information. So, the next step is to use HHfilter to complete the extraction of a representative of sequences from an alignment. The length of each alignment should be equal to the length of the sequence.



```
--ADIAFLIDGSFNIGQRRFNLQKNFVGKVALMLGIGTEGPHVGLVQASEHPKIEFYLKNFTSAKDVLFAIKE--
--ADIAFLMDSSGSIGVRDYKKEKQFVQGLSDIFDISPGQSRASLIIYSDFPKLIFDLEDGVTNQNITSVLKNL-
--ADIAFYVDVSGNLGQSNLERVIEYILKFLDRSDVAQDKNRVAVVGYDVVPHIKLTLQ----------------
--ADIAVVVDASH-ITKKQLKQVKDFVREVLENFQISSSQTAVSVASYGFNLFLASNFTNASD-TSVVEAIKSI-
--ADIFFLVDSG--LNPTDFQQVKTTLSRLVNQMNFNAYTYRLGLAQYGQNIDVKFLFNTHQTKEELLKAIKA--
--ADIFFLVDSG--LNPTDFQQVKTTLSRLVNQMNFNAYTYRLGLAQYGQNIDVKFLFNTHQTKEELLKAIKAV-
--ADIGFLVDESSSIGWSNFNKVKDFLFRIISYFKIGPEGTQVAVAQYSEEPRAAFHFNQHQDRNGALKAVKEL-
--ADIHVLVDGSKSVKTRNFPAVRQFILKLAAGFEIGPDKARIGVYQFAEDMQTEFKMNQYNNR-----------
--ADIHVLVDGSKSVKTRNFPAVRQFILKLAAGFEIGPDKARIGVYQFAKDMQTEFKMNQYNNREI---------
--ADIHVLVDGSKSVKTRNFPAVRQFILKLAAGFEIGPNKARFGVYQFAKDMQTEFKMNQYNNREALLDAIKKI-
--ADIHVLVDGSKSVKTRNFPAVRQFILKLAAGFEIGPNKARIGVYQFAKDMQTEFKMNQYNNR-----------
--ADIIFLIDGSESIKESNFEKMKEFMKLMVNMSNIGPENVRIGVLQFSSSPREEFMLNKYTTKEDLSRAISDI-
--ADIIFLIDGSESISPEDFEKMKRFVASMVNQSNIGTDGIQIGLLQFSSIPQEEFRLNQYSSKVDIYSAIFD--
--ADIIFLIDGSESISPKDFEKMKRFVESMVDIFDVQQDGTR---------------------------------
--ADIIFLIDGSESISPKDFEKMKRFVESMVNQSNIGTDGIQIGLLQFSSIPLEEFRLNQYSSKVDIYRA-----
--ADIIFLIDVSGSISDDGFNTEREFVSSLLSKISVQPSAARIAVVTFGRDINKDIDYIDYG-------------
--ADIIFVLDGSGSVK-QQFKQMTNMASDIAKQFDIDKKEHRIAILEFSSKKWLRYPFDRIKTNNDMEKVIQNL-
--ADIILLVDGSWSIGRLNFKTIRNFIARTVSVFDIGPQRVQIGLAQYSGDPKTEWHLNAHPNRESLLKAVSNL-
--ADIILLVDGSWSIGRMNFKIIRNFIARTVSVFNIGPGRVQIGLAQYSGDPKTEWHLNAHPTKESLLDAVANL-
--ADIIMLFDASNSILLENFDKQFIFAKRLIKNFKIGSNDVRFGGVVFSQKTQLLFNLKDHDDFDGLSKGLT---
--ADILFLVDGSERINTRDFDKMKEFMMQMVNKSDLGPEKVQIGLLQFSSNPQEEFRLNTYYSKVDILRAITGM-
--ADILFVVDGSSSIPPEEFEKVKTFLNNIVGHFDIGPTATQVGVVQYSSSPQQEF-------------------
--ADIMFLVDGSSSIGYANFEKMKNFMQTLLAKIQIGADKTQIGVAQFSDYNKEEFPLNKYFTQKEISDAIDRMK
--ADIMFLVDSSGSIGHDNFGKMKTFMKNLLAKIQIGPDSTQIGVVQFSDINQEEFQLNKYFTQNETSDAIDRMK
--ADIMFLVDSSGSIGLENFGKMKTFMKSLVSKSQIGAHRVQIGVVQFSHINKEEFQLDTFMSQSDISNAIDRMK
--ADIMFLVDSSGSIGLENFIKMKTFMKNLVSKSQIGADRVQIGVVQFSDINKEEFQLNRYMSQNEISNAIDRMK
--ADIMFLVDSSGSIGLENFIKMKTFMKNLVSKSQIGADRVQIGVVQFSDVNKEEFQLNRYMSQNEISNAIDRMK
```

Figure4.5: Format of the alignments.

After reformatting the alignments, using those multiple sequence alignments as input to get the prediction of direct couplings.

4.3 Using CCMpred to predict direct couplings

CCMpred is free and open-source software that can predict protein residue-residue contact [20]. Compared with other published methods, it can predict contacts faster and with the same precision.

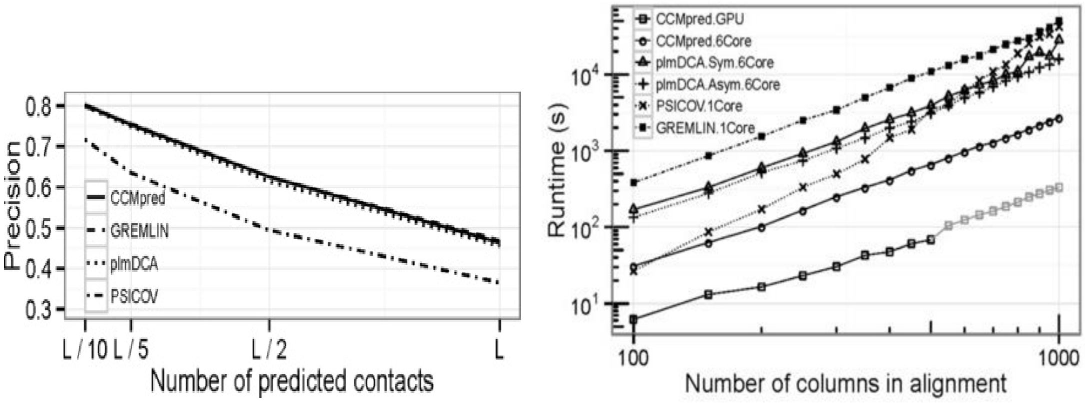
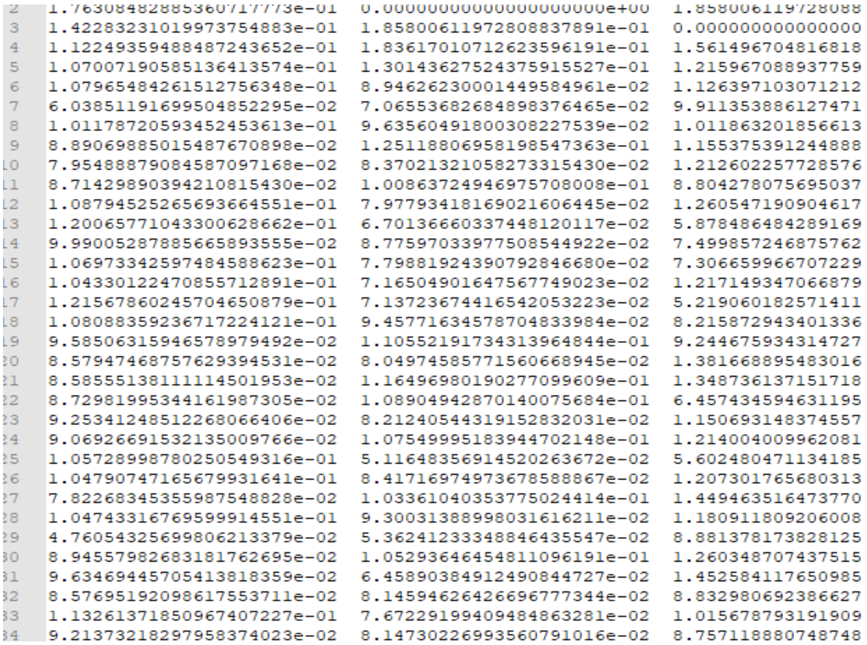


Figure4.6: CCMpred runtime and accuracy compared with other methods.

Protein structure can maintain stability is crucial under evolutionary pressure, which gives rise to correlated mutations between contacting residue pairs. These correlated mutations can be used to predict residue-residue contacts. The output file is a direct couplings matrix that contains the contact information.



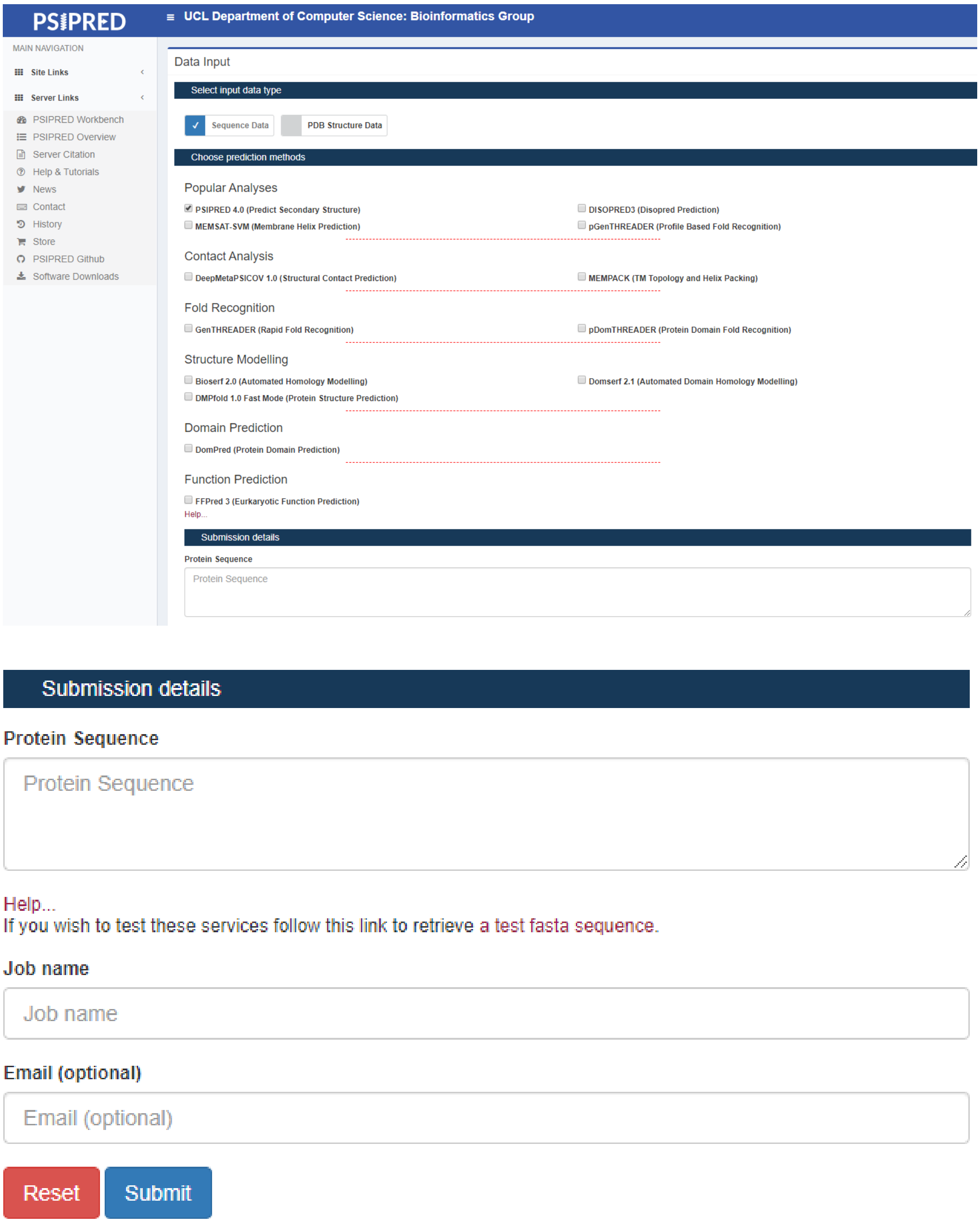
```
1.76308482885360717773e-01   0.00000000000000000000e+00   1.858006119728088
1.42283231019973754883e-01   1.85800611972808837891e-01   0.000000000000000
1.12249359488487243652e-01   1.83617010712623596191e-01   1.561496704816818
1.07007190585136413574e-01   1.30143627524375915527e-01   1.215967088937759
1.07965484261512756348e-01   8.94626230001449584961e-02   1.126397103071212
6.03851191699504852295e-02   7.06553682684898376465e-02   9.911353886127471
1.01178720593452453613e-01   9.63560491800308227539e-02   1.011863201856613
8.89069885015487670898e-02   1.25118806958198547363e-01   1.155375391244888
7.95488879084587097168e-02   8.37021321058273315430e-02   1.212602257728576
8.71429890394210815430e-02   1.00863724946975708008e-01   8.804278075695037
1.08794525265693664551e-01   7.97793418169021606445e-02   1.260547190904617
1.20065771043300628662e-01   6.70136660337448120117e-02   5.878486484289169
9.99005287885665893555e-02   8.77597033977508544922e-02   7.499857246875762
1.06973342597484588623e-01   7.79881924390792846680e-02   7.306659966707229
1.04330122470855712891e-01   7.16504901647567749023e-02   1.217149347066879
1.21567860245704650879e-01   7.13723674416542053223e-02   5.219060182571411
1.08088359236717224121e-01   9.45771634578704833984e-02   8.215872943401336
9.58506315946578979492e-02   1.10552191734313964844e-01   9.244675934314727
8.57947468757629394531e-02   8.04974585771560668945e-02   1.381668895483016
8.58555138111114501953e-02   1.16496980190277099609e-01   1.348736137151718
8.72981995344161987305e-02   1.08904942870140075684e-01   6.457434594631195
9.25341248512268066406e-02   8.21240544319152832031e-02   1.150693148374557
9.06926691532135009766e-02   1.07549995183944702148e-01   1.214004009962081
1.05728998780250549316e-01   5.11648356914520263672e-02   5.602480471134185
1.04790747165679931641e-01   8.41716974973678588867e-02   1.207301765680313
7.82268345355987548828e-02   1.03361040353775024414e-01   1.449463516473770
1.04743316769599914551e-01   9.30031388998031616211e-02   1.180911809206008
4.76054325699806213379e-02   5.36241233348846435547e-02   8.881378173828125
8.94557982683181762695e-02   1.05293646454811096191e-01   1.260348707437515
9.63469445705413818359e-02   6.45890384912490844727e-02   1.452584117650985
8.57695192098617553711e-02   8.14594626426696777344e-02   8.832980692386627
1.13261371850967407227e-01   7.67229199409484863281e-02   1.015678793191909
9.21373218297958374023e-02   8.14730226993560791016e-02   8.757118880748748
```

Figure4.7: Format of the CCMpred result.

The columns number of direct couplings matrix is equal to the length of the sequence, then bbcontacts can predict the β-strands pairing by detecting patterns in the matrix of predicted couplings corresponding to interactions between secondary structure elements.

4.4 Using Psipred to predict secondary structure.

Psipred is a method used to predict a protein’s secondary structure from the primary sequence. There are three stages in the prediction algorithm: generating a sequence profile, predicting the initial secondary structure and filtering the predicted structure. The web service is convenient to use.



Data Input

Select input data type

☑ Sequence Data □ PDB Structure Data

Choose prediction methods

Popular Analyses

☑ PSIPRED 4.0 (Predict Secondary Structure)
□ DISOPRED3 (Disopred Prediction)
□ MEMSAT-SVM (Membrane Helix Prediction)
□ pGenTHREADER (Profile Based Fold Recognition)

Contact Analysis

□ DeepMetaPSICOV 1.0 (Structural Contact Prediction)
□ MEMPACK (TM Topology and Helix Packing)

Fold Recognition

□ GenTHREADER (Rapid Fold Recognition)
□ pDomTHREADER (Protein Domain Fold Recognition)

Structure Modelling

□ Bioserf 2.0 (Automated Homology Modelling)
□ Domserf 2.1 (Automated Domain Homology Modelling)
□ DMPfold 1.0 Fast Mode (Protein Structure Prediction)

Domain Prediction

□ DomPred (Protein Domain Prediction)

Function Prediction

□ FFPred 3 (Eurkaryotic Function Prediction)

Help...

Submission details

Protein Sequence

Protein Sequence

Submission details

Protein Sequence

Protein Sequence

Help...
If you wish to test these services follow this link to retrieve a test fasta sequence.

Job name

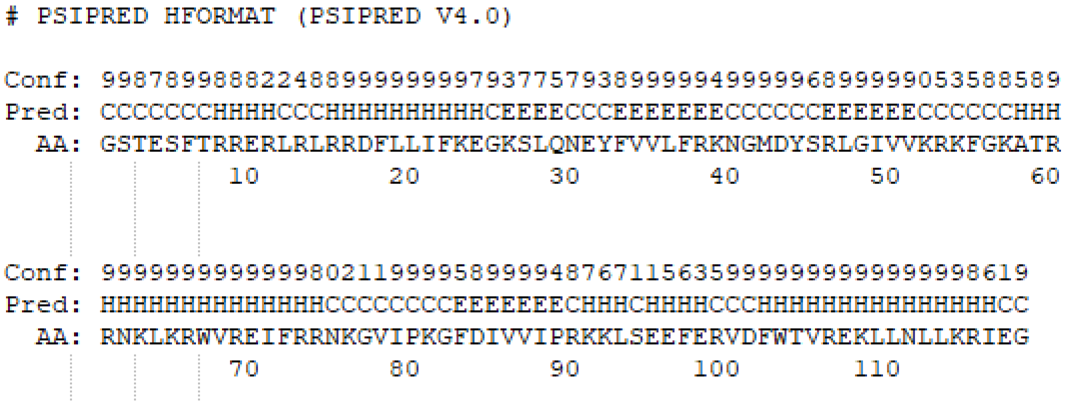
Job name

Email (optional)

Email (optional)

Figure4.8: Psipred web service.

It is very convenient to use the Psipred web server. We can just submit our protein sequence and the email address when the work finished, and we will receive an email that contains the information of the prediction.



```
# PSIPRED HFORMAT (PSIPRED V4.0)

Conf: 998789988822488999999997937757938999994999996899999053588589
Pred: CCCCCCCHHHHCCCHHHHHHHHHHCEEEECCCEEEEEEEECCCCCCEEEEEEECCCCCCHHH
  AA: GSTESFTRRERLRLRRDFLLIFKEGKSLQNEYFVVLFRKNGMDYSRLGIVVKRKFGKATR
              10        20        30        40        50        60

Conf: 99999999999998021199995899994876711563599999999999999998619
Pred: HHHHHHHHHHHHHHCCCCCCCCEEEEEEECHHHCHHHHCCCHHHHHHHHHHHHHHCC
  AA: RNKLKRWVREIFRRNKGVIPKGFDIVVIPRKKLSEEFERVDFWTVREKLLNLLKRIEG
              70        80        90       100       110
```

Figure4.9: Prediction of the secondary structure.

In the secondary structure files, “E” represents an extended strand, participates in the beta ladder. So, we need to identify the relationship between the “E” parts, and if they are contacted, we can regard it as β-sheet contact.

4.5 Using BBcontacts to predict β-sheet contacts.

The Hidden Markov Model (HMM) architecture is used for parallel and antiparallel β-sheet contacts. To run bbcontacts for a given protein, we need a matrix of predicted couplings and a three-state secondary structure prediction. Because when CCMpred performs the average product correction (APC) step [15], the minimum coupling value gets subtracted from all coupling values, we should make sure to use a smoothing range when running bbcontacts.

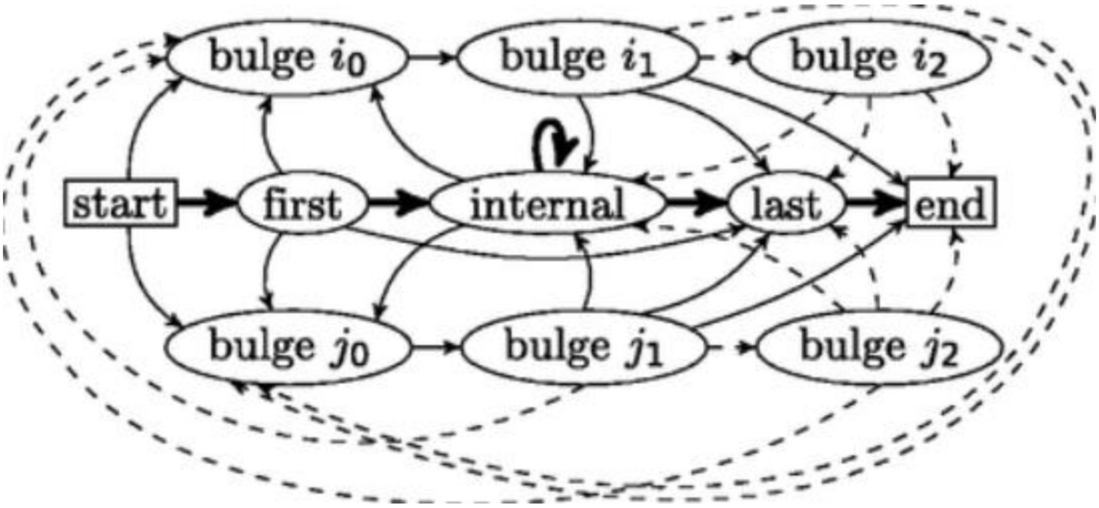
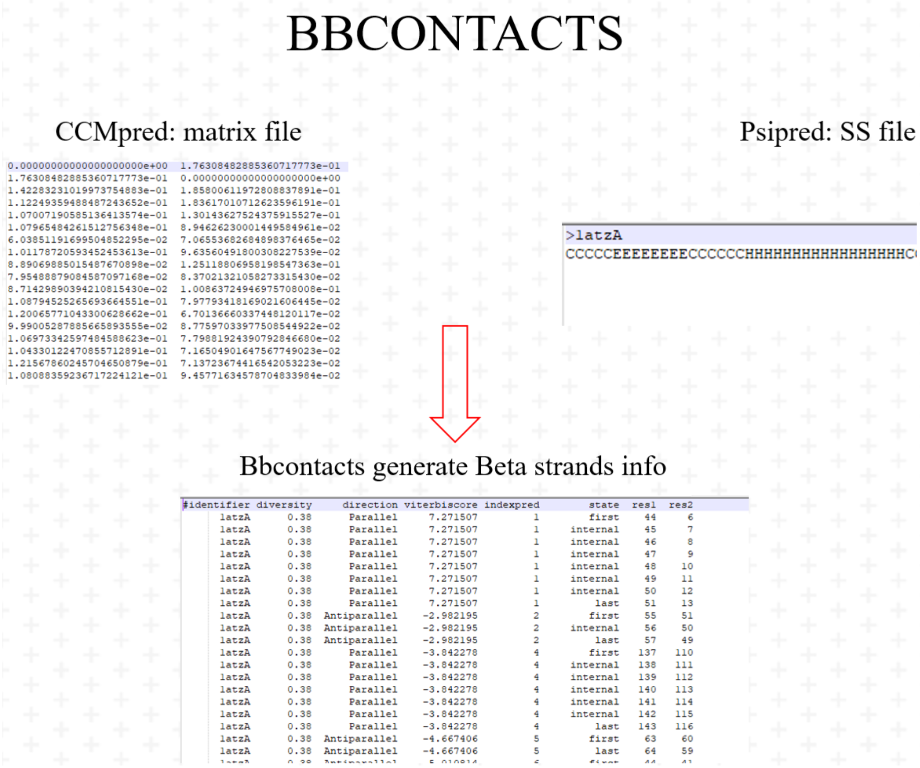


Figure4.10: HMM architecture used in bbcontacts.



BBCONTACTS

CCMpred: matrix file

```
0.000000000000000000000e+00  1.763084828853607177773e-01
1.763084828853607177773e-01  0.000000000000000000000e+00
1.422832310199737548883e-01  1.858006119728088378910e-01
1.122493594884872436520e-01  1.836170107126235961910e-01
1.070071905851364135740e-01  1.301436275243759155270e-01
1.079654842615127563480e-01  8.946262300014495849610e-02
6.038511916995048522950e-02  7.065536826848983764650e-02
1.011787205934524536130e-01  9.635604918003082275390e-02
8.890698850154876708980e-02  1.251188069581985473630e-01
7.954888790845870971680e-02  8.370213210582733154300e-02
8.714298903942108154300e-02  1.008637249469757080080e-01
1.087945252656936645510e-01  7.977934181690216064450e-02
1.200657710433006286620e-01  6.701366603374481201170e-02
9.990052878856658935550e-02  8.775970339775085449220e-02
1.069733425974845886230e-01  7.798819243907928466800e-02
1.043301224708557128910e-01  7.165049016475677490230e-02
1.215678602457046508790e-01  7.137236744165420532230e-02
1.080883592367172241210e-01  9.457716345787048339840e-02
```

Psipred: SS file

```
>latzA
CCCCCEEEEEEEECCCCCCHHHHHHHHHHHHHHHHHC
```

Bbcontacts generate Beta strands info

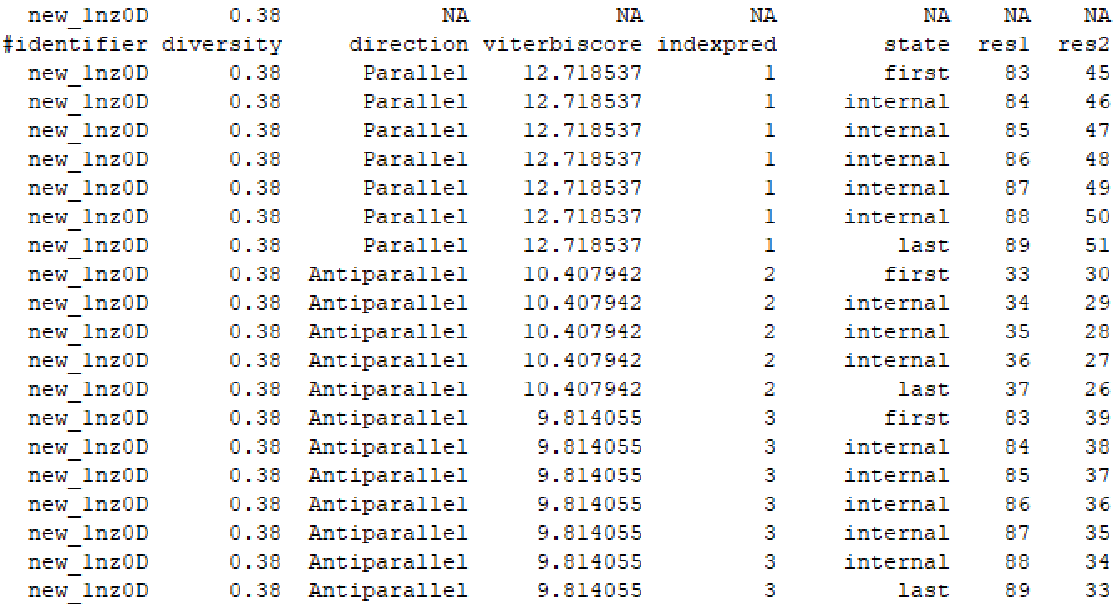
```
#identifier diversity    direction viterbiscore indexpred       state  res1  res2
      latzA       0.38     Parallel     7.271507         1       first    44     6
      latzA       0.38     Parallel     7.271507         1    internal    45     7
      latzA       0.38     Parallel     7.271507         1    internal    46     8
      latzA       0.38     Parallel     7.271507         1    internal    47     9
      latzA       0.38     Parallel     7.271507         1    internal    48    10
      latzA       0.38     Parallel     7.271507         1    internal    49    11
      latzA       0.38     Parallel     7.271507         1    internal    50    12
      latzA       0.38     Parallel     7.271507         1        last    51    13
      latzA       0.38 Antiparallel    -2.982195         2       first    55    51
      latzA       0.38 Antiparallel    -2.982195         2    internal    56    50
      latzA       0.38 Antiparallel    -2.982195         2        last    57    49
      latzA       0.38     Parallel    -3.842278         4       first   137   110
      latzA       0.38     Parallel    -3.842278         4    internal   138   111
      latzA       0.38     Parallel    -3.842278         4    internal   139   112
      latzA       0.38     Parallel    -3.842278         4    internal   140   113
      latzA       0.38     Parallel    -3.842278         4    internal   141   114
      latzA       0.38     Parallel    -3.842278         4    internal   142   115
      latzA       0.38     Parallel    -3.842278         4        last   143   116
      latzA       0.38 Antiparallel    -4.667406         5       first    63    60
      latzA       0.38 Antiparallel    -4.667406         5        last    64    59
      latzA       0.38 Antiparallel    -5.010814         6       first    44    41
```

Figure4.11: BBCONTACTS processing requirements.

The output file contains the β-sheet contact predictions, and there are three key messages that we will use in the next step:

- Direction: Parallel or Antiparallel.
- State: First, Internal or Last.
- Residue position: residue_1, residue_2.

The direction of beta-strand determines the connection order, and three states tell us the begin and end position. The residue position is the information required by CNS solve.



```
  new_1nz0D      0.38             NA          NA         NA            NA    NA    NA
#identifier diversity      direction viterbiscore indexpred         state  res1  res2
  new_1nz0D      0.38       Parallel    12.718537         1         first    83    45
  new_1nz0D      0.38       Parallel    12.718537         1      internal    84    46
  new_1nz0D      0.38       Parallel    12.718537         1      internal    85    47
  new_1nz0D      0.38       Parallel    12.718537         1      internal    86    48
  new_1nz0D      0.38       Parallel    12.718537         1      internal    87    49
  new_1nz0D      0.38       Parallel    12.718537         1      internal    88    50
  new_1nz0D      0.38       Parallel    12.718537         1          last    89    51
  new_1nz0D      0.38   Antiparallel    10.407942         2         first    33    30
  new_1nz0D      0.38   Antiparallel    10.407942         2      internal    34    29
  new_1nz0D      0.38   Antiparallel    10.407942         2      internal    35    28
  new_1nz0D      0.38   Antiparallel    10.407942         2      internal    36    27
  new_1nz0D      0.38   Antiparallel    10.407942         2          last    37    26
  new_1nz0D      0.38   Antiparallel     9.814055         3         first    83    39
  new_1nz0D      0.38   Antiparallel     9.814055         3      internal    84    38
  new_1nz0D      0.38   Antiparallel     9.814055         3      internal    85    37
  new_1nz0D      0.38   Antiparallel     9.814055         3      internal    86    36
  new_1nz0D      0.38   Antiparallel     9.814055         3      internal    87    35
  new_1nz0D      0.38   Antiparallel     9.814055         3      internal    88    34
  new_1nz0D      0.38   Antiparallel     9.814055         3          last    89    33
```

Figure4.12: Format of the bbcontacts output file.

The beta-sheets have three directions, parallel, antiparallel, and mix beta-sheets. In parallel beta-sheets, all the beta strands run in the same direction. And in antiparallel beta-sheets, the beta strands run in the opposite directions. The antiparallel beta-sheets are more stable than the parallel beta-sheets because parallel sheets are less twisted than antiparallel, and the antiparallel sheets can bear more enormous distortions.

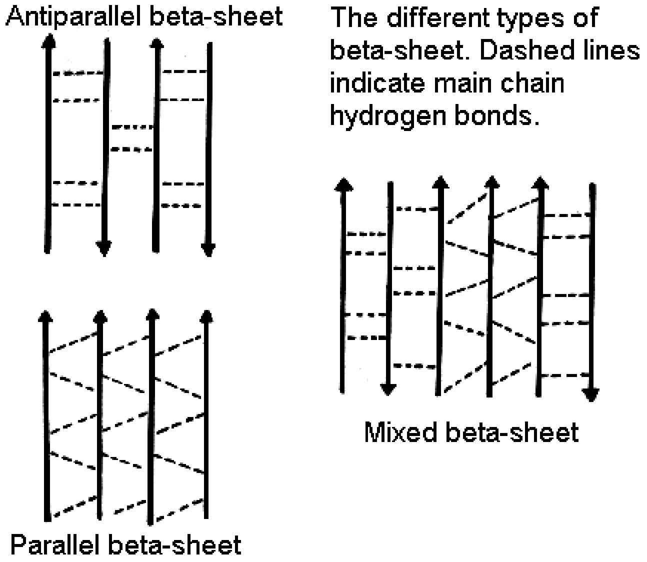

Antiparallel beta-sheet
The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.
Mixed beta-sheet
Parallel beta-sheet

Figure4.13: The diagrams of parallel and antiparallel beta-sheets.

The parallel and antiparallel beta-sheets use the same HMM architecture, but the parameters are different.

4.6 Adding β-sheet contacts information.

From the previous steps, we got the required information. Now we are trying to integrate the data into “dgas.inp” file to calculate the structure. In CONFOLD Version 2, the beta contact information can be detected from the stage one model. But the drawback of this way is that we cannot avoid the mistakes in stage one. In order to solve this problem, we choose to add the β-sheet prediction information in the first stage, so that the models in the early stage can use the position of β-sheet residues and improve the TM-score of the reconstructed protein models.

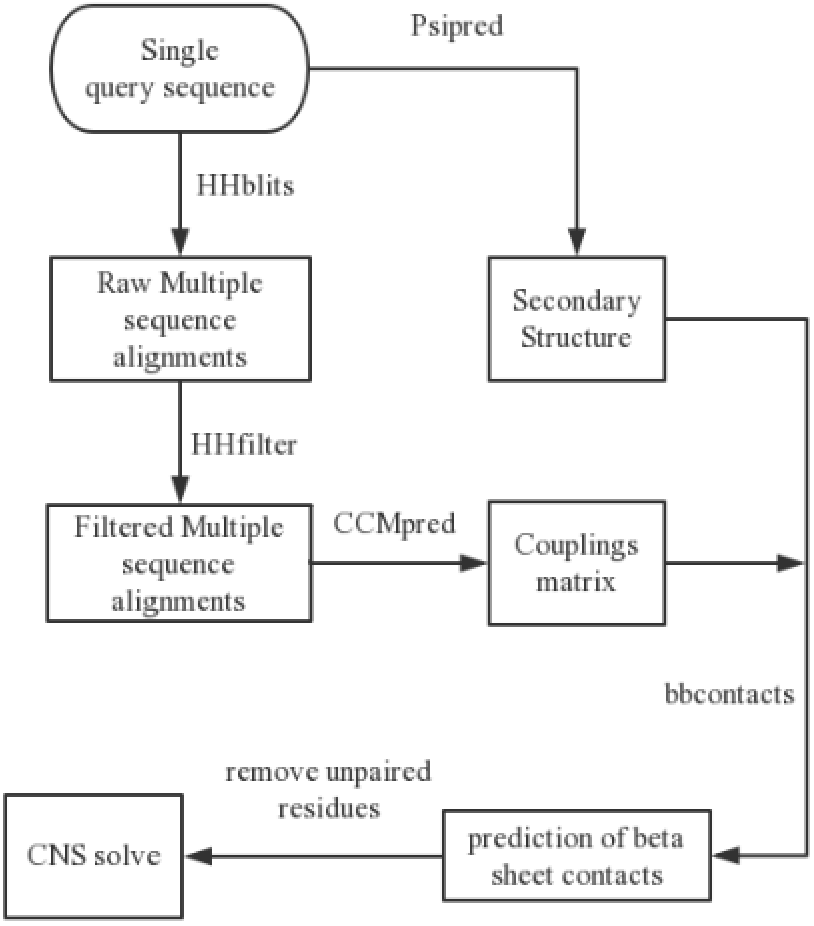
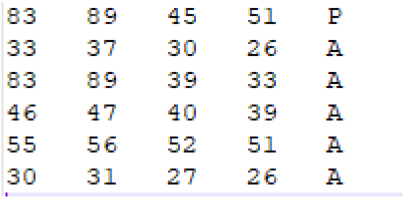


Figure4.14: Process of adding beta-sheet contacts information.

First, reading the required messages from the prediction file. We need to recognize the state of the residue if the residue state is “first” we will start reading the next residues into this strand until the state is “last”. Then the direction of the strand will be attached; the symbol of parallel is “P” and the symbol of antiparallel is “A”.

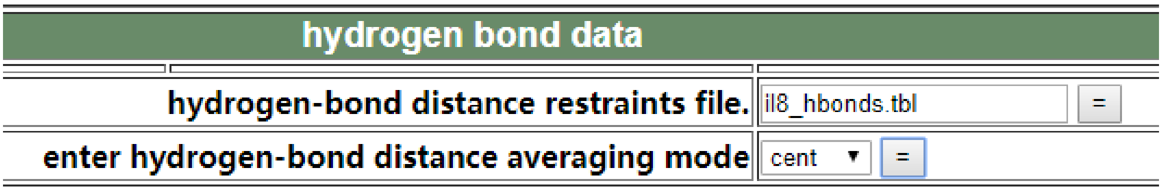


```
83    89    45    51    P
33    37    30    26    A
83    89    39    33    A
46    47    40    39    A
55    56    52    51    A
30    31    27    26    A
```

Figure4.15: The prediction information after processing.

Sometimes the prediction of bbcontacts cannot match the secondary structure file. For example, in some cases, the prediction of bbcontacts shows that the residues 35-38 and residues 79-76 are beta contacts, but in the secondary structure file, the state of 79-76 is not “E”. So we need to remove those unpaired strands.

After removing the unpaired strands, we start writing the beta contacts information into “hbond.tbl” file. This file will be called by the “dgsa.inp” which is used to the distance geometry simulated annealing.



hydrogen bond data

hydrogen-bond distance restraints file. il8_hbonds.tbl

enter hydrogen-bond distance averaging mode cent

Figure4.16: The hydrogen bond distance restraints file in CNS.

Figure 4.16 shows the interface of CNS solve calling the hydrogen bond distance restraints file. It contains all the hydrogen bond information, and we need to select the hydrogen bond distance averaging mode. There are four possible modes: R-6, R-3, sum, cent.

- R-6: The distance between the selected sets of atoms is averaged according to:

  $$\mathrm{R} = [\mathrm{distance}]^{-\frac{1}{6}} \;.$$

- R-3: The distance between the selected sets of atoms is averaged according to:

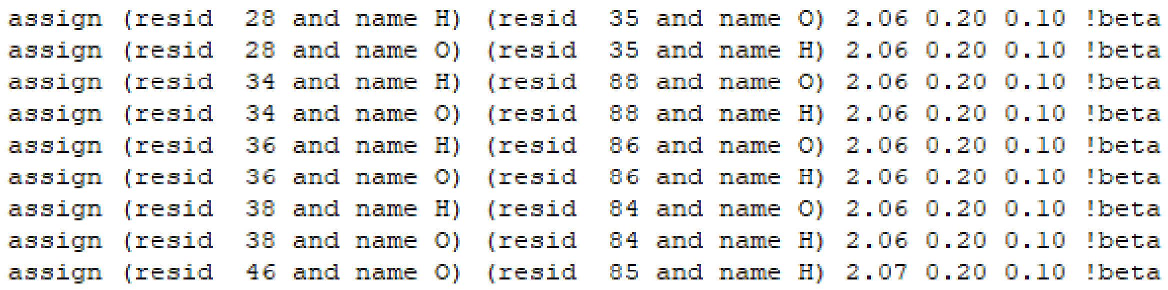
  $$\mathrm{R} = [\mathrm{distance}]^{-\frac{1}{3}} \;.$$

- Sum: The distance between the selected sets of atoms is computed by adding up single contributions: ("nmono" is specified by the monomer statement.)

  $$\mathrm{R} = \mathrm{sum(i,j)}[R_{ij}^{-\frac{6}{nmono}}]^{-\frac{1}{6}}.$$

- Cent: The distance between the selected sets of atoms is set to the difference between the geometric centers of the atoms:

  $$\mathrm{R} = (R_{center1} \;-\; R_{center2}).$$



```
assign (resid  28 and name H) (resid  35 and name O) 2.06 0.20 0.10 !beta
assign (resid  28 and name O) (resid  35 and name H) 2.06 0.20 0.10 !beta
assign (resid  34 and name H) (resid  88 and name O) 2.06 0.20 0.10 !beta
assign (resid  34 and name O) (resid  88 and name H) 2.06 0.20 0.10 !beta
assign (resid  36 and name H) (resid  86 and name O) 2.06 0.20 0.10 !beta
assign (resid  36 and name O) (resid  86 and name H) 2.06 0.20 0.10 !beta
assign (resid  38 and name H) (resid  84 and name O) 2.06 0.20 0.10 !beta
assign (resid  38 and name O) (resid  84 and name H) 2.06 0.20 0.10 !beta
assign (resid  46 and name O) (resid  85 and name H) 2.07 0.20 0.10 !beta
```

Figure4.17: Beta contacts information in the hbond.tbl.

Figure 4.17 shows the format of hydrogen bonds information in “assign” statement. The real number 2.06 means the distance, the 0.20 and 0.10 means the extents either side of this distance.

4.7 Result

In the CONFOLD new version, the beta-sheet contacts information can be accepted by the CNS solve in stage one. And the TM-score of the protein models in stage one improved significantly. But in CONFOLD Version 2, the beta contact information can be detected from the stage one model, so the resulting model’s TM-score is quite similar to the previous version.

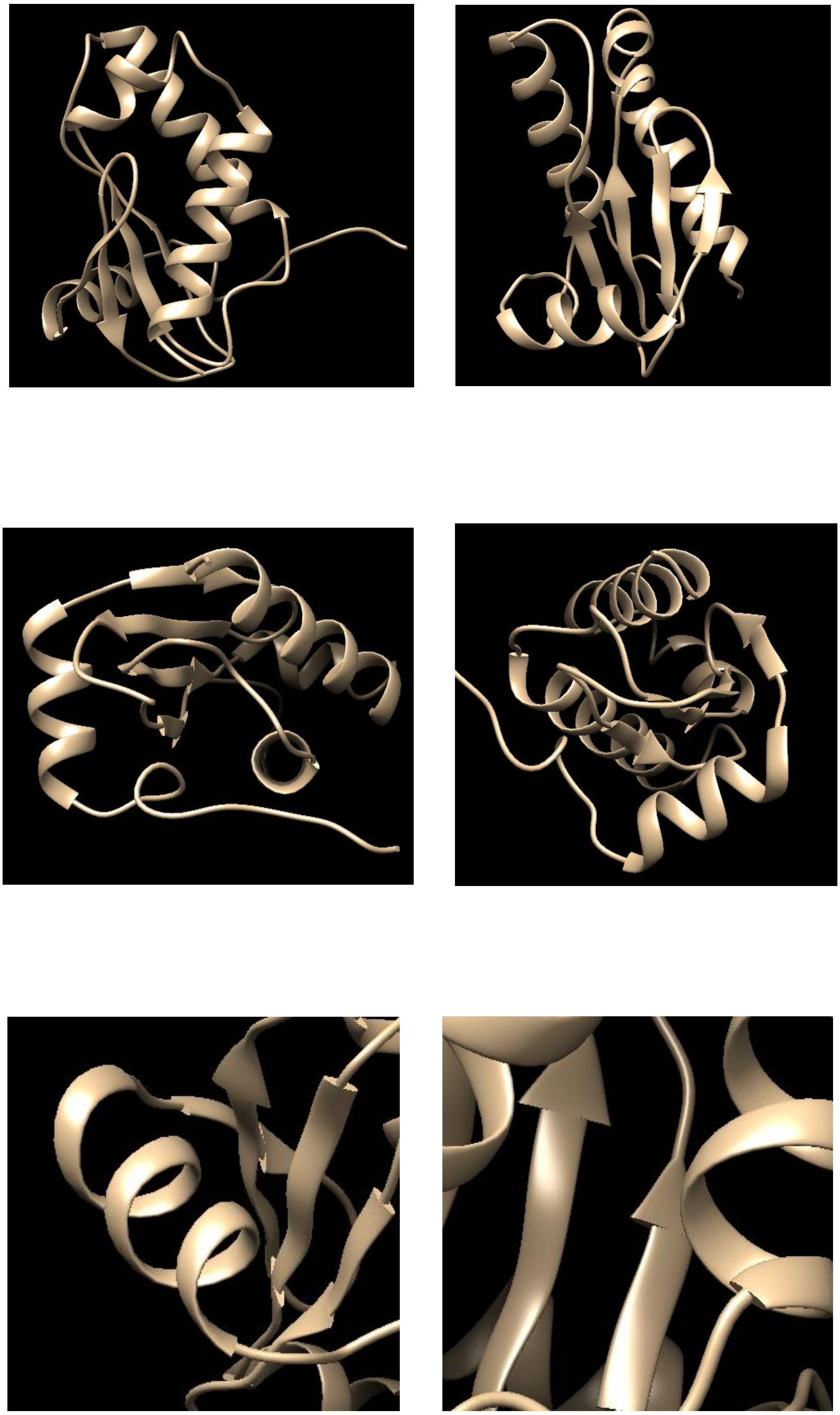
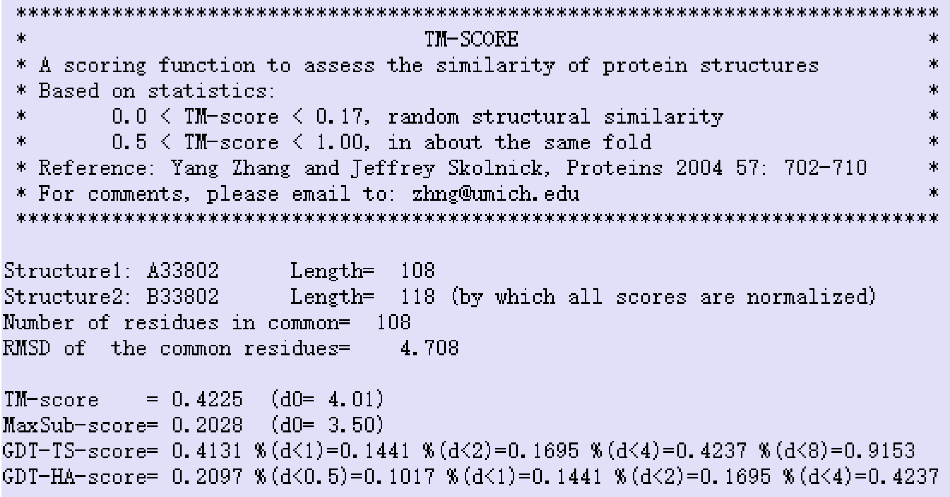


Figure4.18: Reconstructed model.



```
 *****************************************************************************
 *                               TM-SCORE                                    *
 * A scoring function to assess the similarity of protein structures         *
 * Based on statistics:                                                      *
 *       0.0 < TM-score < 0.17, random structural similarity                 *
 *       0.5 < TM-score < 1.00, in about the same fold                       *
 * Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710     *
 * For comments, please email to: zhng@umich.edu                             *
 *****************************************************************************

Structure1: A33802      Length=  108
Structure2: B33802      Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.708

TM-score    = 0.4225  (d0= 4.01)
MaxSub-score= 0.2028  (d0= 3.50)
GDT-TS-score= 0.4131 %(d<1)=0.1441 %(d<2)=0.1695 %(d<4)=0.4237 %(d<8)=0.9153
GDT-HA-score= 0.2097 %(d<0.5)=0.1017 %(d<1)=0.1441 %(d<2)=0.1695 %(d<4)=0.4237
```

Figure 4.19: Protein model TM-score before adding beta sheet contacts

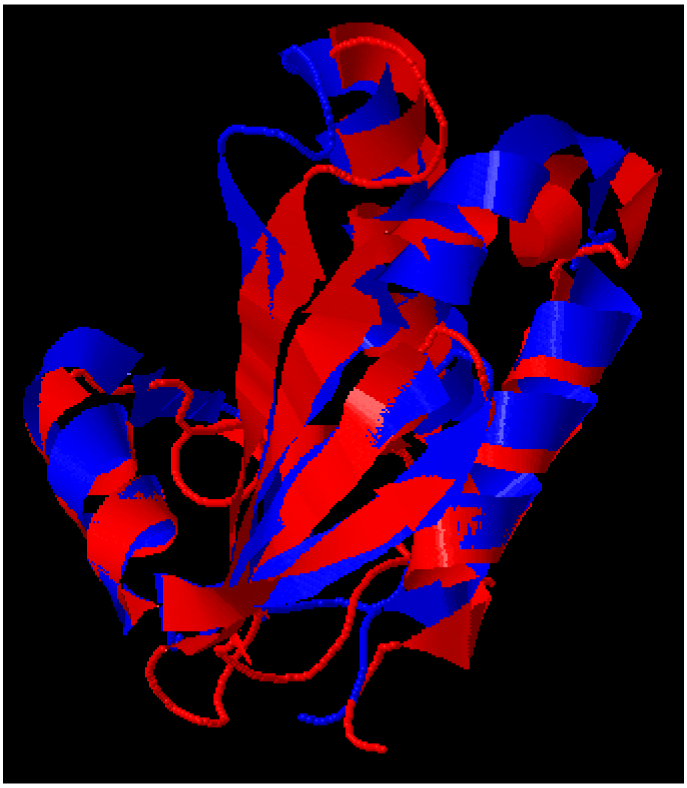
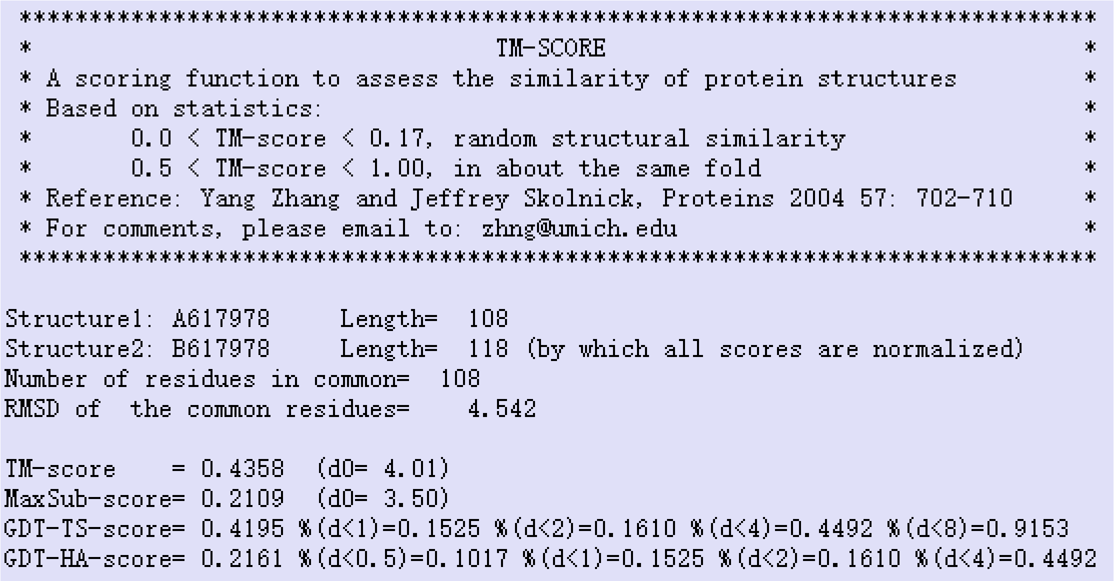


Figure4.20: Visualization of TM-score superposition



```
*****************************************************************************
*                                 TM-SCORE                                  *
* A scoring function to assess the similarity of protein structures        *
* Based on statistics:                                                      *
*       0.0 < TM-score < 0.17, random structural similarity                 *
*       0.5 < TM-score < 1.00, in about the same fold                       *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710     *
* For comments, please email to: zhng@umich.edu                             *
*****************************************************************************

Structure1: A617978     Length=  108
Structure2: B617978     Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.542

TM-score    = 0.4358  (d0= 4.01)
MaxSub-score= 0.2109  (d0= 3.50)
GDT-TS-score= 0.4195 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4492 %(d<8)=0.9153
GDT-HA-score= 0.2161 %(d<0.5)=0.1017 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4492
```

Figure4.21: Protein model TM-score after adding beta-sheet contacts

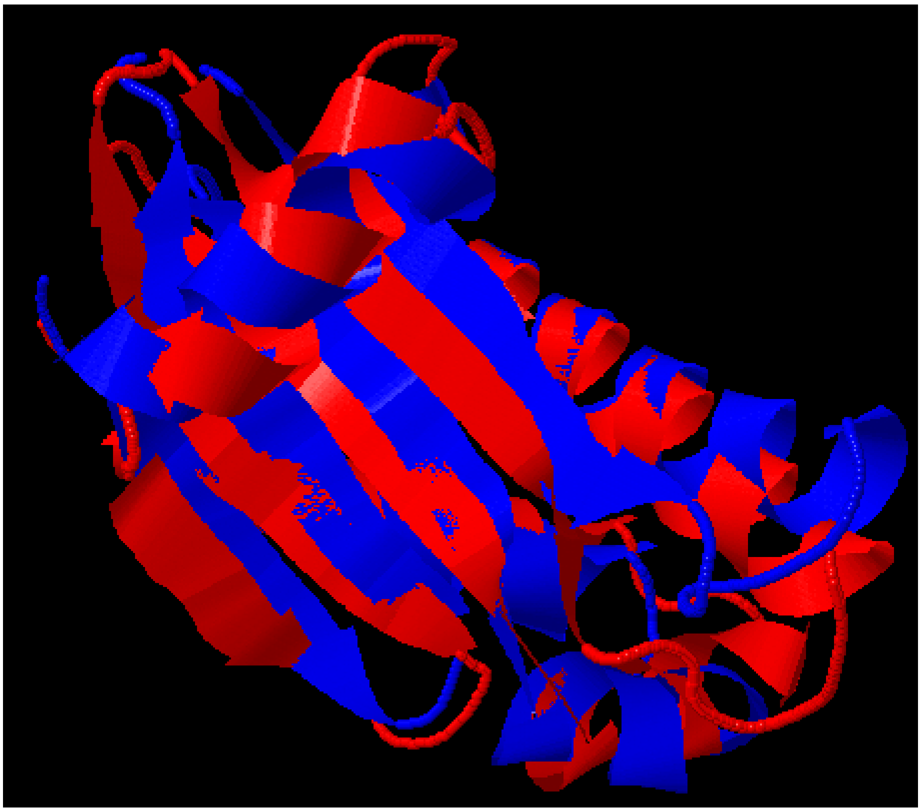


Figure4.22: Visualization of TM-score superposition

The TM-score shows us that the score has not been significantly improved. I compared the TM-scores of the first stage and found that the improvement of the models is visible.

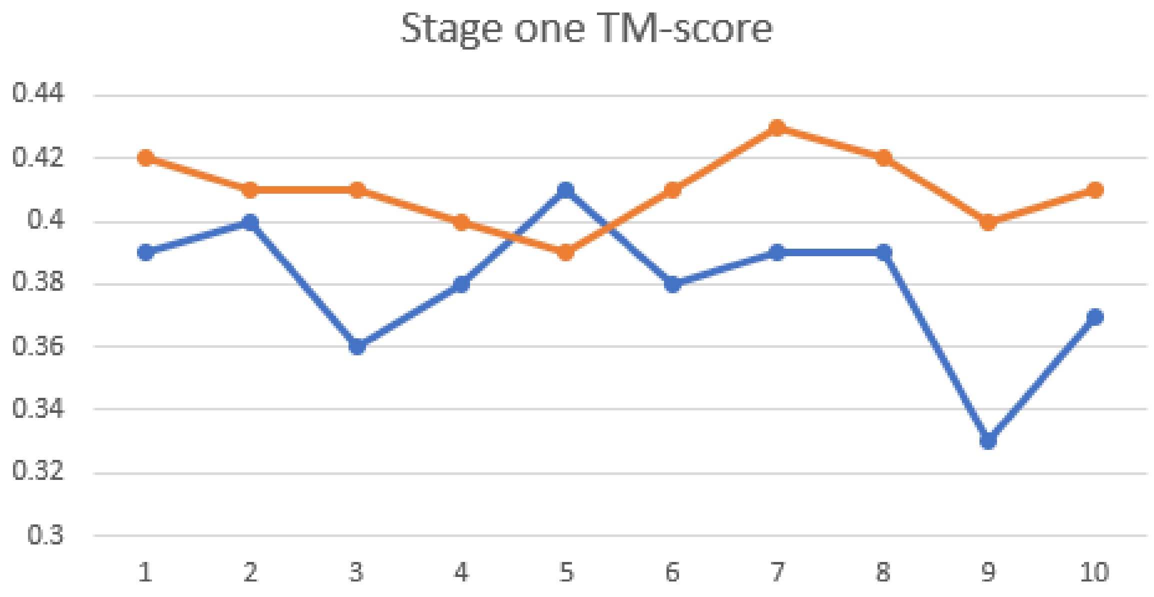


Figure4.23: TM-score in the first stage. (red: new version, blue: previous version)

From the line chart, we can see that the performance of our new version can be more stable. There are too many factors that can affect our result, such as prediction accuracy; it's challenging to improve TM-score significantly. In our new version, we improved our best model TM-score from 0.4225 to 0.4359. It is about 3.14%. And I believe if the model contains more beta-sheet contacts, the performance can be better.

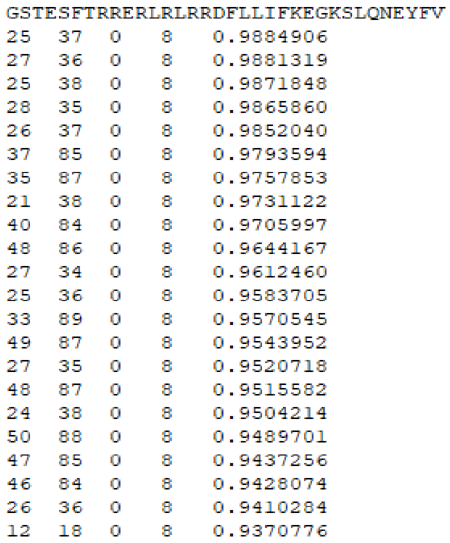
And I think the reason why the resulting model's TM-score is quite similar to the previous version is that the earlier version can detect the hydrogen bonds from generated models. CONFOLD can recognize the beta contacts from the model created by stage 1 and using this information into the next stage. In the future, maybe we can try to reconstruct a protein which contains many beta-sheet contacts, I think the TM-score can be improved more significantly.

Chapter 5 Contacts probability

5.1 background.

One of the input files is the contact prediction results in an “id.rr” file, which contains the residue-residue separation prediction. There are five columns in the RR file: residue number indices i, residue number indices j, distance 1, distance two and probability.

Residues number indices i and j are used for distance specification, the distance one and distance 2 indicate the range of Cβ-Cβ distance predicted for the residue pair (Cα for glycine), and the probability suggests the probability of the distance falling between the predicted range.



```
GSTESFTRRERLRLRRDFLLIFKEGKSLQNEYFV
25   37   0    8    0.9884906
27   36   0    8    0.9881319
25   38   0    8    0.9871848
28   35   0    8    0.9865860
26   37   0    8    0.9852040
37   85   0    8    0.9793594
35   87   0    8    0.9757853
21   38   0    8    0.9731122
40   84   0    8    0.9705997
48   86   0    8    0.9644167
27   34   0    8    0.9612460
25   36   0    8    0.9583705
33   89   0    8    0.9570545
49   87   0    8    0.9543952
27   35   0    8    0.9520718
48   87   0    8    0.9515582
24   38   0    8    0.9504214
50   88   0    8    0.9489701
47   85   0    8    0.9437256
46   84   0    8    0.9428074
26   36   0    8    0.9410284
12   18   0    8    0.9370776
```

Figure5.1: Format of the contacts prediction results.

Figure 5.1 shows an example of the contact prediction results. In this example, the Cβ-Cβ less than 8 Å so that it can be predicted with the format as

i j 0 8 p

In the previous CONFOLD version, the value of probability is not fully utilized. It can generate 40 different subsets of predicted contacts results by selecting top xL contacts. In some cases, this method may miss some essential contact information.

5.2 CONFOLD Version 2

CONFOLD Version 2 generates 40 different subsets and selects 5 top models from each of them. So, it can predict 200 models using a various subset of input contacts. Each subset selects top xL contacts from the RR file, x = 0.1, 0.2, 0.3, …, 4.0 (total 40 items) and L is the length of the protein sequence.

Under this pattern, if the length of the protein sequence is 1000, in the 3.0L stage, input RR file needs (3 * 1000) 3000 contacts distance. After using multiple thresholds contacts probability pattern, we need to provide 1000 contacts restrains as input. It can save a lot of running time.

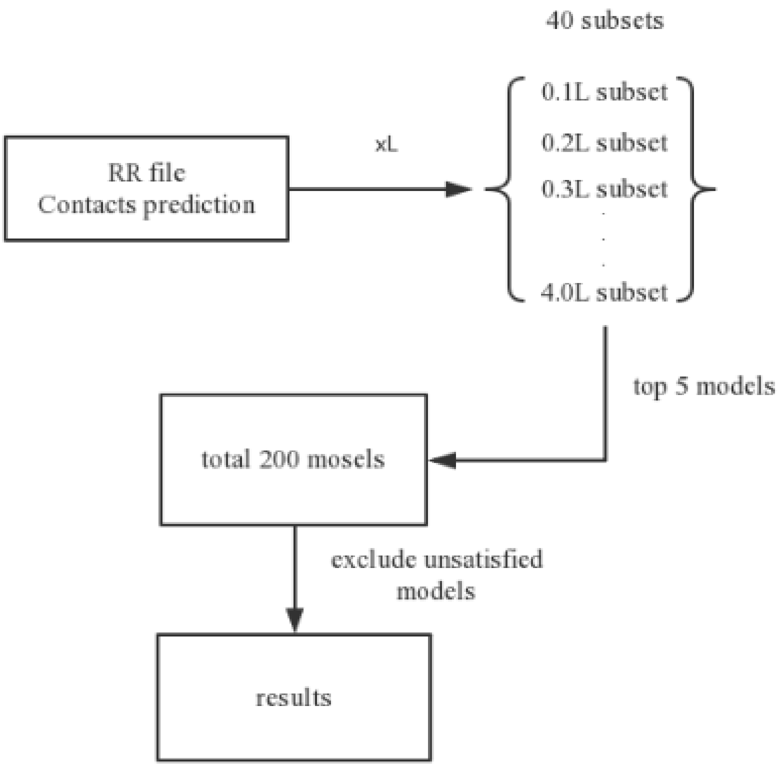
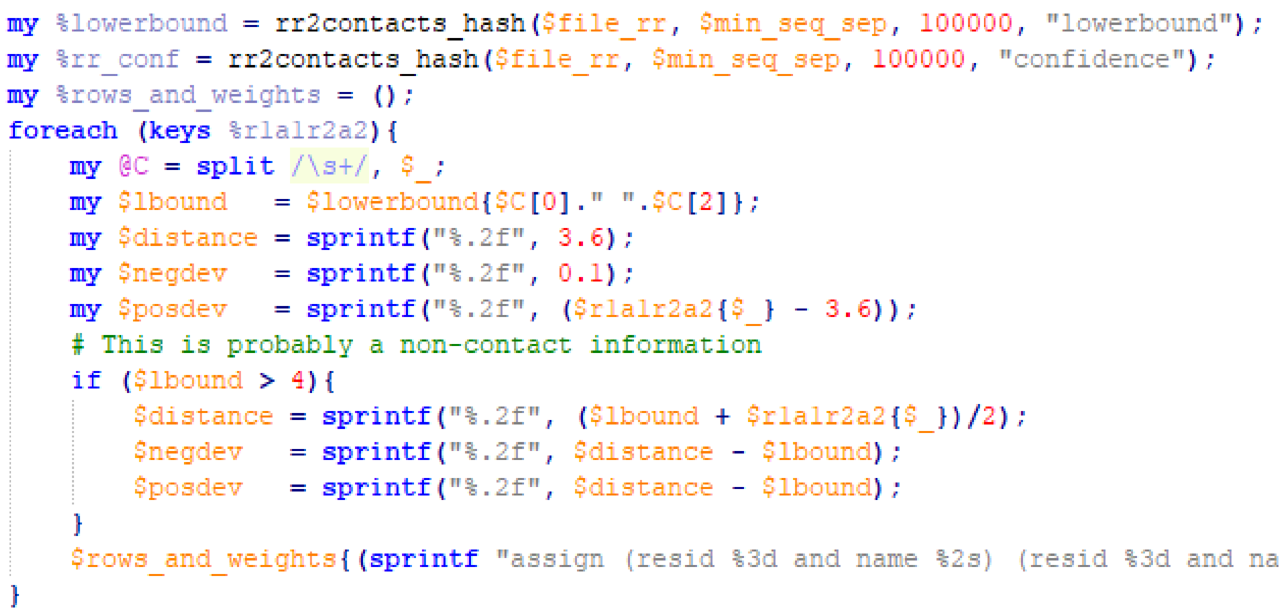


Figure5.2: Process of CONFOLD Version 2.

Figure 5.2 shows the process of CONFOLD 2, dealing with the prediction of the contact. After resulting in a total of 200 models, it calculates the contact satisfaction score using top L/5 long-range contacts and sorts, and the top 50 models will be selected. Then the 50 models separate into 5 clusters and choose the best model from each cluster to form the results.

This way can significantly improve the performance, but it will also make the entire program running time too long. And the program only selects the top xL contacts, and this method may lose some vital information.



```
my %lowerbound = rr2contacts_hash($file_rr, $min_seq_sep, 100000, "lowerbound");
my %rr_conf = rr2contacts_hash($file_rr, $min_seq_sep, 100000, "confidence");
my %rows_and_weights = ();
foreach (keys %rlalr2a2){
    my @C = split /\s+/, $_;
    my $lbound    = $lowerbound{$C[0]." ".$C[2]};
    my $distance = sprintf("%.2f", 3.6);
    my $negdev    = sprintf("%.2f", 0.1);
    my $posdev    = sprintf("%.2f", ($rlalr2a2{$_} - 3.6));
    # This is probably a non-contact information
    if ($lbound > 4){
        $distance = sprintf("%.2f", ($lbound + $rlalr2a2{$_})/2);
        $negdev    = sprintf("%.2f", $distance - $lbound);
        $posdev    = sprintf("%.2f", $distance - $lbound);
    }
    $rows_and_weights{(sprintf "assign (resid %3d and name %2s) (resid %3d and na
}
```

Figure5.3: CONFOLD 2 using probability value.

In CONFOLD version 2, the probability value is used to detect whether this column is non-contact information. In the new version, we are trying to use the probability value as thresholds to judge how many distances should be added.

5.3 Multiple Thresholds contacts probability.

The probability of the distance between Cβ atoms is within the range of 0 to 1. To make sure every contact prediction has the chance to be selected, we choose to use multiple threshold methods to select the contacts.

Multiple thresholds method is to divide the entire data set into several clusters by using the value of probability as an indicator, different clusters have different weights, and the weight is used to determine the proportion of the cluster.

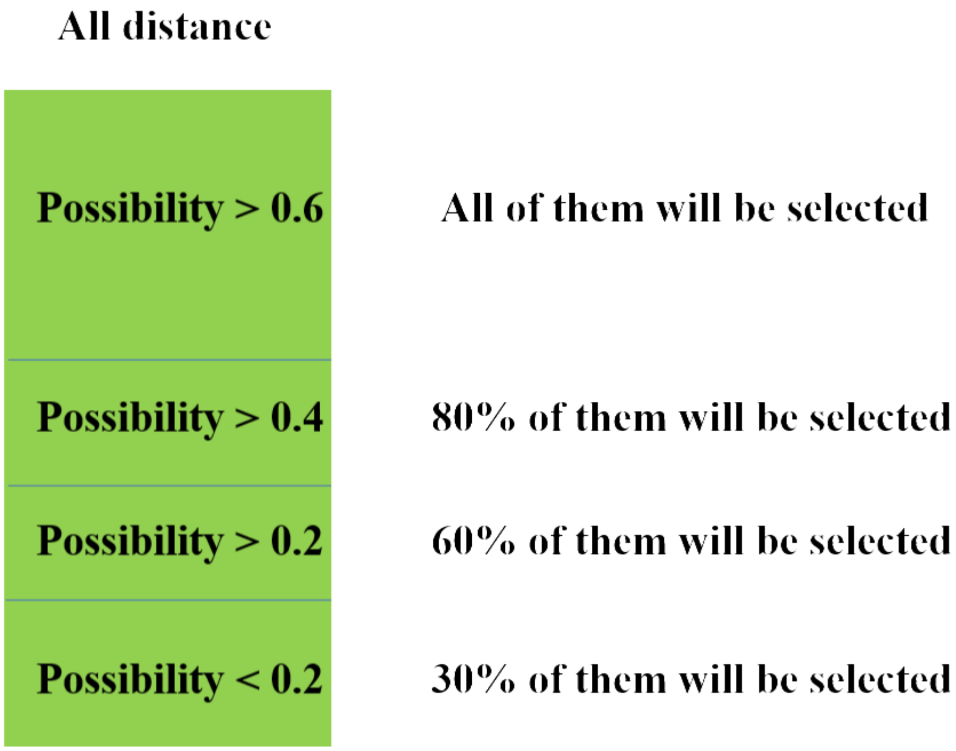
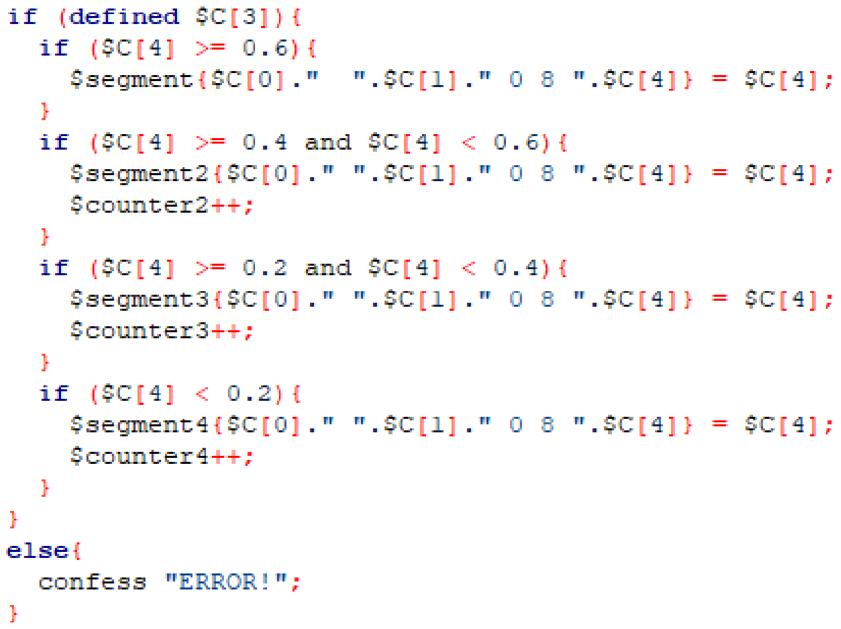

All distance
Possibility > 0.6
All of them will be selected
Possibility > 0.4
80% of them will be selected
Possibility > 0.2
60% of them will be selected
Possibility < 0.2
30% of them will be selected

Figure5.4: Example about how to set thresholds

Figure 5.4 shows an example of how to set the probability thresholds. In this example, we set three thresholds: 0.6, 0.4 and 0.2. If a cluster's probability of the residue-residue contacts is greater than 0.6, it means that the confidence of this cluster is high so that we will select all the residue-residue contacts. If a cluster's probability of the residue-residue contacts is between 0.6-0.8, we will choose 80% of them into protein reconstruction. And if a cluster's possibility is less than 0.2, it means that the confidence of this cluster is low, so that we will select only 30% of them.

In a RR file, most of the residue-residue contacts' probability is between 0 to 0.3, and this method can exclude most of the low probability contacts and save the running time.



```
if (defined $C[3]){
  if ($C[4] >= 0.6){
    $segment{$C[0]."  ".$C[1]." 0 8 ".$C[4]} = $C[4];
  }
  if ($C[4] >= 0.4 and $C[4] < 0.6){
    $segment2{$C[0]." ".$C[1]." 0 8 ".$C[4]} = $C[4];
    $counter2++;
  }
  if ($C[4] >= 0.2 and $C[4] < 0.4){
    $segment3{$C[0]." ".$C[1]." 0 8 ".$C[4]} = $C[4];
    $counter3++;
  }
  if ($C[4] < 0.2){
    $segment4{$C[0]." ".$C[1]." 0 8 ".$C[4]} = $C[4];
    $counter4++;
  }
}
else{
  confess "ERROR!";
}
```

Figure5.5: Coding to divide the clusters.

After separating the entire data set into four clusters, we need to select the contacts from each cluster. Since there are 40 subsets in the program, we require to promise every contact has the chance to be chosen. So, the best way is to select contacts from the cluster randomly.

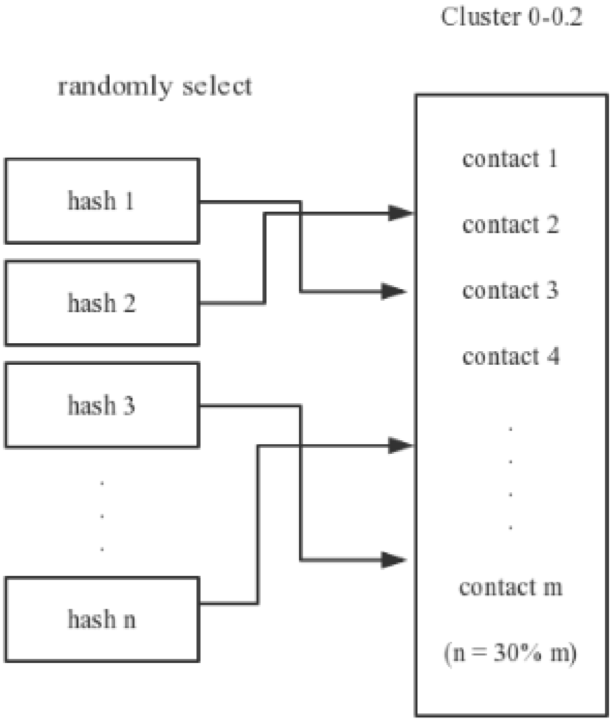
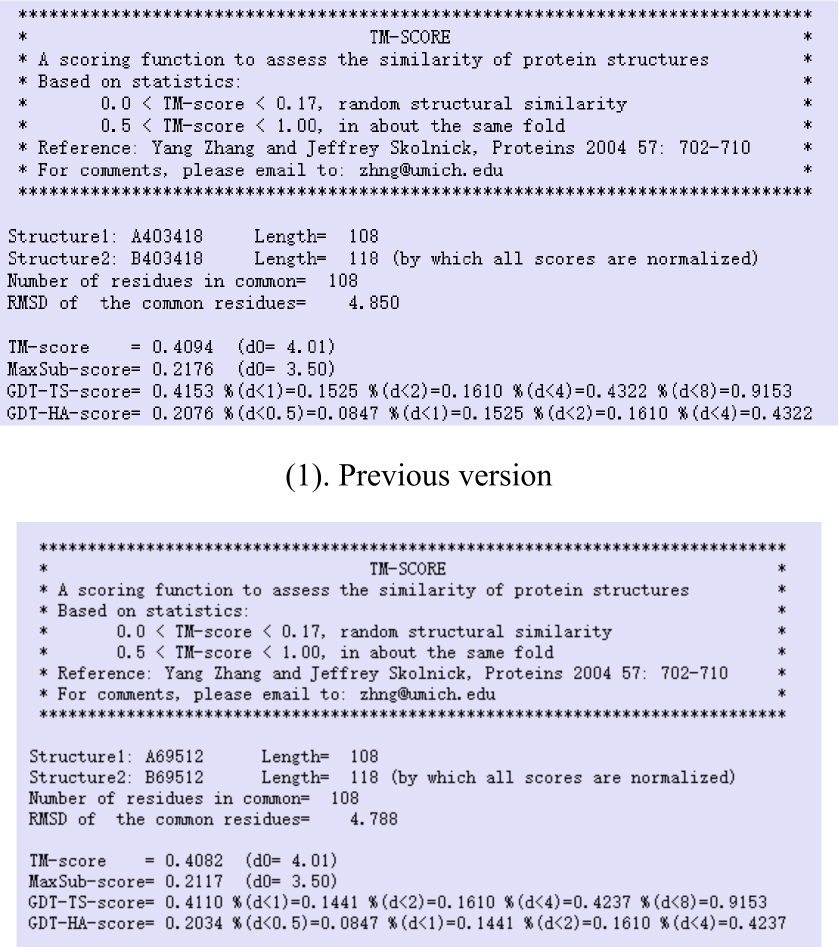


Figure5.6: the process of randomly selecting contacts from the cluster.

Then we need to integrate the selected residue-residue contacts to form a complete RR file. But this RR file is unsatisfied with the requirement of CONFOLD, and it should be sorted and add the protein sequence (fasta file) on the first line.

5.4 Results.

To compare the performance with the previous version, we tested the program on CASP 12. Because of the limitation of the machine, we just selected some of the datasets to get the TM-score of the resulting models and the running time of the entire program.



```
 ******************************************************************************
 *                                 TM-SCORE                                   *
 * A scoring function to assess the similarity of protein structures          *
 * Based on statistics:                                                       *
 *       0.0 < TM-score < 0.17, random structural similarity                  *
 *       0.5 < TM-score < 1.00, in about the same fold                        *
 * Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710      *
 * For comments, please email to: zhng@umich.edu                              *
 ******************************************************************************

Structure1: A403418     Length=  108
Structure2: B403418     Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.850

TM-score    = 0.4094  (d0= 4.01)
MaxSub-score= 0.2176  (d0= 3.50)
GDT-TS-score= 0.4153 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4322 %(d<8)=0.9153
GDT-HA-score= 0.2076 %(d<0.5)=0.0847 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4322
```

(1). Previous version

```
 ******************************************************************************
 *                                 TM-SCORE                                   *
 * A scoring function to assess the similarity of protein structures          *
 * Based on statistics:                                                       *
 *       0.0 < TM-score < 0.17, random structural similarity                  *
 *       0.5 < TM-score < 1.00, in about the same fold                        *
 * Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710      *
 * For comments, please email to: zhng@umich.edu                              *
 ******************************************************************************

Structure1: A69512      Length=  108
Structure2: B69512      Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.788

TM-score    = 0.4082  (d0= 4.01)
MaxSub-score= 0.2117  (d0= 3.50)
GDT-TS-score= 0.4110 %(d<1)=0.1441 %(d<2)=0.1610 %(d<4)=0.4237 %(d<8)=0.9153
GDT-HA-score= 0.2034 %(d<0.5)=0.0847 %(d<1)=0.1441 %(d<2)=0.1610 %(d<4)=0.4237
```

(2). Multiple thresholds version.

Figure5.7: The TM-score from a different version.

Figure 5.7 shows that the TM-scores from different version is quite similar; it means that the multiple thresholds probability method is a reliable way to select the prediction of the contact from the RR file.

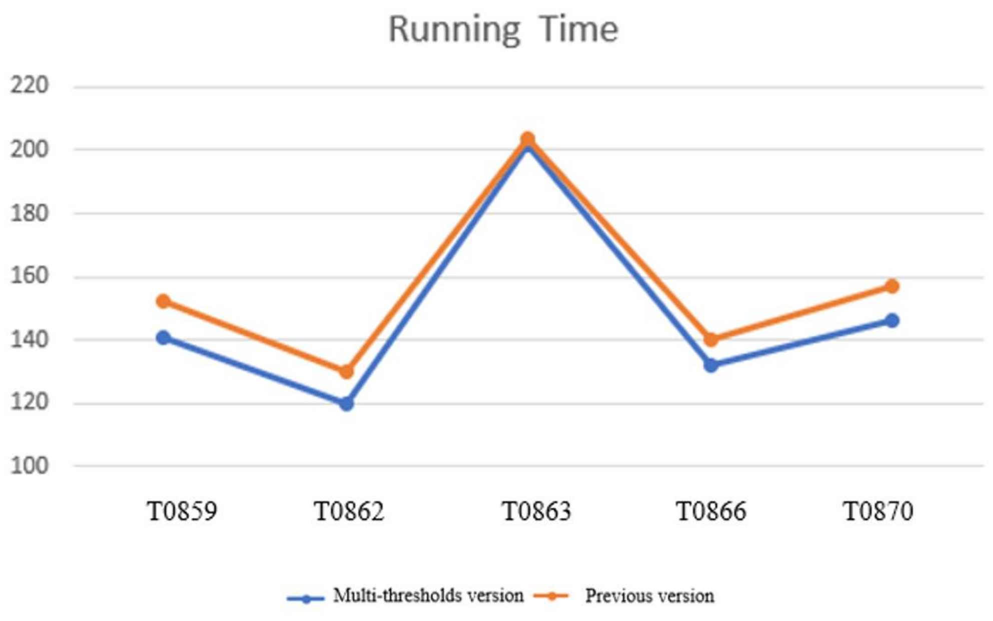


Figure5.8: The comparison of the two versions.

Figure 5.8 shows that the multiple thresholds probability version can be faster than the previous version. We take protein T0859 and T0870 as examples, the running time of two protein sequences in the last version is 154.36 minutes and 159.63 minutes, and the running time in multiple thresholds version is 140.58 minutes and 143.74 minutes. The improvement of these two examples is 8.9% and 9.95%. It can improve efficiency while keeping accuracy.

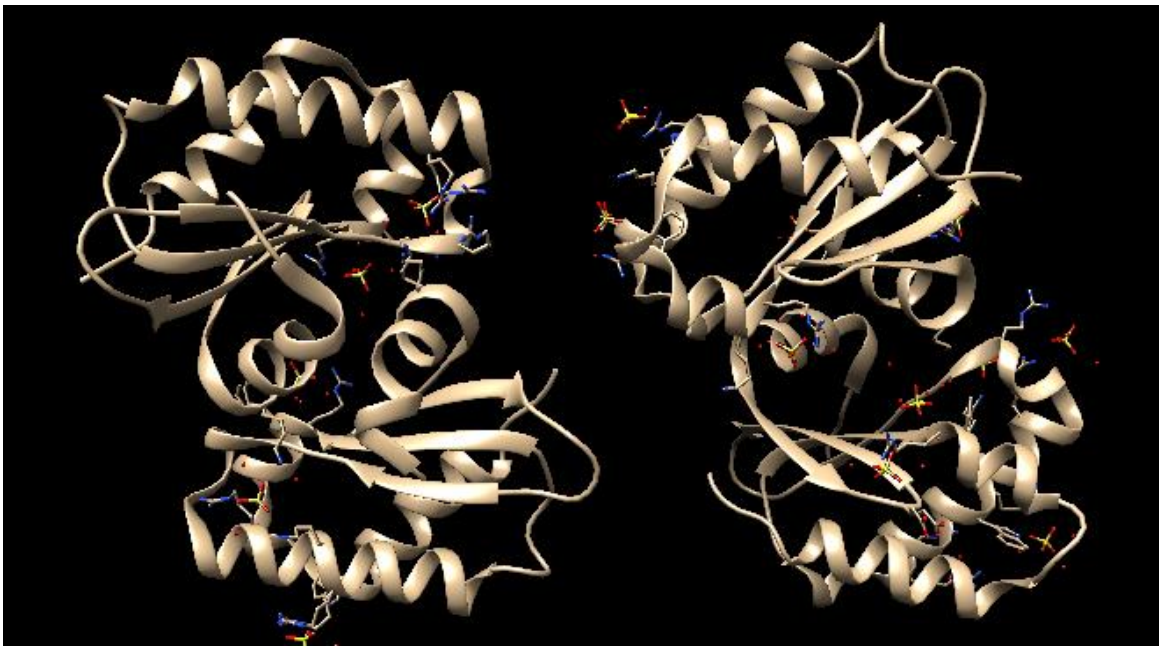
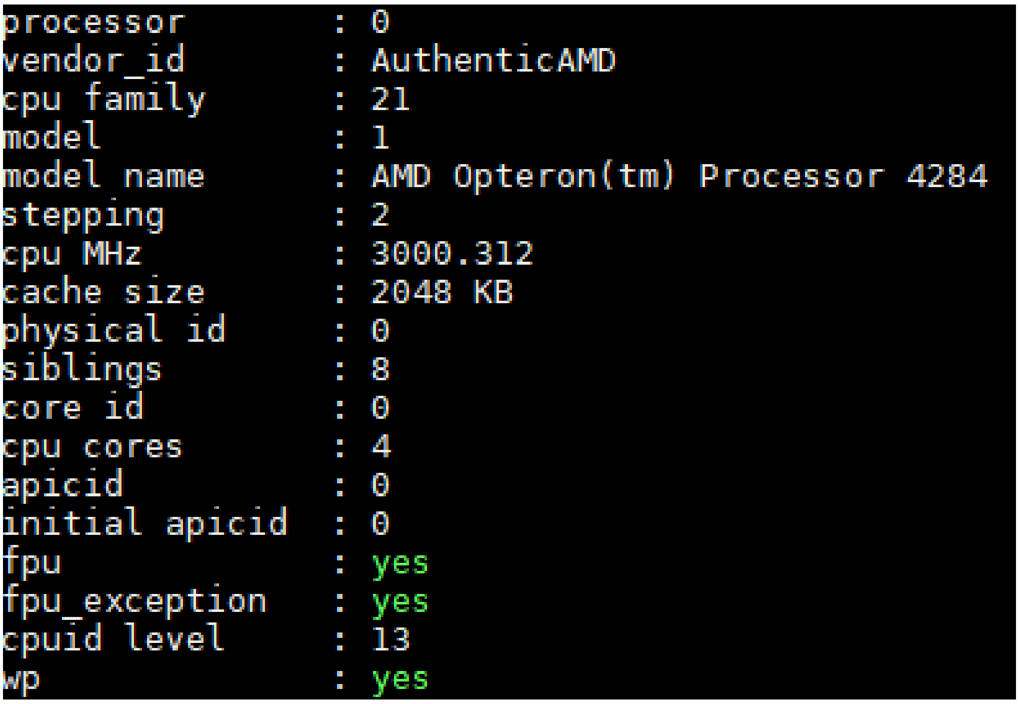


Figure5.9: Visualize the 1nz0 protein model.

This TM-score is generated based on the thresholds 0.6, 0.4, and 0.2, in the future we can try some different thresholds. And we can use different weights in different clusters to improve the performance.

Chapter 6 Summary

In this research, we ran our system based on the Red Hat Enterprise Linux Server release 6.4 (Santiago), and CPU has four cores. We tested our results based on the dataset CASP 12, which can provide research groups with the opportunity to test the structure prediction methods. CASP is a Critical Assessment of protein Structure Prediction, and it can help advance the methods of identifying protein 3-D structure from its amino acid sequence.



```
processor       : 0
vendor_id       : AuthenticAMD
cpu family      : 21
model           : 1
model name      : AMD Opteron(tm) Processor 4284
stepping        : 2
cpu MHz         : 3000.312
cache size      : 2048 KB
physical id     : 0
siblings        : 8
core id         : 0
cpu cores       : 4
apicid          : 0
initial apicid  : 0
fpu             : yes
fpu_exception   : yes
cpuid level     : 13
wp              : yes
```

Figure6.0.1: Linux system information.

Under this system, we tested our new version CONFOLD, and we get the running time information. The length of protein T0859 is 129, in the CONFOLD version 2 running time is 154.36 minutes, and in the new version, the running time is 140.58 minutes. We have improved efficiency by 8.9%.

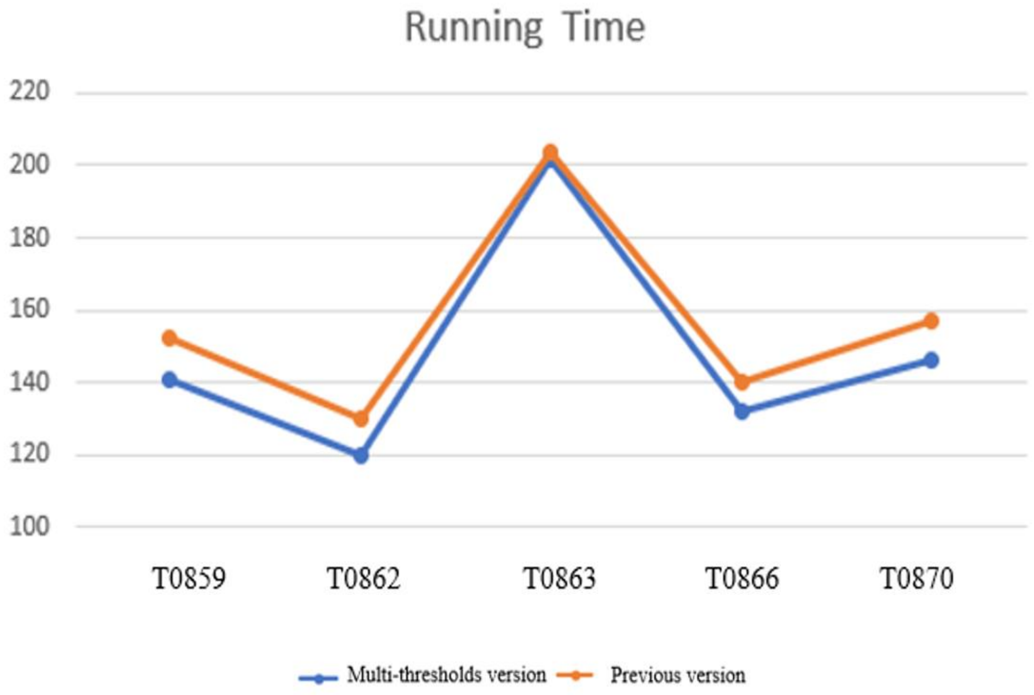
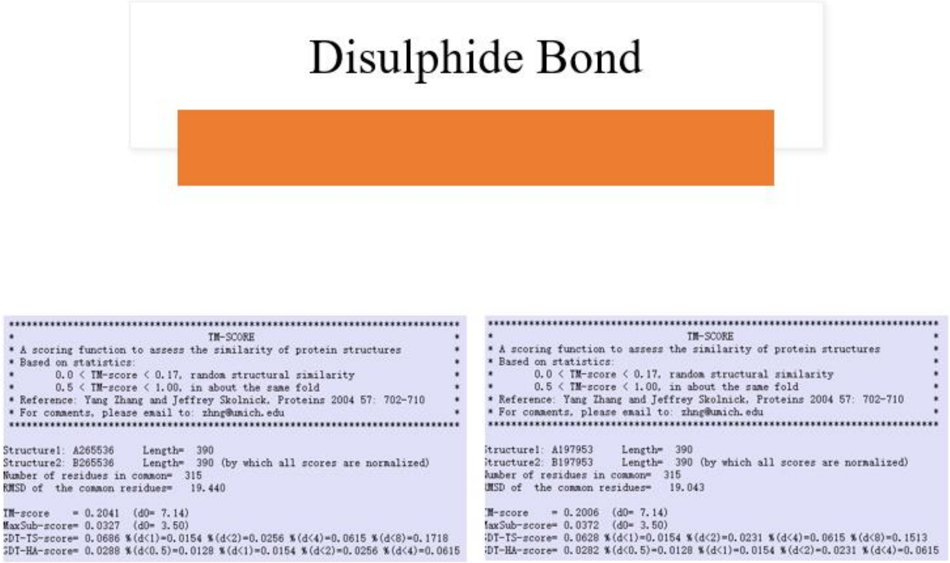


Figure6.2: Running time improvement.

And for the protein which contains the disulfide bond information such as protein 1a4g, the TM-score can be improved from 0.2006 to 0.2041. We have developed the accuracy by 1.74%.



Disulphide Bond

```
**********************************************************************
*                               TM-SCORE                             *
* A scoring function to assess the similarity of protein structures  *
* Based on statistics:                                               *
*       0.0 < TM-score < 0.17, random structural similarity          *
*       0.5 < TM-score < 1.00, in about the same fold                *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710 *
* For comments, please email to: zhng@umich.edu                      *
**********************************************************************

Structure1: A265536    Length=  390
Structure2: B265536    Length=  390 (by which all scores are normalized)
Number of residues in common=  315
RMSD of  the common residues=   19.440

TM-score    = 0.2041  (d0= 7.14)
MaxSub-score= 0.0327  (d0= 3.50)
GDT-TS-score= 0.0686 %(d<1)=0.0154 %(d<2)=0.0256 %(d<4)=0.0615 %(d<8)=0.1718
GDT-HA-score= 0.0288 %(d<0.5)=0.0128 %(d<1)=0.0154 %(d<2)=0.0256 %(d<4)=0.0615
```

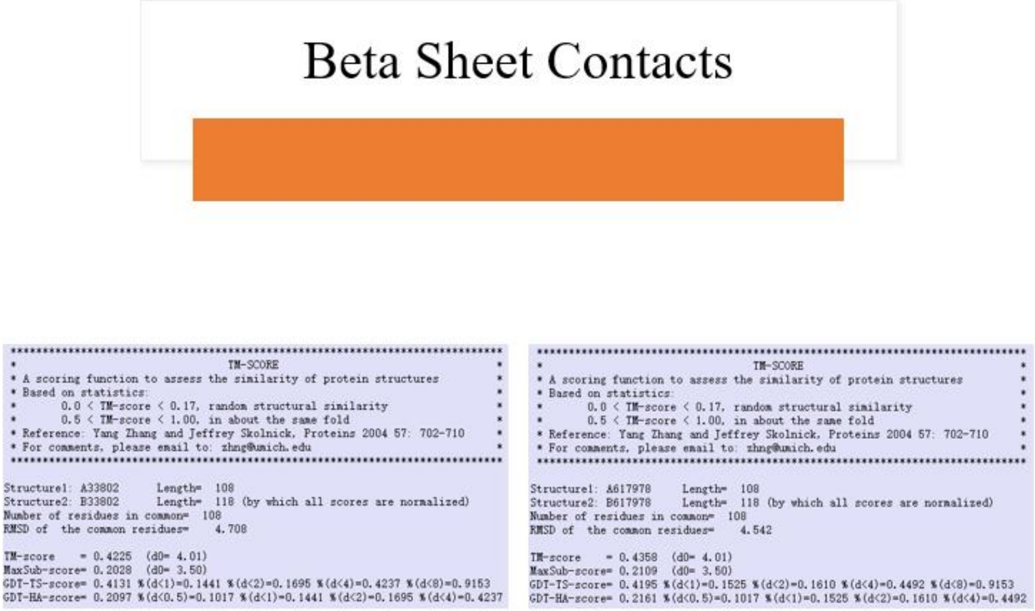
```
**********************************************************************
*                               TM-SCORE                             *
* A scoring function to assess the similarity of protein structures  *
* Based on statistics:                                               *
*       0.0 < TM-score < 0.17, random structural similarity          *
*       0.5 < TM-score < 1.00, in about the same fold                *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710 *
* For comments, please email to: zhng@umich.edu                      *
**********************************************************************

tructure1: A197953    Length=  390
tructure2: B197953    Length=  390 (by which all scores are normalized)
umber of residues in common=  315
MSD of  the common residues=   19.043

M-score    = 0.2006  (d0= 7.14)
axSub-score= 0.0372  (d0= 3.50)
DT-TS-score= 0.0628 %(d<1)=0.0154 %(d<2)=0.0231 %(d<4)=0.0615 %(d<8)=0.1513
DT-HA-score= 0.0282 %(d<0.5)=0.0128 %(d<1)=0.0154 %(d<2)=0.0231 %(d<4)=0.0615
```

Figure6.3: Disulfide bond feature improvement.

For the protein which contains the beta-sheets contacts information such as 1nzD, the TM-score can be improved from 0.4225 to 0.4358. We have developed the accuracy by 3.14%.



Beta Sheet Contacts

```
**********************************************************************
*                               TM-SCORE                             *
* A scoring function to assess the similarity of protein structures  *
* Based on statistics:                                               *
*       0.0 < TM-score < 0.17, random structural similarity          *
*       0.5 < TM-score < 1.00, in about the same fold                *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710 *
* For comments, please email to: zhng@umich.edu                      *
**********************************************************************

Structure1: A33802     Length=  108
Structure2: B33802     Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.708

TM-score    = 0.4225  (d0= 4.01)
MaxSub-score= 0.2028  (d0= 3.50)
GDT-TS-score= 0.4131 %(d<1)=0.1441 %(d<2)=0.1695 %(d<4)=0.4237 %(d<8)=0.9153
GDT-HA-score= 0.2097 %(d<0.5)=0.1017 %(d<1)=0.1441 %(d<2)=0.1695 %(d<4)=0.4237
```

```
**********************************************************************
*                               TM-SCORE                             *
* A scoring function to assess the similarity of protein structures  *
* Based on statistics:                                               *
*       0.0 < TM-score < 0.17, random structural similarity          *
*       0.5 < TM-score < 1.00, in about the same fold                *
* Reference: Yang Zhang and Jeffrey Skolnick, Proteins 2004 57: 702-710 *
* For comments, please email to: zhng@umich.edu                      *
**********************************************************************

Structure1: A617978    Length=  108
Structure2: B617978    Length=  118 (by which all scores are normalized)
Number of residues in common=  108
RMSD of  the common residues=    4.542

TM-score    = 0.4358  (d0= 4.01)
MaxSub-score= 0.2109  (d0= 3.50)
GDT-TS-score= 0.4195 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4492 %(d<8)=0.9153
GDT-HA-score= 0.2161 %(d<0.5)=0.1017 %(d<1)=0.1525 %(d<2)=0.1610 %(d<4)=0.4492
```

Figure6.4: Beta sheet contacts feature improvement.

The reconstruction of the protein model is a very complicated task, which contains many influencing factors. In this research, we improved the CONFOLD system with three new features, and the results show that in the new version, it can perform better.

Chapter 7 Future Work

In the future, we can continue to improve the performance of CONFOLD. According to the disulfide bonds part, now we can use DIpro2 to predict the position of pairs of cysteines, but we still cannot recognize the residue is thioredoxin or peptide. CNS can receive this kind of information to make the structure more reliable.

What’s more, we can modify the thresholds and then observe which set of thresholds will get the best TM-score and the fastest running time. In this way, the performance of CONFOLD will be improved.

We can also integrate the HHblits, HHfilter, CCMpred, and bbcontacts into one program so that the user can save a lot of time generating multiple sequences and direct couplings.

Chapter 8 Conclusion

In this research, we got familiar with the CONFOLD and added some new features to it. Because the CONFOLD is built based on CNS solve, we also spent plenty of time studying how to use CNS.

The first feature of this new version is disulfide bond prediction, using DIpro2 to predict the information is my first step. Then the next step is to figure out how to make CNS using the prediction, and the tutorial told me to write those cysteines position into the Molecular topology file. Since then, the new version can recognize disulfide bond information from input files.

Adding Beta-sheet contacts prediction into CONFOLD is also a new feature. During this part, we focused on how to recognize the direction of the beta-strands. After a long period of research, we found “bbcontacts” which can predict the position and direction of the beta-sheet contacts. The “bbcontacts” require the secondary structure file and direct couplings matrix, so we need to use HHblits, CCMpred and Psipred to generate the required input files. After getting the prediction of beta-sheet contacts, we need to exclude the unpaired strands. And we are then writing the beta strands information into “hbond.tbl” file to reconstruct protein.

And the last new feature is multiple thresholds contacts probability. CONFOLD is a residue-residue contact-guided ab initio protein folding method, but the value of probability in the RR file is not used. In this research, we separated contacts into different clusters and gave each cluster a weight. It can make the program faster and keep the TM-score.

Reference

[1] Jianlin Cheng, Hiroto Saigo, Pierre Baldi, "Large-Scale Prediction of Disulphide Bridges Using Kernel Methods, Two-Dimensional Recursive Neural Networks, and Weighted Graph Matching". Proteins: Structure, Function, Bioinformatics, vol 62, no. 3, pp. 617-629, 2006.

[2] Jianlin Cheng, Hiroto Saigo, Pierre Baldi, "Large-Scale Prediction of Disulphide Bridges Using Kernel Methods, Two-Dimensional Recursive Neural Networks, and Weighted Graph Matching". Proteins: Structure, Function, Bioinformatics, vol 62, no. 3, pp. 617-629, 2006.

[3] Pierre Baldi, Jianlin Cheng, Alessandro Vullo, "Large-Scale Prediction of Disulphide Bond Connectivity", Advances in Neural Information Processing Systems (NIPS 2004) 17, L. Saul, Y. Weiss, and L. Bottou editors, pp.97-104, MIT Press, Cambridge, MA, 2005.

[4] J. Cheng, A. Randall, M. Sweredoski, P. Baldi, SCRATCH: A Protein Structure and Structural Feature Prediction Server, Nucleic Acids Research, vol. 33 (web server issue), w72-76, 2005.

[5] Gene Center, LMU Munich, Feodor-Lynen-Strasse 25, 81377, Munich and Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany.

[6] HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Remmert M, Biegert A, Hauser A, Söding J. Nat Methods. 2011 Dec 25;9(2):173-5.

[7] A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.

[8] Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. J Mol Biol. 2018 Jul 20. S0022-2836(17)30587-9.

[9] Remmert M., Biegert A., Hauser A., Söding J. (2011) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 9(2):173-5. doi: 10.1038/nmeth.1818. PMID: 22198341

[10] Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. J Mol Biol. 2018 Jul 20. S0022-2836(17)30587-9.

[11] M. Remmert, A. Biegert, A. Hauser, and J. Soeding (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9, 173-175.

[12] J. Cheng and P. Baldi (2005). Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms. Bioinformatics, 21 Suppl 1, 75-84. Link to the BetaSheet916 dataset file (last accessed 12 September 2014)

[13] D. T. Jones (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292, 195-202.

[14] Proteins: Structure, Function, and Bioinformatics, 2015.B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng.

[15] Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins Struct Funct Bioinforma. 2016; 84(S1):131–44.

[16] Michel M, Hurtado DM, Uziela K, Elofsson A. Large-scale structure prediction by improved contact predictions and model quality assessment. bioRxiv. 2017;128231.

[17] Adhikari B, Bhattacharya D, Cao R, Cheng J. CONFOLD: Residue-residue contact-guided ab initio protein folding. Proteins. 2015; 83(8):1436–49.

[18] Nilges M, Gronenborn AM, Brünger AT, Clore GM. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. Protein Eng Des Sel. 1988; 2(1):27–38.

[19] Michel M, Hayat S, Skwarek MJ, Sander C, Marks DS, Elofsson A. Pconsfold: improved contact predictions improve protein models. Bioinformatics. 2014; 30(17):482–8.

[20] Kosciolek T, Jones DT. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE. 2014; 9(3):e92197.

[21] Mabrouk M, Werner T, Schneider M, Putz I, Brock O. Analysis of free modeling predictions by rbo aleph in casp11. Proteins Struct Funct Bioinforma. 2016; 84(S1):87–104.