CHAPTER 3

METHODOLOGY

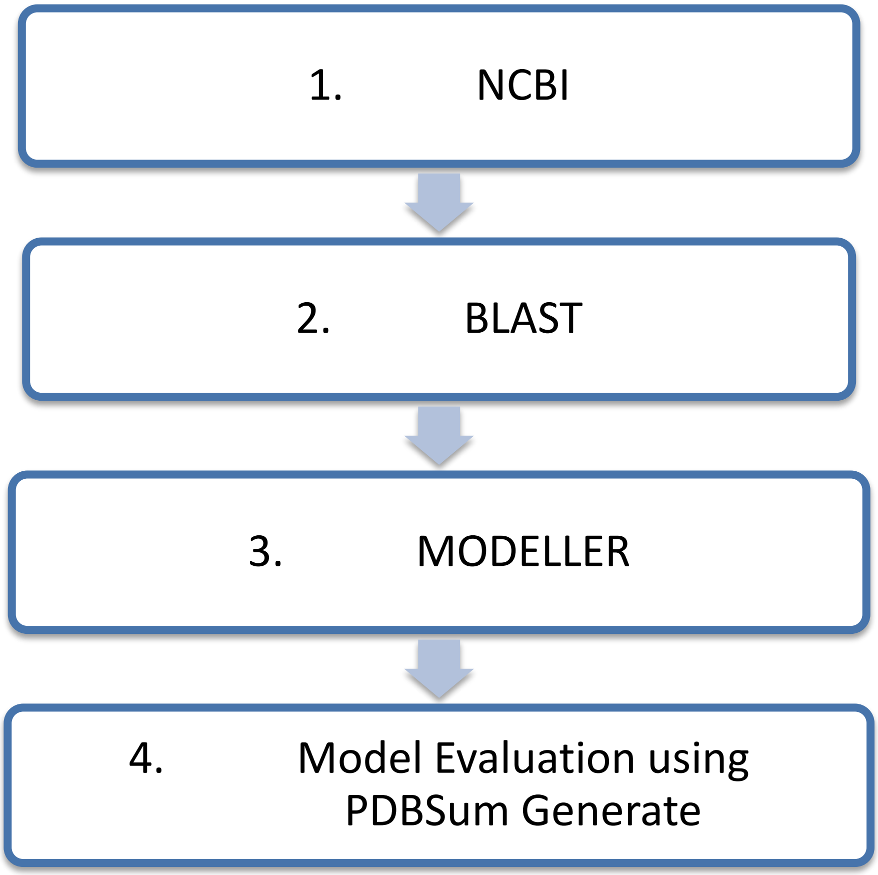

Figure 3.1: Flowchart of Methodology

3.1 Research Design

Template-based modeling technique or known as homology modeling used the concept of evolutionary relationship between target and template sequence whereby the template sequence has been known of its three dimensional structure. This is because sequences that are evolutionary related have the high chance of having similar 3D structure. In homology modeling there are several steps in establishing the most correct and satisfactory model.

1. Identify suitable template sequences with approved 3D structures that related to the target sequence and by aligning the target and template sequences either pairwise alignment or multiple alignment,
2. The target is modeled according to the structure of template sequences. The modeling of structures can be done using many software that available,
3. Refinement of model,
4. Evaluate the model.

This research is conducted in order to determine the structure of DmGSTD3 as well as identifying the amino acid residue that involve in glutathione conjugation. I adopt the homology modeling techniques in determining and building the three dimensional structure of the model. The methodology section is divided into four parts (Figure 3.1):

1. NCBI – to search for protein sequence,
2. BLAST – to investigate the template sequences for target sequence,
3. MODELLER – to align the template against target sequence, build model for target sequence, and evaluate the model.

4. Model Evaluation using PDBSum Generate – evaluate the model generate from Modeller by sending it through PDBSum Generate to produce the Ramachandran plot of the model.

3.2 Searching for target sequence

National Centre for Biotechnology Information is an organization that stores and provides access to biomedical and genetic information. It contains other databases that interlink among each other and updated on frequent basis. NCBI contains thousands of protein sequences of different species ranging from single-celled microorganisms, archaea, to multi-cellular organisms, human.

The protein sequence that I want to study is originated from *Drosophila melanogaster* with the length of 199 amino acid. There are other databases provided that can be used in protein sequence searching. In investigating the GSTD3 sequence, I used NCBI Protein database provided by NCBI consortium. Searching through the NCBI database, I obtained the glutathione S-transferase D3 with accession number AA041561 and the NCBI reference sequence is NP_788656.1

3.3 Searching for template sequences

To facilitate the search of template sequence, the target sequence undergone blast program in which the database that used as search set is UniProt. Meanwhile, the program is BLAST program in which the query sequence was BLAST against the protein sequences in UniProt database. The reason of choosing UniProt as the search set database is to compare and identify the protein sequences as a whole by taking into account the protein sequences without structures as well. Even though researchers have decided that 40% of sequence identities between target and template sequence may produce a good structure model for target sequence, however, I decided to select template sequences having more than 50% sequence identity since it may produce a better and reliable model for my target sequence Therefore, the template sequence that selected is having above 50% sequence identity with the target sequence.

Although 25% sequence identities are enough in selecting a suitable template, I decided to choose the template sequences of having above 50% sequence identities. After the template sequences have been identified, both of the sequences, template and query, were aligned in pairwise mode and multiple sequence alignment modes. In pairwise alignment, the purpose is to assess the degree of similarity between target and template sequences as well as determining the homology, structure prediction, function prediction, and others. Meanwhile, in multiple sequence alignment, the purpose is to identify the conserved protein by aligning all residues that originated from the same ancestor.

3.4 Modeling the target sequence

In Modeller, it obtained the information in constructing the 3D structures from two sources which are alignment with template sequences and molecular mechanics force field in which are responsible in deriving the data of distance and dihedral angles as well as bond length and bond angle respectively (Figure 3.2) (Webb 2010). The Python scripts in modeling the query sequence were organized in 5 scripts which are PIR file, comparetemplates.py, targettemplatealign.py, and buildmodel.py.

1. The fasta file which contains the target sequence was converted in PIR format file. PIR file stand for protein information resource is readable by the Modeller whereby this type of file is usually used by Modeller to read and write the sequences and alignments. The format of PIR file is the header begin with '>' sign followed by

sequence code which is target (referring to the target sequence, DmGSTD3). Next, in the second line, it consists of ten fields which are separated by colons. Each field specified the details about the structure. In this case, there is no prior knowledge about the structure, except for the first two fields which contain “sequence” (telling the Modeller that this file consist of sequence with unknown structure) and “target”, the model file name. The rest of the lines were protein sequences in uppercase - one letter protein code, enclosed with ‘*’ sign signifying the end of file.

2. In comparetemplates.py, the templates sequences obtained from UniProt database were selected in order to choose an appropriate template for modeling the target sequence. In this Python script file, the pdb file of template sequences were specified and provided since the Modeller read through the protein sequences and information about them. In comparetemplates.py, Modeller calculated the score of multiple sequence alignment in order to align them. From the multiple sequence alignment, several measures were calculated such as score of RMS and DRMS deviations between atomic positions and distances, differences between sidechain and mainchain dihedral angles, sequence identities, and others, which later be used in constructing the clustering tree.
3. After an appropriate template have been identified and selected from those several templates, the template sequence was aligned with the target sequence. Referring to this file, it employed dynamic programming algorithm which is Needleman Wunsch algorithm as method in aligning target and template sequences. In addition, Modeller takes into account the information about structure of the template. In targettemplatealignment.py, Modeller created an empty object to append the target sequence and template sequence in this object. Later, the alignments of target and

template sequences were written into two files which are PIR and PAP file. PIR file is used for model building whereas the PAP file is for visualizing. The identical positions were marked with '*' sign.

4. Subsequent to the construction of target-template alignment, the Modeller calculates the 3D model structure of the target sequence. A set of 100 models were generated based on the information about template and its target-template alignment. One input and two strings were needed which are alignment file in PIR format, as well as target and template sequences, respectively. In determining the best model generated, DOPE score and GA341 were assessed using Modeller by choosing the lowest score of DOPE or highest score of GA341.

The Python scripts in modeling the target sequence using Modeller are attached into the attachment.

3.5 Model Evaluation using PDBSum Generate

Model of query sequence that has been generated by Modeller was evaluated using PDBSum database which provide by European Bioinformatics Institute. The PDBSum generates PROCHECK results which consist of Ramachandran plot, secondary structure prediction, and topology information as well as information about clefts.

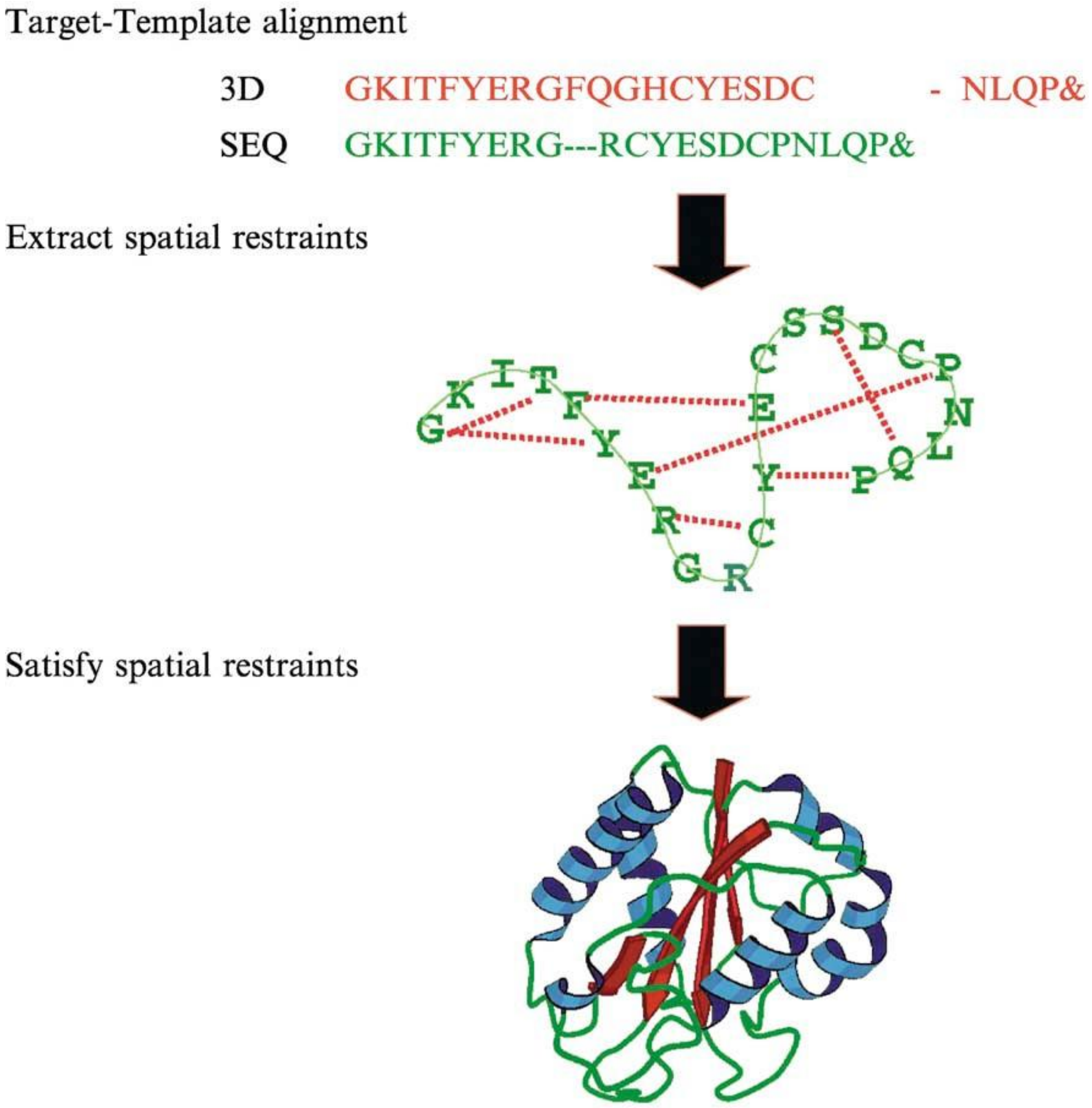
Target-Template alignment
3D GKITFYERGFQGHCYESDC - NLQP&
SEQ GKITFYERG---RCYESDCPNLQP&
Extract spatial restraints
G K I T F Y E R G R C Y E C S S D C P N L Q P
Satisfy spatial restraints

Figure 3.2: Comparative model building programmed by MODELLER