Architectural Frame Work for Homology Modeling in Grid Environment

P.Pradeep Kumar
Research Scholar,
Department of Computer Science,
Bharathiar University,
Coimbatore, Tamilnadu, India.

D.Ramyachitra
Assistant Professor,
Department of Computer Science,
Bharathiar University,
Coimbatore, Tamilnadu, India.

ABSTRACT
This paper shows how homology modeling works for given protein sequence and how it is implemented in grid environment. Homology modeling involves taking a known sequence with an unknown structure and mapping against a known structure of one or several similar protein. The quality of homology modeling is dependent on the quality of the sequence alignment and template structure. Based on this modeling, the unknown protein sequence can be predicted based on the relevance match found from the database. This homology modeling is very useful for molecular docking process. The number of sequences is inputted to the grid environment where homology modeling is designed by FCFS strategy.

Keywords
Grid, Grid Computing, Grid Scheduling, Homology Modeling.

1. INTRODUCTION
Grid computing enables the sharing of hardware and data resources to create a cohesive resource environment for executing distributed applications. Although it has been used within the academic and scientific community for some time, standards, enabling technologies, toolkits, and products are becoming available that allow businesses to use and reap the advantages of Grid computing.

Grid is a hardware and software infrastructure that involves the integrated and collaborative use of resources such as networks, databases and scientific instruments owned and managed by multiple organizations. It also provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities [1]. Grid applications often involve large amounts of data and/or computing resources that require secure resource sharing across organizational boundaries [2]. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. Grids computing offer a way to solve Grand Challenge problems such as protein folding, financial modeling, earthquake simulation, and climate/weather modeling. Grids offer a way of using the information technology resources optimally inside an organization.

The grid infrastructure forms the core foundation for successful grid applications. This infrastructure is a complex combination of a number of capabilities and resources identified for the specific problem and environment being addressed.

Grid scheduling is a process of mapping grid tasks to grid resources over multiple administrative domains. The grid scheduler has four phases, which consists of resource discovery, resource selection, job selection and job execution. The responsibility of a scheduler is selecting resources and scheduling tasks in such a way that the user and application constraints are satisfied, in terms of overall execution time and cost of the resources utilized [3]. In parallel job scheduling on single parallel machines, simple first-come-first-served (FCFS) strategy is often followed.

Homology modeling, also known as compound modeling of protein or knowledge-based modeling, refers to constructing an atomic-resolution model of the target protein from its amino acid sequence. It relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence [4].

Homology modeling is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment [5].

We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

2. HOMOLOGY MODELING
Homology has a precise definition as having a common evolutionary origin [6]. Homology is a qualitative description of the nature of the relationship between two or more things, and it cannot be partial.

Homology modeling is a computational approach for three-dimensional protein structure modeling and prediction. Proteins whose structures are still uncharacterized can be modeled using homology modeling. This method builds an atomic model based on experimentally determined known structures that have sequence homology of more than 40% with the target molecule. Modeling structures with less then 40% template similarity would result in less reliable models and hence ignored. Homology modeling is also known as comparative modeling [7].

The *principle* governing this approach is that if two proteins share a high sequence similarity, they are more likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then this structure can be

superimposed onto the unknown protein with a high degree of confidence. Protein sequences are more conserved than DNA and hence attribute to greater evolutionary significance.

The method of homology modeling is based on the observation that protein tertiary structure is better conserved than amino acid sequence[8].Homology modeling can produce high-quality structural models when the target and template are closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds[9]. The chief inaccuracies in homology modeling, which worsen with lower sequence identity, derive from errors in the initial sequence alignment and from improper template selection [10].

Like other methods of structure prediction, current practice in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

3. HOMOLOGY MODELING STEPS

Homology modeling approaches consists of three steps

1. Finding homologus PDB files.
2. Creation of the alignment, using single or multiple sequence alignments. (if more than one known is involved, sometimes the knowns are aligned together, then the unknown sequence aligned with the group; this helps ensure better domain conservation) Analysis of alignments; gap deletions and additions; secondary structure weighting
3. Structure calculation and model refinement

The sequence alignment and template structure are then used to produce a structural model of the target. The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The modeling process can be subdivided into 9 stages[7]

- template recognition
- alignment
- alignment correction
- backbone generation
- generation of canonical loops (data based)
- side chain generation plus optimization
- ab initio loop building (energy based)
- overall model optimization (energy minimization)

model verification with optional repeat of previous steps

3.1 Step I – Template Selection

Template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures. The search can be performed using a heuristic pair wise alignment search program like BLAST or FASTA. As a rule of thumb, a database protein should have at least 40% sequence identity, high resolution and the most appropriate cofactors for it to be considered as a *template sequence*. The protein sequence whose 3D structure is to be predicted is called the "target sequence".

3.2 Step II – Sequence Alignment

Once the template is identified, the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment. The alignment gives specific alignment scores.

3.3 Step III – Backbone Model Building

Once optimal alignment is achieved the corresponding coordinate's residues from the template proteins can be simply copied onto the target protein. If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms.

If multiple templates selected, then average coordinate values of the templates are used.

3.4 Step IV – Loop Modeling

After the sequence alignment, there are often regions created by insertions and deletions that lead to gaps in alignment. These gaps are modeled by loop modeling, which is less accurate, a major source of error. Currently, two main techniques are used to approach the problem:

- The database searching method - this involves finding loops from known protein structures and superimposing them onto the two stem regions (main chains mostly) of the target protein. Some specialized programs like FREAD and CODA can be used.
- The ab initio method - this generates many random loops and searches for one that has reasonably low energy and φ and ψ angles in the allowable regions in the Ramachandran plot.

3.5 Step V – Side Chain Refinement

After the main chain atoms are built, the positions of side chains must be determined. This is important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface.

A side chain can be built by searching every possible conformation for every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. A rotamer library can also be used, which has all the favorable side chain torsion angles extracted from known protein crystal structures can also be used for this purpose.

3.6 Step VI – Model Refinement and Model Evaluation

This step carries out the energy minimization procedure on the entire model, which adjusts the relative position of the atoms so that the overall conformation of the molecule has the lowest possible energy potential. The goal of energy minimization is to relieve steric collisions without altering the overall structure. In these loop and side chain modeling steps, potential energy calculations are applied to improve the model. Model refinement can also be done by Molecular Dynamic simulation which moves the atoms toward a global minimum by applying various stimulation conditions (heating, cooling, considering water molecules) thus having a better chance at finding the true structure.

3.7 Steps in Model Production

The homology modeling procedure can be broken down into four sequential steps [8]

(1) template selection
(2) target-template alignment
(3) model construction
(4) model assessment

The first two steps are often essentially performed together, as the most common methods of identifying templates rely on the production of sequence alignments; however, these alignments may not be of sufficient quality because database search techniques prioritize speed over alignment quality. The critical first step in homology modeling is the identification of the best template structure, if indeed any are available. The simplest method of template identification relies on serial pair wise sequence alignments aided by database search

techniques such as FASTA and BLAST. Figure 1 shows the flow in homology modeling.

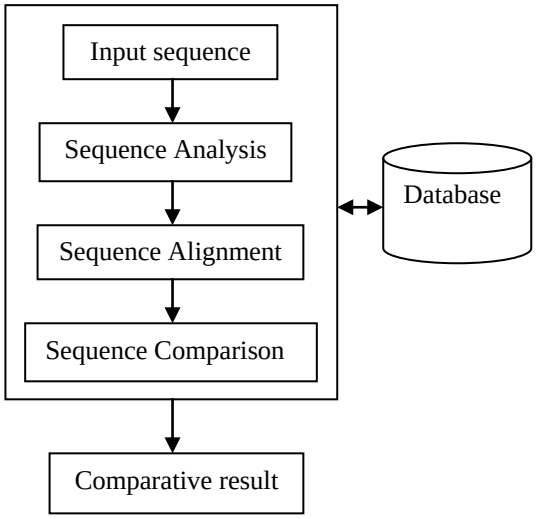

Fig 1: Homology modeling work flow

In homology modeling, once the sequence is inputted, the input sequence is analyzed and then the alignment is done for the sequence. The input sequence is compared with the sequence available in the Protein Data Bank and the comparative result is obtained.

4. TYPES OF HOMOLOGY MODELING

- Basic Modeling - Modeling using a template with very high similarity with the target sequence.
- Advanced Modeling - In this case, the target is modeled using more than one template such that regions of the template proteins that share a high identity with portions of the target are used individually to model these sections.
- Iterative Modeling - This method generates models by using data gathered from previous target-template alignments to generate the probable range of values that can be considered significant for each criterion used in the modeling. This increases the accuracy of the model predicted [7].

5. HOMOLOGY MODELING IN GRID ENVIRONMENT

Three systems are connected in grid environment and homology modeling is implemented in grid environment. When the protein sequence is inputted, through grid environment the sequence is inputted to the system and based on the search in PDB (Protein Databank), the search result will be produced as the percentage of sequence matches. Figure 2 shows how homology model in implemented in the grid environment.

The general flows of creating homology model are as follows

- identify homologous proteins and determine the extent of their sequence similarity with one another and the unknown
- align the sequences
- identify structurally conserved and structurally variable regions
- generate coordinates for core (structurally conserved) residues of the unknown structure from those of the known structure(s)
- generate conformations for the loops (structurally variable) in the unknown structure
- build the side-chain conformations
- refine and evaluate the unknown

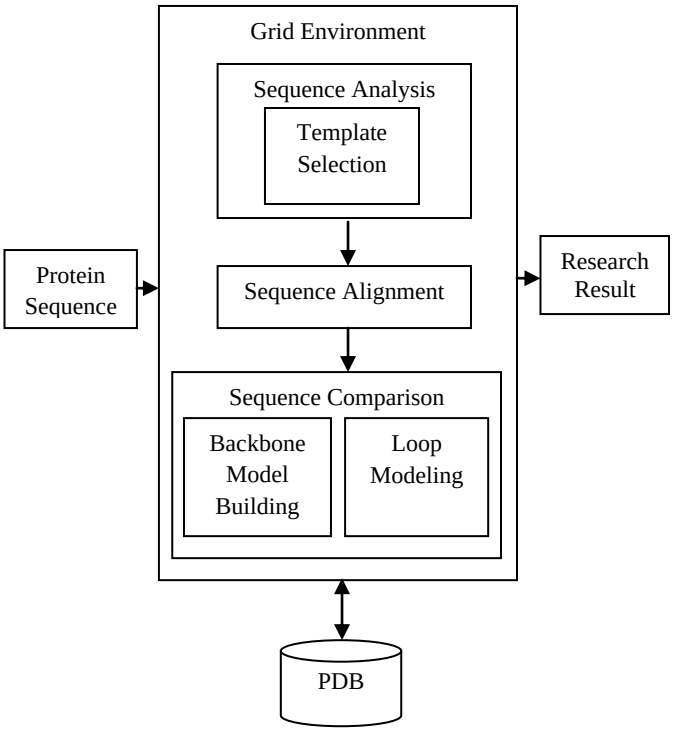

Fig 2: Architectural Framework for Homology Modeling in Grid

5.1 Evaluation and Refinement of the Structure

For a homology model from any source, it is important to demonstrate that the structural features of the model are reasonable in terms of what is know about protein structures in general. That is, researchers have analyzed three-dimensional structures of proteins from which basic principles of protein structure and folding have been developed. Several programs are available to assist in this analysis of correctness of a homology model [11, 12].

The criteria for analysis of correctness can include

- main chain conformations in acceptable regions of the Ramachandran map
- planar peptide bonds
- side chain conformations that correspond to those in the rotamer library
- hydrogen-bonding of polar atoms if they are buried
- proper environments for hydrophobic and hydrophilic residues
- no bad atom-atom contacts
- no holes inside the structure.

6. FUTURE WORK AND CONCLUSION

To improve the efficiency of the target protein, the new algorithm is designed as based on the systems in grid and used instead of FCFS strategy, so that each input sequence is allocated to the respective system to get a target protein in short time.

The unknown structure predicted from the known structure, is verified using the Ramachandran plot. Homology modeling is used in microbiological research and also in bioinformatics research process. This homology modeling is very useful for molecular docking process. It can find the location of alpha carbons of key residues inside the folded protein. It can help to guide the mutagenesis experiments, or hypothesize structure-function relationships. The positions of conserved regions of the protein surface can help identify putative active sites, binding pockets and ligands.

The protein sequence search in a single system will result in time delay and the user has to wait for long time to find each sequence result. To overcome this time delay and waiting time of a user, the homology modeling is implemented in grid environment, where more than one sequence can be inputted to the system in the grid environment and search in PDB is done, so that more than one protein sequence will result at a time.

Homology models are unable to predict conformations of insertions or deletions, or side chain positions with a high level of accuracy. Homology models are not useful in modeling and ligand docking studies necessary for the drug designing and development process. However, it may be helpful for the same, if the sequence identity with the template is greater than 70%[7]Our thanks to the experts who have contributed towards development of the template.

7. REFERENCES

[1] Ian Foster, What is the Grid? A Three Point Checklist, Argonne National Laboratory & University of Chicago, July 20, 2002.

[2] L.J. zhang, J.Y.Chung, and Q. Zhou, Developing grid computing applications, part 1: Introduction of a grid architecture and toolkit for building grid solutions, IBM Corporation New York, October 2002.

[3] Maozhen Li, Mark Baker, The Grid Core Technologies, A John Wiley & Sons, Inc., 2005.

[4] Chothia C and Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*5:823–6.

[5] Zhang Y and Skolnick J (2005). "The protein structure prediction problem could be solved using the current PDB library". *Proc Natl Acad Sci USA* 102 (4):1029–34. doi:10.1073/pnas.0407152101. PMC 545829. PMID 15653774.

[6] Reeck, G.R. et al. (1987) "Homology" in Proteins and Nucleic Acids: A Terminology Muddle and a Way out of It. Cell 50: 667.

[7] http://iitb.vlab.co.in/?sub=41&brch=118&sim=657&cnt=1.

[8] Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29: 291–325.

[9] Williamson AR. (2000). Creating a structural genomics consortium. *Nat Struct Biol* 7 S1(11s):953.

[10] Venclovas C, Margelevičius M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins* 61(S7):99–105.

[11] Luthy, R., Bowie, J.U., and Eisenberg, D. (1992) Assessment of Protein Models with Three-Dimensional Profiles. Nature 356: 83-85.

[12] Bowie, J.U., Luthy, R., and Eisenberg, D. (1991) A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. Science 253: 164-170.

AUTHOR'S PROFILE

Ms. D.Ramyachitra has completed MCA from Madras University in 2000 and M.Phil in Computer Science from Bharathiar University in 2004. She is currently pursuing Ph.D in Computer Science in Bharathiar University. She has published papers in 10 national and international Conferences and in 7 national and international journals. Her area of interest includes grid computing and image processing.

Mr. P.Pradeep Kumar has completed M.Sc in Computer Technology from Kongu Engineering College. He is currently pursuing his M.Phil in Computer Science in the School of Computer Science and Engineering Bharathiar University, Coimbatore. His has published papers in 5 national and international conferences. His are of interest includes grid computing and data mining.