THEORY

Homology modeling is the computational approaches for protein three-dimensional structure modeling and prediction. Homology modeling builds an atomic model based on experimentally determined known structures that have sequence homology of more than 40%. It is also known as comparative modeling.

The principle behind it is that if two proteins share a high sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.

Homology Modeling is moderately accurate for the positions of alpha carbons and inaccurate for side chain positions and loops. The others approaches are threading for <40% similarity and ab initio prediction for no homolog.

Types of Modeling:

- Basic Modeling - Model a sequence with high identity to a template.
- Advanced Modeling - Model based on multiple templates and bound to a ligand.
- Iterative Modeling - Increase the accuracy of the modeling exercise by iterating search.

The overall homology modeling procedure consists of six steps-

**Step I - Template Selection**

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

**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.

**Step III - Backbone Model Building**

Once optimal alignment is achieved, the corresponding coordinates residues of the template proteins selected 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.

**Step IV – Loop Modeling**

After the sequence alignment, there are often regions caused by insertions and deletions leads to gaps in sequence alignment. The gaps are modeled by loop modeling, which is a very problem and is also a major source of error. Currently, there are two main techniques used to approach the problem:

- The database searching method - this involves finding loops from known protein structures and aligning 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 the one that has reasonably low energy and φ and ψ angles in the allowable regions in the Ramachandran plot.
Step V - Side Chain Refinement

After the main chain atoms are built, the positions of side chains should 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 by 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.

Step VI - Model Refinement and Model Evaluation

This step includes the energy minimization procedure on the entire model, which moves the atoms in such a way that the overall conformation has the lowest energy potential. The goal of energy minimization is to relieve steric collisions without altering the overall structure. In these loop modeling 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 atoms toward a global minimum by applying various stimulation conditions (heating, cooling, considering water molecules) and has a better chance at finding the true structure.

The final model has to be evaluated for checking the φ–ψ angles, chirality, bond lengths, close contacts and also the stereochemical properties. Various online protein validation software packages are available such as Procheck, WHATIF, ANOLEA, Verify3D, PROSA.

- Successful model depends on Template selection, algorithm used and the validation of the model.

- Various Comprehensive Modeling Programs are available like Modeller, SWISS MODEL, Schrodinger, 3D-JIGSAW.
ADVANTAGES -

- To be able to find the location of alpha carbons of key residues inside the folded protein.
- They help in guiding the mutagenesis experiments, or hypothesize structure-function relationships.
- The software builds the model containing all non-hydrogen atoms.
- The positions of conserved regions of the protein surface can help to identify putative active sites, binding pockets and ligands.

DISADVANTAGES -

- Homology models are unable to predict conformations of insertions or deletions, or side chain positions.

- Homology models will not be useful in modeling and ligand docking necessary for the drug design and development process. However, it may be helpful for the same if the sequence identity with the template is >70%.