Comparative Protein Structure Modeling of Genes and Genomes
Marc A. Martí-Renom, Ashley C. Stuart, András Fiser, Roberto Sánchez, Francisco Melo, and Andrej Šali

Presented by Barry Sexton, Amanda Caster, Stephanie Lee
Bioinformatics and Bioengineering Summer Institute 2005
University of Pittsburgh, Carnegie Mellon University, Duquesne University, Pittsburgh Supercomputing Center

Comparative Modeling Introduction
Also known as Homology Protein Structure Modeling
Three dimensional structure prediction of a given protein sequence (target) based on its alignment to one or more proteins of known structure (templates)
Two conditions must be met in order to construct a useful model:
1) Must be "detectable" similarity between target sequence and template structure
2) A correct alignment between the target and template must be calculated (CRUCIAL)
Models based on homologous alignments are justified because 3D protein structure is more conserved than amino acid sequence

Comparative Modeling Utility
Technique is becoming increasingly applicable and accurate because:
1) Rapid increase in the number of known protein structures
2) Recent improvements in modeling software
Estimated that approximately 1/3 of all known protein sequences can be modeled with useful accuracy using comparative modeling
100,000 out of 500,000 proteins with known sequences
This is because at least 1/3 of all sequences are recognizably related to at least one known structure
The number of unique folds proteins can adapt is limited
Estimated that in less than 5 years, one example of most structural folds will be known, making comparative modeling applicable to most protein sequences

Steps of Comparative Modeling
1) Fold Assignment & Template Selection
2) Target-Template Alignment
3) Model Construction
4) Model Evaluation
Fold Assignment & Template Selection
- Identify all protein structures related to the target sequence to be used as templates

- GenBank
  - Run by National Center for Biotechnology Information (NCBI)
  - Contains over 23 million sequences and continues to double every 10 months (Release 134)

- Basic Local Alignment Search Tool (BLAST)
  - Search algorithm emphasizes speed over sensitivity

BLAST

Fold Assignment & Template Selection
- Structural Classification of Proteins (SCOP)
  - Aims to provide structural & evolutionary relationships between proteins with known structure
  - Broad survey of protein folds and detailed information about protein relatedness

- Protein Data Bank (PDB)
  - Repository for 3D structural data of proteins and nucleic acids

[Image of GenBank and BLAST interfaces]
Factors to Consider

- **The family or subfamily of proteins**: Phylogenetic trees aid in selecting a template(s) which are most closely related to the target.

- **Template environment should be similar to the desired environment for the target**: Solvent, pH, Ligands, etc.

- **Quality of the template structure**: Resolution (~2.5 Å), Experimentally determined R-factor.

Target-Template Alignment

- Most template search tools automatically produce a template-template alignment.

- However, search tools are geared more towards detecting remote relationships than constructing optimal alignments.

- Other alignment tools such as BLAST, CLUSTAL, or GeneDoc should be used to manually construct optimal alignments.

More Information:

- CLUSTAL: [http://www.ebi.ac.uk/clustalw/newwindow.html](http://www.ebi.ac.uk/clustalw/newwindow.html)
- GeneDoc: [http://www.psc.edu/biomed/genedoc/](http://www.psc.edu/biomed/genedoc/)

When sequence identity is above 40% the alignment is almost always correct.

The alignment becomes much more difficult in the “twilight zone” of less than 30% sequence identity.

Maximal effort is needed to obtain the most accurate alignment because no comparative modeling method can recover from an incorrect alignment.

Quite often multiple sequence alignments are conducted in which many homologous sequences are aligned to obtain more reliable results.

In more difficult cases it is beneficial to include structural information when determining models (results in a more accurate alignment):

- Structural information helps avoid gaps in secondary structure elements, in buried regions, or between two distant residues.
Model Building

Three Major Types

1) Modeling by rigid-body assembly
2) Modeling by segment matching
3) Modeling by satisfaction of spatial restraints

Accuracies of the three methods are relatively similar.
Modeling methods must allow flexibility, automation, and easy model recalculation based on changes made to the alignment.

Rigid-Body Assembly

Builds the model based on a small number of rigid bodies obtained from conserved regions on aligned structures.

1) Select template structures and superpose
2) Calculate framework based on the average coordinates of Cα atoms in structurally conserved regions

Superimposition of Homologous Proteins

1) Select template structures and superpose
2) Calculate framework based on the average coordinates of Cα atoms in structurally conserved regions
3) Superpose the template most similar in sequence to the target on the framework to generate main chain atoms of the target
4) Generate loops based on a database search which maximizes compatibility of variable regions (loops) with the anchor core regions
5) Model side-chains based on their conformational preferences and on the conformation of equivalent side-chains in the templates
6) Improve stereochemistry of model based on energy minimization or a molecular dynamics refinement
Segment Matching

- Based on the finding that most hexapeptide sequences can be clustered into about 100 structural classes
- A subset of atomic positions from templates are used as ‘guiding positions’
  - Cα backbone of template(s) usually taken as guiding positions
- All-atom segments which fit the guiding positions can be constructed by:
  - searching databases of known protein structures or
  - a conformational search restrained by an energy function

Satisfaction of Spatial Restraints

- Assumes distances and angles between aligned residues in the template and target are similar
- Based on stereochemical restraints on:
  - bond lengths
  - bond angles
  - dihedral angles
  - non-bonded atom-atom contacts obtained from a molecular mechanics force field
- Model derived by minimizing the violation of these restraints
- Considered the strongest modeling method because:
  - Many types of information from the target can be used
  - Restraints derived from many sources can easily be added to homology-driven restraints

Loop Modeling

- Loops defined as regions which connect secondary structure elements in a fold
- No generally reliable method currently exists for modeling loops longer than 5 residues
- Luckily, most loops are fairly short

Importance of Loop Modeling

- In a given protein family, structural variability results from substitutions, insertions, and deletions
- These changes occur most often in loop regions because they are more exposed
- Thus, loops often determine the functional specificity of a protein and usually possess most active and binding sites
- Therefore, loop modeling accuracy is a major determining factor of model usefulness
  - Especially when experimenting with ligand binding and docking

Myoglobin (loop examples)
Methods of Loop Modeling

- Loop conformation affected not only by its own sequence but also largely by the core regions which surround it.

**Ab Initio Loop Prediction**
- Based on a search of known conformations in a given environment, governed by an energy function.
- Search algorithms include global & local energy minimizations, MD simulations, and Monte Carlo analysis.

**Database Approach to Loop Prediction**
- Searches for main chain segments which match the two stem regions surrounding the target loop.
- Many matching segments are obtained and sorted according to geometric criteria or similarity between template and target loop sequences.

Sidechain Modeling

- Similar to loop prediction in that side chain conformations are predicted from similar sequences/structures and from steric or energetic considerations.
- Most difficult aspect of modeling because many exposed sidechains are highly flexible and are without a single dominant conformation.
- A template backbone with less than 30% sequence identity to the target is insufficient to produce the correct packing of sidechains.

For more information on sidechain modeling techniques:

Errors in Comparative Models

- **Distortions and Shifts in Correctly Aligned Regions**
- Main chain conformation can change even if overall fold remains the same.
- Even with correctly aligned sequences, model can have local differences from template.
- Often due to structure determination in different environments (e.g. crystal).

- **Side Chain Packing**
- As sequences diverge the packing of sidechains changes.
- Especially critical if they occur in functionally significant regions.

Errors in Comparative Modeling

- **Regions without a template**
- i.e. insertions or loops.
- Very difficult to model.
- Current techniques can accurately predict backbone conformation of an insertion of less than 9 residues.

- **Misalignments**
- Largest source of error.
- Significant when target-template sequence identity is less than 30%.
- Multiple alignments typically more accurate than pairwise alignments.

For more information on errors in comparative modeling techniques:
Errors in Comparative Modeling

- Incorrect Templates
  - Only a problem when distantly related proteins are used as templates (<25% seq. identity)
  - In general conservation of key functional or structural residues in the target increases the likeliness of a correct model

Model Evaluation

- Models formulated before the experimental structures were available
- % Structure Overlap is defined as fraction of equivalent residues
- Residues are considered equivalent when their Cα atoms are within 3.5 Å of each other

Applications of Comparative Modeling

- Increasingly efficient way to obtain useful information about proteins
- Examples
  - Designing mutants to test hypotheses about a protein’s function
  - Identifying active and binding sites
  - Identifying, designing, and improving ligands for a given binding site
  - Modeling substrate specificity
  - Simulating protein-protein docking
  - Testing remote structural relationships
  - Many others...

However, the types of biological question that a given model can address depend on its accuracy...