Homology Modeling

• Presentation
• Fold recognition
• Model building
  – Loop building
  – Sidechain modeling
  – Refinement
• Testing methods: the CASP experiment
Homology Modeling

• Presentation

Why do we need homology modeling?

To be compared with:

PDB Current Holdings Breakdown

<table>
<thead>
<tr>
<th>Exp. Method</th>
<th>Molecule Type</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/NA Complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray</td>
<td></td>
<td>27430</td>
<td>808</td>
<td>1272</td>
<td>85</td>
<td>29995</td>
</tr>
<tr>
<td>NMR</td>
<td></td>
<td>4430</td>
<td>674</td>
<td>148</td>
<td>17</td>
<td>5238</td>
</tr>
<tr>
<td>Electron Microscopy</td>
<td></td>
<td>79</td>
<td>5</td>
<td>27</td>
<td>0</td>
<td>115</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>72</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>32009</td>
<td>1485</td>
<td>1420</td>
<td>162</td>
<td>35826</td>
</tr>
</tbody>
</table>
Structural Genomics project

- Aim to solve the structure of all proteins: this is too much work experimentally!

- Solve enough structures so that the remaining structures can be inferred from those experimental structures

- The number of experimental structures needed depend on our abilities to generate a model.
Global versus Local

Global alignment

Global versus Local (2)

motif

Local alignment
Measuring protein structure similarity

Given two “shapes” or structures A and B, we are interested in defining a distance, or similarity measure between A and B.

• Visual comparison
• Dihedral angle comparison
• Distance matrix
• RMSD (root mean square distance)

Is the resulting distance (similarity measure) D a metric?

\[ D(A,B) \leq D(A,C) + D(C,B) \]

Comparing dihedral angles

Torsion angles \((\phi, \psi)\) are:
- local by nature
- invariant upon rotation and translation of the molecule
- compact (O(n) angles for a protein of n residues)

But...

Add 1 degree To all \(\phi, \psi\)
Distance matrix

• Advantages
  - invariant with respect to rotation and translation
  - can be used to compare proteins

• Disadvantages
  - the distance matrix is $O(n^2)$ for a protein with $n$ residues
  - comparing distance matrix is a hard problem
  - insensitive to chirality
Root Mean Square Distance (RMSD)

To compare two sets of points (atoms) $A=\{a_1, a_2, \ldots, a_N\}$ and $B=\{b_1, b_2, \ldots, b_N\}$:

- Define a 1-to-1 correspondence between $A$ and $B$
  
  for example, $a_i$ corresponds to $b_i$, for all $i$ in $[1,N]$

- Compute RMS as:

$$RMS(A,B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d(a_i, b_i)^2}$$

$d(A_i, B_i)$ is the Euclidian distance between $a_i$ and $b_i$.

Homology Modeling: why it works

High sequence identity

High structure similarity
From a computational point of view:

'Protein folding problem'

Sequence ----------------------------> Structure

Function

A fundamental paradigm of protein science:
The amino acid sequence encodes the structure; the structure determines the function.

http://www.csb.yale.edu/toc.html

FASTA
Sequence Comparison & Searching

CLUSTALW
Sequence Comparison & Searching

GenBank
NIH genetic sequence database

BLAST

Psi_BLAST
BLAST (Basic Local Alignment Search Tool) is a popular program for searching biosequences against databases. BLAST was developed and is maintained by a group at the National Center for Biotechnology Information (NCBI). Salient characteristics of BLAST are:

**Local alignments**
BLAST tries to find patches of regional similarity, rather than trying to find the best alignment between your entire query and an entire database sequence.

**Ungapped alignments**
Alignments generated with BLAST do not contain gaps. BLAST’s speed and statistical model depend on this, but in theory it reduces sensitivity. However, BLAST will report multiple local alignments between your query and a database sequence.

**Explicit statistical theory**
BLAST is based on an explicit statistical theory developed by Samuel Karlin and Steven Altschul (PNAS 87:2284-2268, 1990). The original theory was later extended to cover multiple weak matches between query and database entry (PNAS 90:5873, 1993).

**CAUTION**: the repetitive nature of many biological sequences (particularly naive translations of DNA/RNA) violates assumptions made in the Karlin & Altschul theory. While the P values provided by BLAST are a good rule-of-thumb for initial identification of promising matches, care should be taken to ensure that matches are not due simply to biased amino acid composition.

**Rapid**
BLAST is extremely fast. You can either run the program locally or send queries to an E-mail server maintained by NCBI: blast@ncbi.nlm.nih.gov

**Heuristic**
BLAST is not guaranteed to find the best alignment between your query and the database; it may miss matches. This is because it uses a strategy which is expected to find most matches, but sacrifices complete sensitivity in order to gain speed. However, in practice few biologically significant matches are missed by BLAST which can be found with other sequence search programs. BLAST searches the database in two phases. First it looks for short subsequences which are likely to produce significant matches, and then it tries to extend these subsequences.
BLAST

• **Raw Score**
  The score of an alignment, $S$, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (see PAM, BLOSUM). Gap scores are typically calculated as the sum of $G$, the gap opening penalty and $L$, the gap extension penalty. For a gap of length $n$, the gap cost would be $G+Ln$. The choice of gap costs, $G$ and $L$ is empirical, but it is customary to choose a high value for $G$ and a low value for $L$.

  $E$ value
  Expectation value. The number of different alignments with scores equivalent to or better than $S$ that are expected to occur in a database search by chance. The lower the $E$ value, the more significant the score.

  \[ E = Kmn e^{-\lambda S} \]

PSI-BLAST

• **Position specific iterative BLAST (PSI-BLAST)** refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each “iteration” used to refine the profile. This iterative searching strategy results in increased sensitivity.
FASTA
Salient characteristics of FASTA are:

Local alignments
FASTA tries to find patches of regional similarity, rather than trying to find the best alignment between your entire query and an entire database sequence.

Gapped alignments
Alignments generated with FASTA can contain gaps.

Rapid
FASTA is quite fast. You can either run the program locally or send queries to an E-mail server

Heuristic
FASTA is not guaranteed to find the best alignment between your query and the database; it may miss matches. This is because it uses a strategy which is expected to find most matches, but sacrifices complete sensitivity in order to gain speed.

• In general, for a database, which contains many unrelated as well as related proteins, two sequences might be considered related in evolutionary terms (i.e. diverged from a common ancestor and share common three-dimensional structure), when the $E$-value of the FASTA query is less than 0.02, Pearson, 1996.
Homology Modeling: How it works

- Find template
- Align target sequence with template
- Generate model:
  - add loops
  - add sidechains
- Refine model
Homology Modeling

- Fold recognition

Fold Recognition

Homology modeling refers to the easy case when the template structure can be identified using BLAST alone.

What to do when BLAST fails to identify a template?

- Use more sophisticated sequence methods
  - Profile-based BLAST: PSIBLAST
  - Hidden Markov Models (HMM)

- Use secondary structure prediction to guide the selection of a template, or to validate a template

- Use threading programs: sequence-structure alignments
Fold Recognition

Blast for PDB search
Full homology modeling packages
Profile based approach
HMM
Structure-derived profiles
Fold recognition and secondary structure prediction

Homology Modeling

• Model building
  – Loop building
Very short loops: Analytic Approach

Medium loops: A database approach

\[ C_A + C_{1} \cos(\Omega_{1}) + C_{2} \cos(\xi_{2}) + \\
C_{3} \cos(\Omega_{3}) \cos(\phi_{3}) + C_{4} \sin(\Omega_{4}) \sin(\xi_{4}) = 0 \]

Scan database and search protein fragments with correct number of residues and correct end-to-end distances
Medium loops: A database approach

Long loops: A fragment-based approach

1) Clustering Protein Fragments to Extract a Small Set of Representatives (a Library)
Generating Loops

Fragment library

Generating Loops

Fragment library
Generating Loops

Fragment library

Generating Loops

Fragment library
Generating Loops

Fragment library

Long loops: A fragment-based approach

Test cases: 20 loops for each loop length
Methods: database search, and fragment building, with fragment libraries of size L
Loop building: Other methods

Heuristic sampling (Monte Carlo, simulated annealing)
Inverse kinematics
Relaxation techniques
Systematic sampling


Homology Modeling

- Model building
  - Sidechain modeling
Self-Consistent Mean-Field Sampling

\[ P(i,1) + P(i,2) + P(i,3) = 1 \]
Self-Consistent Mean-Field Sampling

**Multicopy Protein**

\[ E(i,k) = U(i,k) + U(i,k,\text{Backbone}) + \sum_{j-l \neq i} \sum_{l} P(j,l) U(i,k,j,l) \]
Self-Consistent Mean-Field Sampling

\[ E(i,k) = U(i,k) + U(i,k,Backbone) + \sum_{j=1}^{Nrot(i,2)} \sum_{l=1}^{Nrot(j,2,i,1)} P(j,l)U(i,k,j,l) \]

\[ P_{new}(i,k) = \frac{\exp\left[\frac{-E(i,k)}{RT}\right]}{\sum_{l=1}^{Nrot(j,2,i,1)} \exp\left[\frac{-E(i,l)}{RT}\right]} \]

Dead End Elimination (DEE) Theorem

- There is a global minimum energy conformation (GMEC) for which there is a unique rotamer for each residue.
- The energy of the system must be pairwise.

Each residue has a set of possible rotamers. The notation means residue i has the conformation described by rotamer r.

The energy of any conformation C of the protein is given by:

\[ E_{Tot}(C) = E_{Template} + \sum_{i=1}^{N} E(i_r, Template) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E(i_r, j_s) \]

Note that:

\[ E_{Tot}(C) \geq E_{Tot}(GMEC) \quad \forall \ C \]

Dead End Elimination (DEE) Theorem

Consider two rotamers, \( i_r \) and \( j_s \) at residue i and the set of all other rotamer conformations \( \{S\} \) at all residues excluding i.

If the pairwise energy between \( i_r \) and \( j_s \) is higher than the pairwise energy between \( i_r \) and \( j_s \) for all \( j_s \) in \( \{S\} \), then \( i_r \) cannot exist in the GMEC and is eliminated. Mathematically:

\[ E(i_r, Template) + \sum_{j \neq i} E(i_r, j_s) > E(i_r, Template) + \sum_{j \neq i} E(i_r, j_s) \quad \forall \{S\} \]

then

\( i_r \) does not belong to the GMEC
Dead End Elimination (DEE) Theorem

This is impractical as it requires \( S \). It can be simplified to:

\[
E(i, \text{Template}) + \sum_{j \neq i} \min_s E(i, j_s) > E(i, \text{Template}) + \sum_{j \neq i} \max_s E(i, j_s)
\]

then

\( i \), does not belong to the GMEC

Iteratively eliminate high energy rotamers: proved to converge to GMEC


Other methods for side-chain modeling

- Heuristics (Monte Carlo, Simulated Annealing)
  SCWRL (Dunbrack)

- Pruning techniques

- Mean field methods
Loop building + Sidechain Modeling: generalized SCMF

- Template
- Add multi-copies of candidate “loops”
- Add multi-copies of candidate side-chains
- Final model


---

Homology Modeling

- Model building
  - Refinement
ROSETTA - the most successful approach to \textit{ab initio} prediction

- David Baker, U. Washington, Seattle
- Based on the idea that the possible conformations of any short peptide fragment (3-9 residues) are well-represented by the structures it is observed to adopt in the pdb
- Generate a library of different possible structures for each sequence segment
- Search the possible combinations of these for ones that are protein-like by various criteria

ROSETTA fragment libraries

- Remove all homologs of the protein to be modeled (>25% sequence identity)
- For each 9 residue segment in the target, use sequence similarity and secondary structure similarity (compare predicted secondary structure for target to fragment secondary structure) to select ~25 fragments
- Because secondary structure is influenced by tertiary structure, ensure that the fragments span different secondary structures
- The extent to which the fragments cluster around a consensus structure is correlated with how good a model the fragment is likely to be for the target

\text{LSERTVARS}
Refinement?

CASP5 assessors, homology modeling category:

“We are forced to draw the disappointing conclusion that, similarly to what observed in previous editions of the experiment, no model resulted to be closer to the target structure than the template to any significant extent.”

The consensus is not to refine the model, as refinement usually pulls the model away from the native structure!!
Homology Modeling

- Testing methods: the CASP experiment

The CASP experiment

- CASP= Critical Assessment of Structure Prediction

- Started in 1994, based on an idea from John Moult (Moult, Pederson, Judson, Fidelis, Proteins, 23:2-5 (1995))

- First run in 1994; now runs regularly every second year (CASP6 was held in 2004, CAS7 will be in 2006)
The CASP experiment: how it works

1) Sequences of target proteins are made available to CASP participants in June-July of a CASP year - the structure of the target protein is known, but not yet released in the PDB, or even accessible.

2) CASP participants have between 2 weeks and 2 months over the summer of a CASP year to generate up to 5 models for each of the targets they are interested in.

3) Model structures are assessed against experimental structure.

4) CASP participants meet in December to discuss results.

CASP Statistics

<table>
<thead>
<tr>
<th>Experiment</th>
<th># of Targets</th>
<th># of predictors</th>
<th># of 3D models</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP1</td>
<td>33</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td>CASP2</td>
<td>42</td>
<td>72</td>
<td>947</td>
</tr>
<tr>
<td>CASP3</td>
<td>43</td>
<td>61</td>
<td>1256</td>
</tr>
<tr>
<td>CASP4</td>
<td>43</td>
<td>111</td>
<td>5150</td>
</tr>
<tr>
<td>CASP5</td>
<td>67</td>
<td>175</td>
<td>22909</td>
</tr>
<tr>
<td>CASP6</td>
<td>87</td>
<td>166</td>
<td>28965</td>
</tr>
</tbody>
</table>
CASP

Three categories at CASP

- Homology (or comparative) modeling
- Fold recognition
- Ab initio prediction

CASP dynamics:

- Real deadlines; pressure: positive, or negative?
- Competition?
- Influence on science?


CASP: quality of alignment
Homology Modeling: Practical guide

**Approach 1: Manual**

- Submit target sequence to BLAST; identify potential templates
- For each template:
  - Generate alignment between target and template (Smith-Waterman + manual correction)
  - Build framework
  - build loop + sidechain
  - assess model (stereochemistry, …)
Homology Modeling: Practical guide

**Approach 2:** Submit target sequence to automatic servers

- **Fully automatic:**
  - 3D-Jigsaw: [http://www.bmm.icnet.uk/servers/3djigsaw/](http://www.bmm.icnet.uk/servers/3djigsaw/)

- **Fold recognition:**
  - PHYRE: [http://www.sbg.bio.ic.ac.uk/~phyre/](http://www.sbg.bio.ic.ac.uk/~phyre/)

- **Useful sites:**
  - Meta server: [http://bioinfo.pl/Meta](http://bioinfo.pl/Meta)

- **Homology Modeling Software/Servers**
  - CASP: Protein Structure Prediction Center, Lawrence Livermore National Laboratory, CA.
  - Swiss-Model Server: Free
  - Modeller: Free for Academics Non-profit institutions
  - CIB Server
  - Homology (Accelrys Inc.) Automatic Homology modeling module. The software suite also has Modeller, SeqFold modules, Quanta
  - Wloop: The Loop Homology Modeling Server
  - GeneMine: Formerly known as Xlook
  - MatchMaker: Tripos Sybyl's MatchMaker. Also has Composer and GeneFold software.
  - UCLA/DOE Server: UCLADOE Fold Server
  - Predict Protein Server: EMBL, Predict Protein Server
  - Abagyan Lab Server: Scripps Research Institute
  - NBCR Server: National Biomedical Computation Resource, University of California, San Diego
  - 3D-Jigsaw: Comparative Modelling Server UK Site. Click on submission to submit the sequence.

- **Fold**

- **Visualizer**
  - SPDBV: Swiss-PDB Viewer
  - Gn3D: See in 3-D
  - Rasmol: Rasmol
  - Rasmol: Rasmol Animations

- **Computational Chemistry/Biology Software**
  - Sybyl: Tripos (Flexi-Dock, QSAR, Gen-Fold etc.)
  - InsightII: Accelrys (GCG, InsightII, Homology, Modeller etc.)
  - AutoDock: Automated Docking of Flexible Ligands to Macromolecules (Scripps Research Institute)
  - Dock Small Molecule Receptor docking (UCSF)
  - GROMACS: The Fastest Molecular Dynamics Program.
  - VMD: Visual Molecular Dynamics software (NIH)

  * [http://ncisgi.ncifcrf.gov/~ravichas/HomMod/](http://ncisgi.ncisgi.ncifcrf.gov/~ravichas/HomMod/)