Protein Structure Determination

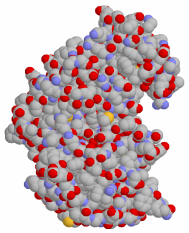
Thomas Funkhouser
Princeton University
CS597A, Fall 2005

Introduction

Goal:

- Given a protein sequence, determine its 3D structure

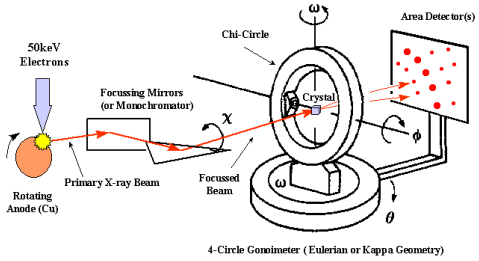
```
  1 MIKLGIVMDP IANINIKKDS SFAMLLEAQR RGYELHYMEM GDLYLINGEA
 51 RAHTRTLNVK QNYEEWFSFV GEQQLPLADL DVILMRKDPP FDTEFIYATY
101 ILERAEEKGT LIVNKPQSLR DCNEKLFTAW FSDLTPETLV TRNKAQLKAF
151 WEKHSDIILK PLDGMGGASI FRVKEGDPNL GVIAETLTEH GTRYCMAQNY
201 LPAIKDGDKR VLVVDGEPVP YCLARIPQGG ETRGNLAAGG RGEPRPLTES
251 DWKIARQIGP TLKEKGLIFV GLDIIGDRLT EINVTSPTCI REIEAEFPVS
301 ITGMLMDAIE ARLQQQ
```



1gsa

Experimental Methods

X-ray crystallography



4-Circle Goniometer (Eulerian or Kappa Geometry)

http://ruppweb.dyndns.org/

Experimental Methods

X-ray crystallography

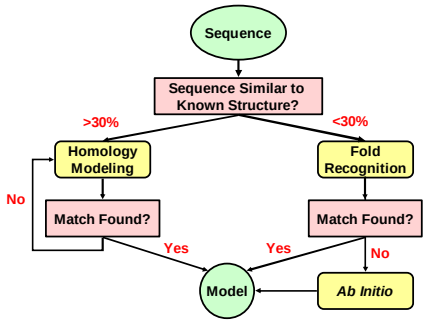
- Beams of x-rays are passed through a crystal of protein. Atoms in the protein crystal scatter the x-rays, which produce a diffraction pattern on a photographic film
- Protein must be crystallizable

NMR spectroscopy

- A solution of protein is placed in a magnetic field and the effects of different radio frequencies on the resonance of different atoms in a protein are measured
- Protein must be small (~120 residues)
- Protein must be soluble

Both methods are expensive, slow, and cannot be applied for all proteins

Computational Methods



http://www.cs.cmu.edu/~cjl/teaching/15872AS05/

Outline

Introduction

Tertiary structure prediction

- Ab initio methods
- Homology modeling
- Fold recognition

Results

Discussion

Outline

Introduction

Tertiary structure prediction

- Ø Ab initio methods
- Homology modeling
- Fold recognition

Results

Discussion

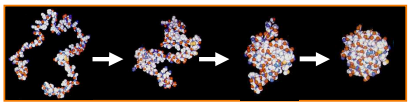
Ab Initio Methods

Goal:

- Predict tertiary structure from first principles

Motivation:

- Thermodynamic hypothesis predicts that the native conformation of a protein corresponds to a global free energy minimum of the protein/solvent system [Afinsen73]



http://www.cs.cmu.edu/~cjl/teaching/15872AS05/

Ab Initio Methods

Approach:

- Find protein conformations minimizing global free energy

Challenges:

- Must search large space of possible conformations
  - § Backbone
  - § Side-chains
- Must be able (at least) to recognize conformations with lowest global free energy
  - § Solvation effects

Ab Initio Methods

Search procedures

- Molecular dynamics
- Simulated annealing
- Genetic algorithms

Scoring functions

- Molecular mechanics
- Empirical functions
- Knowledge-based functions

This is like protein-protein docking

Ab Initio Methods

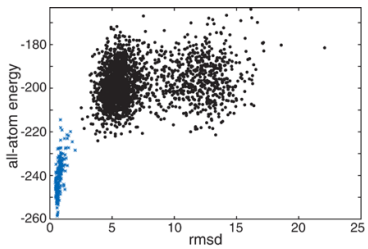
General strategies

- Predict secondary structures first
- Predict coarse representation first (coarse-to-fine)
- Assemble structural fragments extracted from other proteins with similar local sequences

Issues

- Large search space
- Insufficient scoring functions

Ab Initio Methods

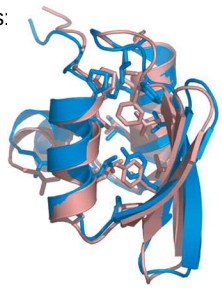


Free-energy landscape for the small protein barstar (PDB code 1a19 [PDB]). Rosetta all-atom energy (y axis) is plotted against C-RMSD (x axis) for models generated by simulations starting from the native structure (refined natives, blue points) or from an extended chain (de novo models, black points) [Baker et al.].

http://www.sciencemag.org/cgi/content/full/309/5742/1868/FIG3

Ab Initio Methods

Some good results:
- Baker et al.

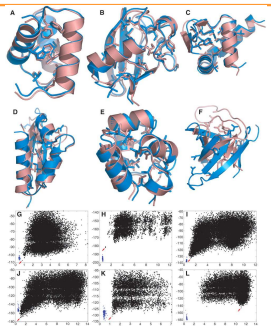


1.6 Å C -RMSD blind structure prediction for CASP6 target T0281

http://www.sciencemag.org/cgi/content/full/309/5742/1868/FIG3

Ab Initio Methods

Some good results
- Baker et al.



http://www.sciencemag.org/cgi/content/full/309/5742/1868/FIG3

Outline

Introduction

Tertiary structure prediction
- Ab initio methods
- Homology modeling
- Fold recognition

Results

Discussion

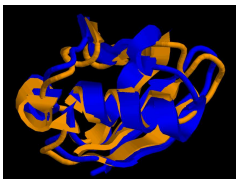
Comparative Modeling

Goal:
- Use existing structure(s) to help determine new structure

Taxonomy:
- Homology modeling
- Fold recognition

Homology Modeling

Motivation:
- If sequence similarity is high,
  then structural similarity is probably high, too



Ubiquitin (blue)
Ubx-Faf1 (gold)

http://www.cs.cmu.edu/~cjl/teaching/15872AS05/

Homology Modeling

Steps:
1. Sequence-sequence alignment
2. Loop modeling
3. Side-chain positioning
4. Structure refinement

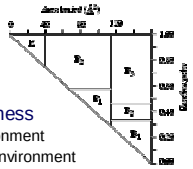
Homology Modeling

Sequence-sequence alignment

- Pairwise or multiple alignment
- Similar to alignment methods we've discussed
  - § Dynamic programming, branch and bound
  - § Amino acid substitution matrices
  - § etc.

Homology Modeling

Use properties of structure to build alignment substitution matrix

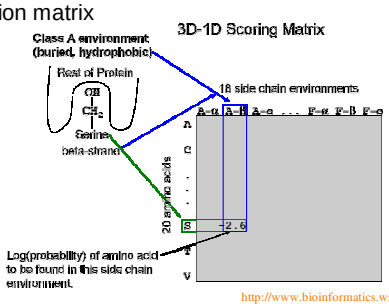


- Secondary structure
  - § α helix
  - § β sheet
  - § Other
- Hydrophobicity, polarity, buriedness
  - § A = B1 buried; hydrophobic environment
  - § B = B2 buried; moderately polar environment
  - § C = B3 buried; polar environment
  - § D = P1 partially buried; moderately polar environment
  - § E = P2 partially buried; polar environment
  - § F = E exposed to solvent

http://www.bioinformatics.wsu.edu/

Homology Modeling

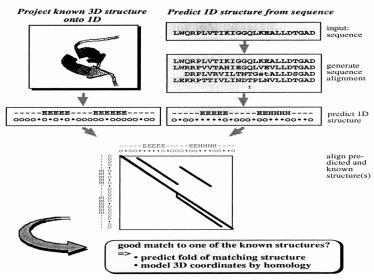
Use properties of structure to build alignment substitution matrix



http://www.bioinformatics.wsu.edu/

Homology Modeling

Use dynamic programming or branch and bound to find best alignment



Homology Modeling

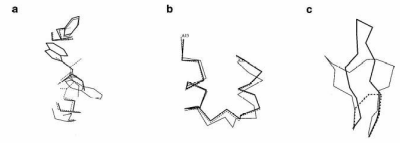
Optimize structure

- Determine positions of residues without alignments
- Adjust side-chain positions

Homology Modeling

Possible errors:

a) Errors in side chain packing
b) Distortions and shifts in correctly aligned regions
c) Errors in regions without template
d) Errors due to misalignment
e) Incorrect templates

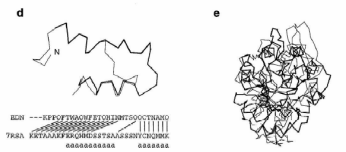


http://www.cs.sunysb.edu/~skiena/549/presentations/protein-folding.ppt

Homology Modeling

Possible errors:

a) Errors in side chain packing
b) Distortions and shifts in correctly aligned regions
c) Errors in regions without template
d) Errors due to misalignment
e) Incorrect templates



http://www.cs.sunysb.edu/~skiena/549/presentations/protein-folding.ppt

Homology Modeling

Issues:

- Usually loops and surface residues have least conservation, and yet they are hardest to place

Results:

- Produce good models if sequence similarity is high enough (>50%)
- Loops and side-chain positioning is problematic, especially if sequence similarity is low

Outline

Introduction

Tertiary structure prediction

- Ab initio methods
- Homology modeling

ØFold recognition

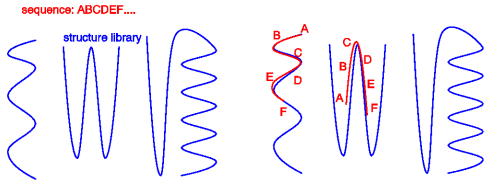
Results

Discussion

Fold Recognition

Goal:

- Thread sequence onto template structure for each fold

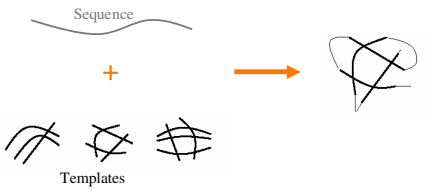


[Huber]

Fold Recognition

Goal:

- Thread sequence onto template structure for each fold

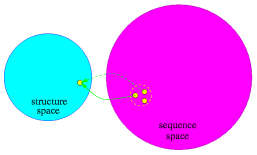


[http://www.bioinformatics.wsu.edu

Fold Recognition

Motivations:

- Structure better conserved that sequence
- There are fewer folds than sequences
- ~30,000 protein structures in PDB
- ~3,000 families (<25% sequence identity)
- ~1000 folds

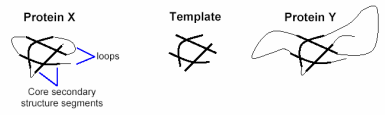


[Huber]

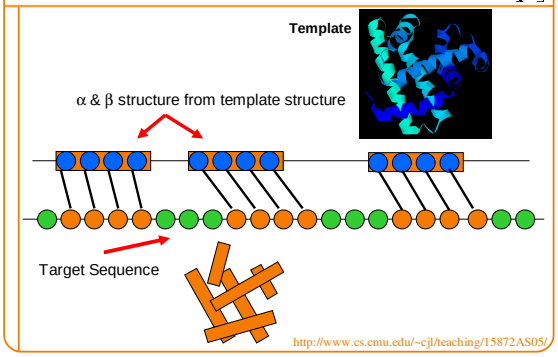
Fold Recognition

Steps:

1. Sequence-structure alignment
2. Partial backbone modeling
3. Loop modeling
4. Side-chain positioning
5. Structure refinement



Fold Recognition



Fold Recognition

After alignment and backbone modeling

- Loop modeling
- Side-chain positioning
- Structure refinement

Use methods similar to ab initio modeling

Outline

Introduction

Tertiary structure prediction

- Ab initio methods
- Homology modeling
- Fold recognition

Results ←

Discussion

CASP

Goal:

- Assess methods for prediction of protein structure from sequence

Methodology:

- Ask experimentalists to delay publication of structure
- Build suite of sequences with unpublished structures
- Allow groups to submit predicted structures
- Evaluate/compare results

CASP

CASP6 targets:

Homology

- Templates can be found with BLAST (CM/easy) = 25
- Templates can be found with PSI-BLAST (CM/hard) = 18
- Templates can be found with profile-profile searchers, significant structural similarity, but not likely convergent evolution (FR/H) = 19

Non-homology

- Template can be found by structure alignment to PDB, but no clear evidence for homology (FR/A) = 15
- No similar structures in PDB (NF) = 10

Results

CASP questions:

- Are the models produced similar to the corresponding experimental structure?
- Is the mapping of the target sequence onto the proposed structure (i.e. the alignment) correct?
- Have similar structures that a model can be based on been identified?
- Are the details of the models correct?
- Has there been progress from the earlier CASPs?
- What methods are most effective?
- Where can future effort be most productively focused?

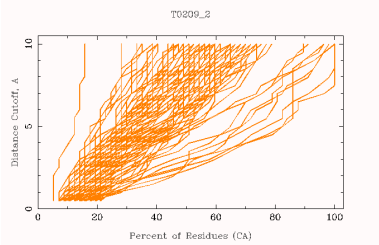
CASP

CASP6 participation:

Prediction format	Number of groups contributing	Number of models designated as 1 for released targets (assessed targets)	Total number of models for released targets (assessed targets)
3D coordinates	166	8686 (6992)	28965 (23119)
Alignments to PDB structures	37	1884 (1455)	5866 (4484)
Residue-residue contacts	17	1050 (830)	1776 (1397)
Structural domains assignments	24	1332 (1033)	1672 (1293)
Disordered regions	20	1429 (1144)	1769 (1420)
Function prediction	26	1067 (867)	1235 (990)
All	**288 (unique)**	**15448 (12321)**	**41283 (32783)**

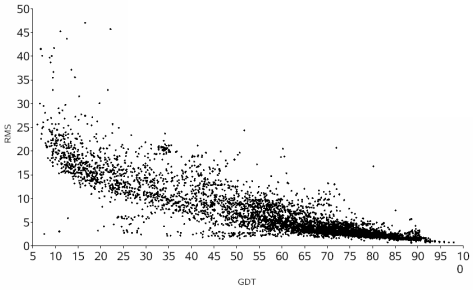
CASP

CASP6 example results:



CASP

CASP6 results:



http://predictioncenter.org/casp6/meeting/presentations/CM_assessment.pdf

Results

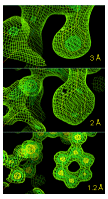
Accuracy comparison:

Approach		req. seq. identity	accuracy
NMR, X-ray		-	1.0 Å
Comparative	sequence	> 50%	1.5 Å
	threading	> 30%	3.5 Å
	threading	< 30%	high error
De novo		insignificant	4-8 Å

Results

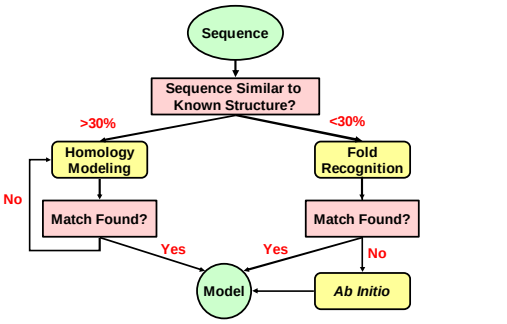
Accuracy comparison:

Approach		req. seq. identity	accuracy
NMR, X-ray		-	1.0 Å
Comparative	sequence	> 50%	1.5 Å
	threading	> 30%	3.5 Å
	threading	< 30%	high error
De novo		insignificant	4-8 Å



http://ruppweb.dyndns.org/

Summary



http://www.cs.cmu.edu/~cjl/teaching/15872AS05/

Discussion

?