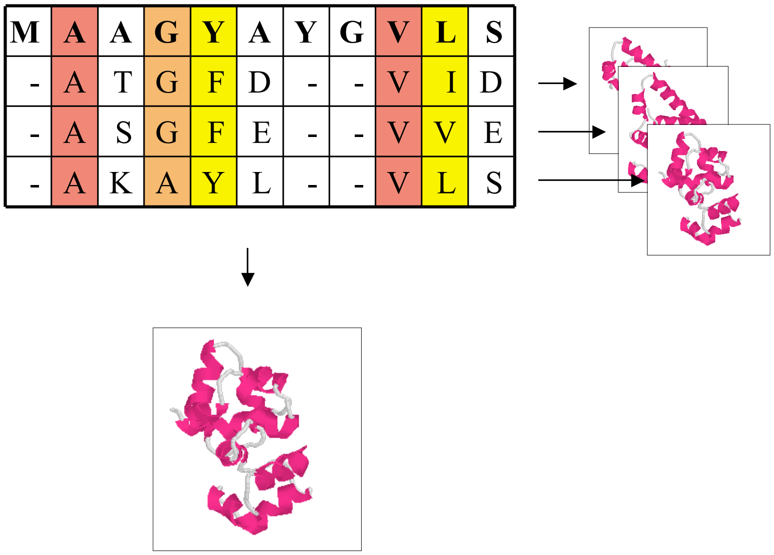
Protein Structure Prediction

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Homology Modeling



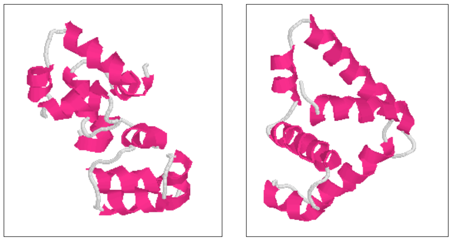
M	**A**	**A**	**G**	**Y**	**A**	**Y**	**G**	**V**	**L**	**S**
-	A	T	G	F	D	-	-	V	I	D
-	A	S	G	F	E	-	-	V	V	E
-	A	K	A	Y	L	-	-	V	L	S

Fold Recognition

Sequence:

M	A	A	G	Y	A	V	L	S

+
Known folds



→

Ab Initio Structure Prediction

M	A	A	G	Y	A	V	L	S

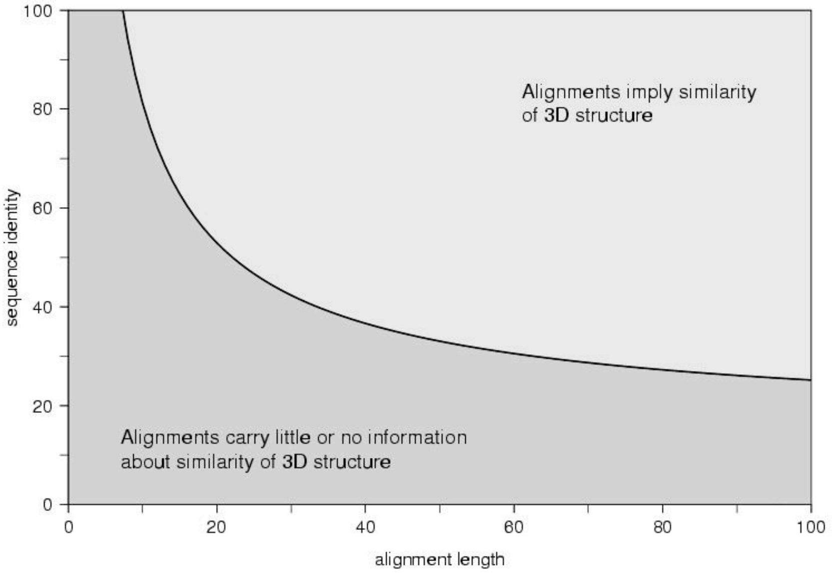
→

Homology Modeling

- Align sequence to protein sequences with known structure.
- Construct and evaluate model of 3D structure from alignment.
- Requirement: Close match to template sequences with known 3D structure (sequence similarity of at least 25%).

Note: about 25% of the protein sequences in the Swiss-Prot database have templates for at least part of the sequence!

Threshold for Structural Homology



Rost B, Protein Engineering 12 (1999).

Homology Modeling Approach

1. Find set of sequences related to target sequence.
2. Align target sequence to template sequences (key step).
3. Construct 3D model for core (backbone):
   - Conserved regions → conserved structure / coordinates.
   - Structure diverges → use sequence similarity, secondary structure prediction, manual prediction, etc. to fill in gaps.
4. Construct 3D models for loops:

   Search loop conformation library, limited protein folding.
5. Model location of side chains

   Search rotamer library, use molecular dynamics.
6. Optimize / verify the model

   Improve likelihood / ensure legality of model.

Homology Modeling Web Pages

MODELLER
`http://salilab.org/modeller/modeller.html`

SWISS-MODEL
`http://www.expasy.org/swissmod/SWISS-MODEL.html`

Quality Assessment

- Goal
  - Ensure predicted 3D structure is possible / probable in practice
  - Based on general knowledge of protein structures
- Criteria
  - Carbon backbone conformations allowed (Ramachandran map)
  - Legal bond lengths, angles, dihedrals
  - Peptide bonds are planar
  - Side chain conformations correspond to ones in rotamer library
  - Hydrogen-bonding of polar atoms if buried
  - Proper environments for hydrophobic / hydrophilic residues
  - No bad atom-atom contacts
  - No holes inside 3D structure
  - Solvent accessibility

Quality Assessment Programs

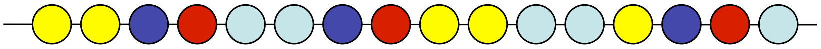
VERIFY3D
`http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D`

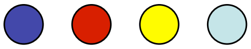
PROCHECK
`http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html`

WHATIF
`http://www.cmbi.kun.nl/whatif/`

Fold Recognition

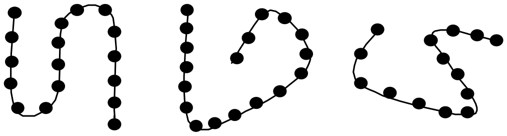
- The input sequence is threaded on different folds from a library of known folds.
- Using scoring functions, we get a score for the compatibility between the sequence and the structures.



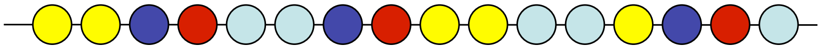


Amino acids with different chemical properties

Library of known folds:



Fold Recognition



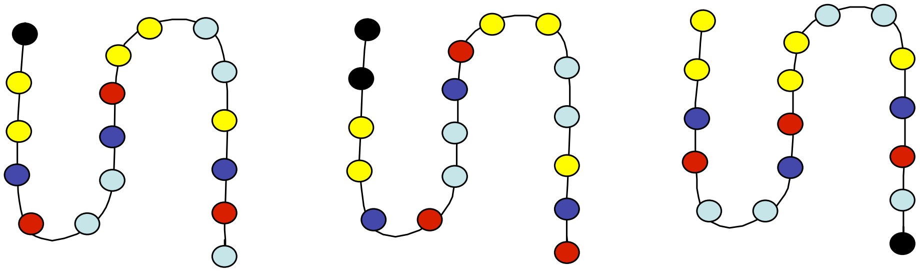
🔵 Hydrogen donor

🔴 Hydrogen acceptor

🟡 Hydrophobic

⚪ Glycin

Good score!



Fold Recognition

- This method is less accurate than homology modeling, but can be applied in more cases.
- When the real fold of the input sequence is not represented in the structural database, we do not get a good solution (duh).
- The most important part is the accuracy of the scoring function. The scoring function is the major difference between the approaches used for fold recognition.

Profile Based Scoring Functions

- In methods based on structural profiles, for every fold a profile is built based on structural features of the fold and the compatibility of every amino acid to the features.
- The structural features of each position are based on the combination of secondary structure, solvent accessibility, and the properties of the local environment (such as hydrophobicity, etc).

Contact Potentials

- This method is based on predefined tables which include (pseudo-energetic) scores for each interaction of two amino acids.
- This method makes use of a distance matrix for the representation of different folds.
- For each pair of amino acids which are close in space, the interaction energy is summed up. The total sum is the indication for the “fitness” of the sequence for the given structure .

Web Sites for Fold Recognition

3D-PSSM

`http://www.bmm.icnet.uk/~3dpssm`

LIBRA I

`http://www.ddbj.nig.ac.jp/htmls/Email/libra/LIBRA_I.html`

UCLA DOE

`http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html`

123D

`http://www-Immb.ncifcrf.gov/~nicka/123D.html`

PROFIT

`http://lore.came.sbg.ac.at/home.html`

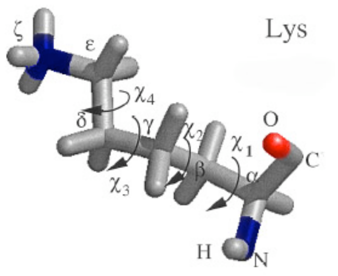
Ab Initio Methods

- Ab initio: “From the beginning”.
- Assumption 1: All the information about the structure of a protein is contained in its sequence of amino acids.
- Assumption 2: The structure that a (globular) protein folds into is the structure with the lowest free energy.
- Finding native-like conformations require:
  - A scoring function (potential).
  - A search strategy.

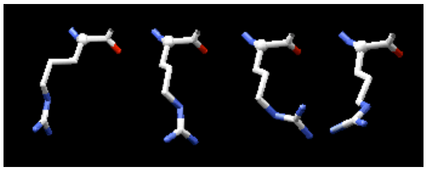
Representations of the Protein

- Sidechain: represented as all atoms, rotamers, carbon α or β, centroids.
- Backbone: torsion angles restricted to discrete values commonly seen in known structures (using a small set of pre-selected ϕ-ψ angles, angels chosen from secondary structure elements, selection of fragments of known structures), secondary structure rigid bodies, lattice models.

Rotamer Libraries



Some members of the rotamer library:



Potential Functions

- So-called “molecular mechanics” potentials model the force that determine protein conformation using physically based functional forms (van der Waals, Coulomb).
- Potentials empirically derived from known structures in the Protein Data Bank.

Search Strategies

- Molecular dynamics. Not really feasible for ab initio prediction per se.
- Probabilistic search algorithms (simulated annealing, genetic algorithms) generate ensembles of candidate structures. Additional methods to discriminate between those are needed.

Rosetta

- The scoring function is a model generated using various contributions. It has a sequence dependent part (including for example a term for hydrophobic burial), and a sequence independent part (including for example a term for strand-strand packing).
- The search is carried out using simulated annealing. The move set is defined by a fragment library for each three and nine residue segment of the chain. The fragments are extracted from observed structures in the PDB.

The Rosetta Scoring Function

$$\mathrm{P}(\text{structure}|\text{sequence}) \propto \mathrm{P}(\text{sequence}|\text{structure}) \times \mathrm{P}(\text{structure})$$

Sequence dependent:

- hydrophobic burial
- residue pair interaction

Sequence independent:

- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

The Sequence Dependent Term

$$\mathrm{P}(\mathrm{aa}_1, \ldots, \mathrm{aa}_n|\mathrm{X}) =$$

$$\prod_i \mathrm{P}(\mathrm{aa}_i|\mathrm{X}) \times$$

$$\prod_{i<j} \frac{\mathrm{P}(\mathrm{aa}_i, \mathrm{aa}_j|\mathrm{X})}{\mathrm{P}(\mathrm{aa}_i|\mathrm{X})\mathrm{P}(\mathrm{aa}_j|\mathrm{X})} \times$$

$$\prod_{i<j<k} \frac{\mathrm{P}(\mathrm{aa}_i, \mathrm{aa}_j, \mathrm{aa}_k|\mathrm{X})\mathrm{P}(\mathrm{aa}_i|\mathrm{X})\mathrm{P}(\mathrm{aa}_j|\mathrm{X})\mathrm{P}(\mathrm{aa}_k|\mathrm{X})}{\mathrm{P}(\mathrm{aa}_i, \mathrm{aa}_j|\mathrm{X})\mathrm{P}(\mathrm{aa}_i, \mathrm{aa}_k|\mathrm{X})\mathrm{P}(\mathrm{aa}_j, \mathrm{aa}_k|\mathrm{X})} \times$$

$$\ldots$$

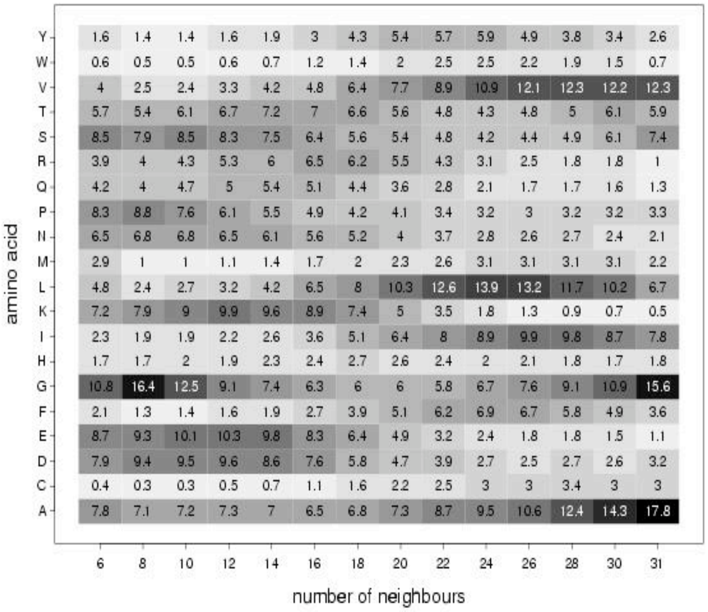
The Sequence Dependent Term

$$\mathsf{P}(\text{sequence}|\text{structure}) \approx \mathsf{P}_{\text{env}} \times \mathsf{P}_{\text{pair}}$$

$$\mathsf{P}_{\text{env}} = \prod_{i} \mathsf{P}(\text{aa}_i|\mathsf{E}_i)$$

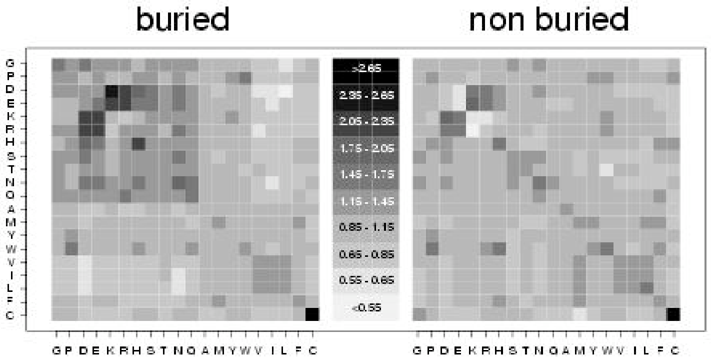
$$\mathsf{P}_{\text{pair}} = \prod_{i<j} \frac{\mathsf{P}(\text{aa}_i, \text{aa}_j|\mathsf{E}_i, \mathsf{E}_j, \mathsf{r}_{ij})}{\mathsf{P}(\text{aa}_i|\mathsf{E}_i, \mathsf{r}_{ij})\mathsf{P}(\text{aa}_j|\mathsf{E}_j, \mathsf{r}_{ij})}$$

Hydrophobic Burial



	6	8	10	12	14	16	18	20	22	24	26	28	30	31
Y	1.6	1.4	1.4	1.6	1.9	3	4.3	5.4	5.7	5.9	4.9	3.8	3.4	2.6
W	0.6	0.5	0.5	0.6	0.7	1.2	1.4	2	2.5	2.5	2.2	1.9	1.5	0.7
V	4	2.5	2.4	3.3	4.2	4.8	6.4	7.7	8.9	10.9	12.1	12.3	12.2	12.3
T	5.7	5.4	6.1	6.7	7.2	7	6.6	5.6	4.8	4.3	4.8	5	6.1	5.9
S	8.5	7.9	8.5	8.3	7.5	6.4	5.6	5.4	4.8	4.2	4.4	4.9	6.1	7.4
R	3.9	4	4.3	5.3	6	6.5	6.2	5.5	4.3	3.1	2.5	1.8	1.8	1
Q	4.2	4	4.7	5	5.4	5.1	4.4	3.6	2.8	2.1	1.7	1.7	1.6	1.3
P	8.3	8.8	7.6	6.1	5.5	4.9	4.2	4.1	3.4	3.2	3	3.2	3.2	3.3
N	6.5	6.8	6.8	6.5	6.1	5.6	5.2	4	3.7	2.8	2.6	2.7	2.4	2.1
M	2.9	1	1	1.1	1.4	1.7	2	2.3	2.6	3.1	3.1	3.1	3.1	2.2
L	4.8	2.4	2.7	3.2	4.2	6.5	8	10.3	12.6	13.9	13.2	11.7	10.2	6.7
K	7.2	7.9	9	9.9	9.6	8.9	7.4	5	3.5	1.8	1.3	0.9	0.7	0.5
I	2.3	1.9	1.9	2.2	2.6	3.6	5.1	6.4	8	8.9	9.9	9.8	8.7	7.8
H	1.7	1.7	2	1.9	2.3	2.4	2.7	2.6	2.4	2	2.1	1.8	1.7	1.8
G	10.8	16.4	12.5	9.1	7.4	6.3	6	6	5.8	6.7	7.6	9.1	10.9	15.6
F	2.1	1.3	1.4	1.6	1.9	2.7	3.9	5.1	6.2	6.9	6.7	5.8	4.9	3.6
E	8.7	9.3	10.1	10.3	9.8	8.3	6.4	4.9	3.2	2.4	1.8	1.8	1.5	1.1
D	7.9	9.4	9.5	9.6	8.6	7.6	5.8	4.7	3.9	2.7	2.5	2.7	2.6	3.2
C	0.4	0.3	0.3	0.5	0.7	1.1	1.6	2.2	2.5	3	3	3.4	3	3
A	7.8	7.1	7.2	7.3	7	6.5	6.8	7.3	8.7	9.5	10.6	12.4	14.3	17.8

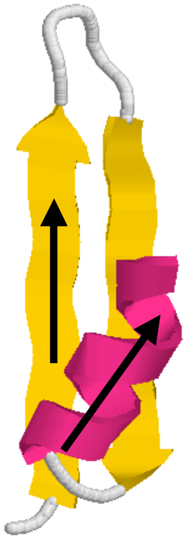
Residue Pair Interaction

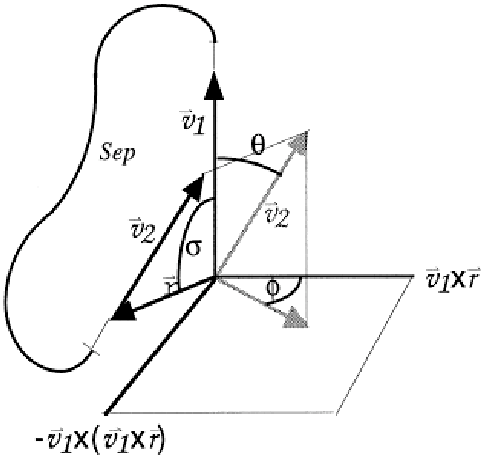


The Sequence Independent Term

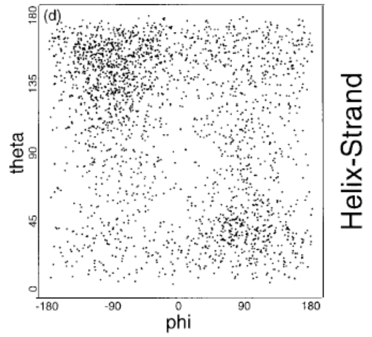
$$P(r, \phi, \theta, \sigma, hb|sep) \approx$$

$$P(\phi, \theta|r,sep) \times P(hb|r,sep) \times P(\sigma|r,sep) \times P(r|sep)$$

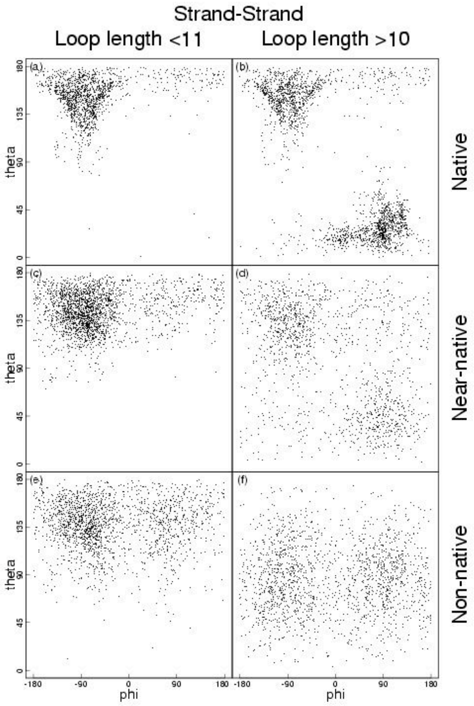


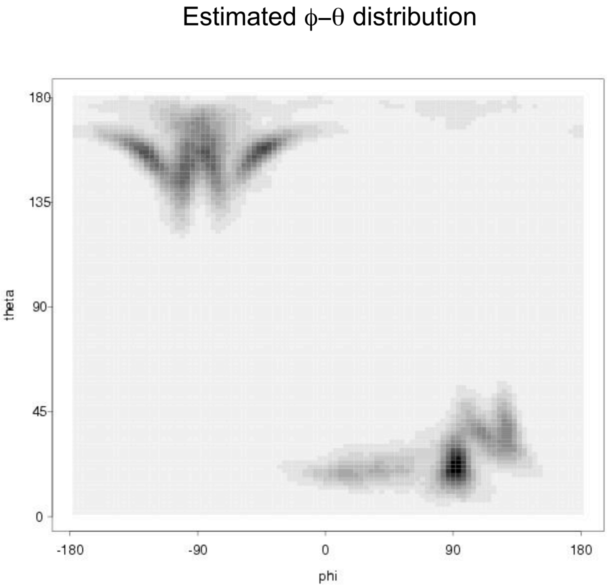

Sep
$\vec{v}_1$
θ
$\vec{v}_2$
σ
$\vec{v}_2$
$\vec{r}$
φ
$\vec{v}_1 \times \vec{r}$
$-\vec{v}_1 \times (\vec{v}_1 \times \vec{r})$

vector representation

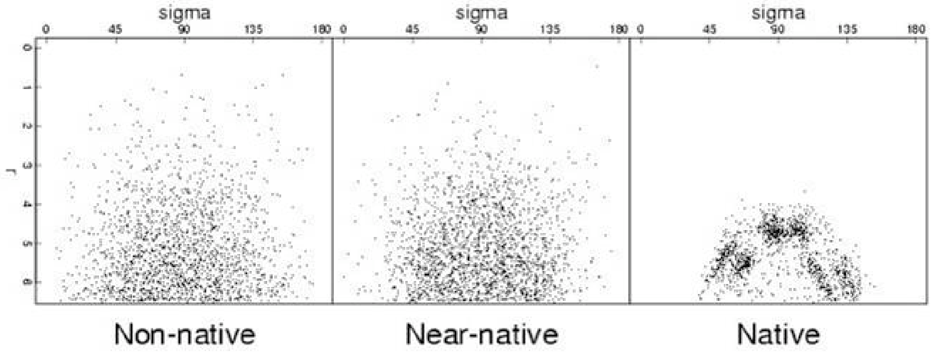


Strand Packing – Helps!





Sheer Angles – Help not!



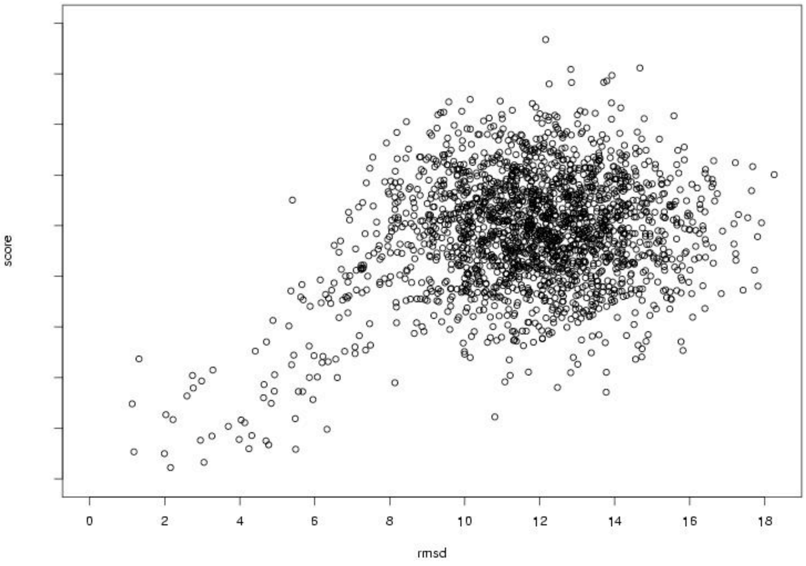
The Model

$$\mathsf{P}(\text{structure}) = \mathsf{P}_{\mathsf{A}}^{w_{\mathsf{A}}}\mathsf{P}_{\mathsf{B}}^{w_{\mathsf{B}}}\mathsf{P}_{\mathsf{C}}^{w_{\mathsf{C}}}, \quad w_{\mathsf{X}} > 0.$$

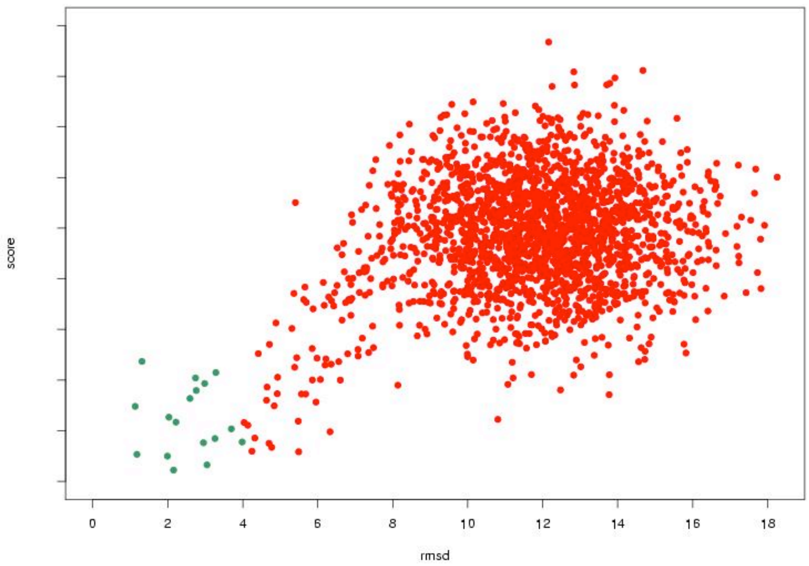
$$- \log \mathsf{P}(\text{structure}|\text{sequence}) \propto$$
$$- \log \mathsf{P}(\text{sequence}|\text{structure}) - \log \mathsf{P}(\text{structure})$$

$$\text{g(rmsd)} = w_{\text{protein}} + w_{\text{HS}} \log \mathsf{P}_{\text{HS}} + w_{\text{SS}} \log \mathsf{P}_{\text{SS}} + w_{\text{VdW}}\,\text{VdW} +$$
$$w_{\text{sheet}} \log \mathsf{P}_{\text{sheet}} + w_{\text{seq}} \left(\log \mathsf{P}_{\text{env}} + \log \mathsf{P}_{\text{pair}}\right)$$

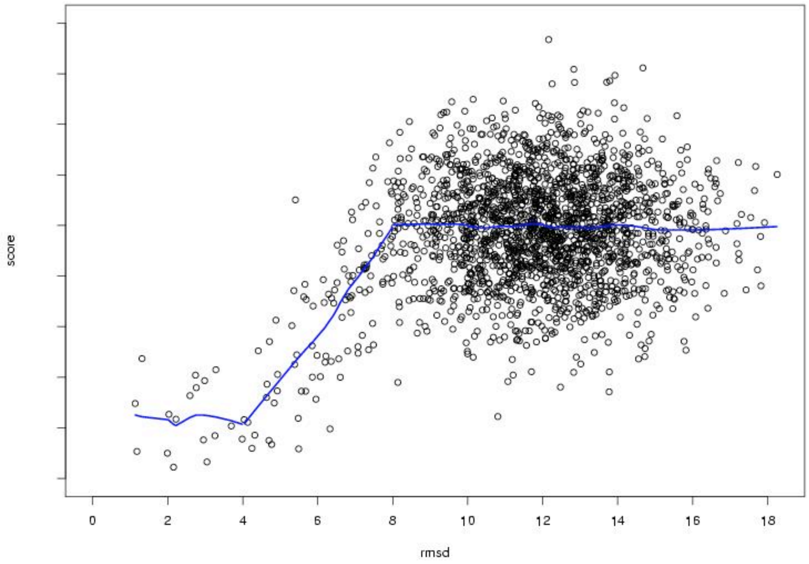
Parameter Estimation



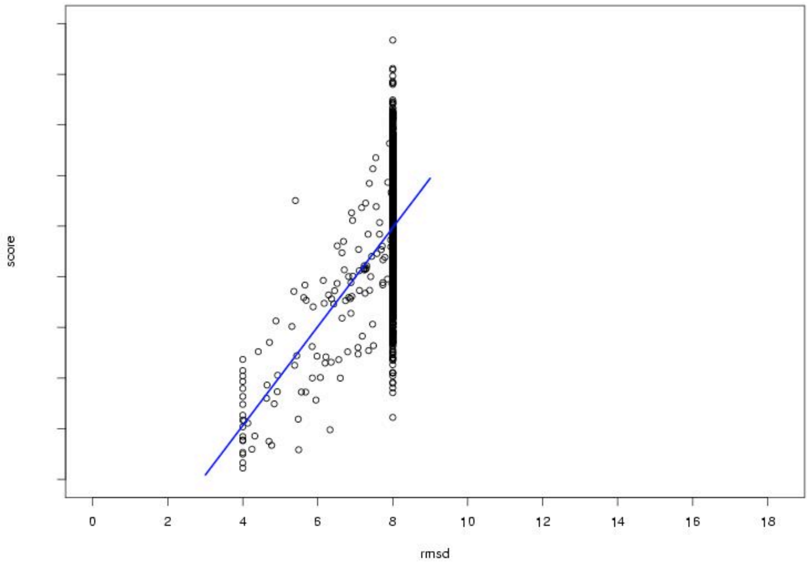
Parameter Estimation

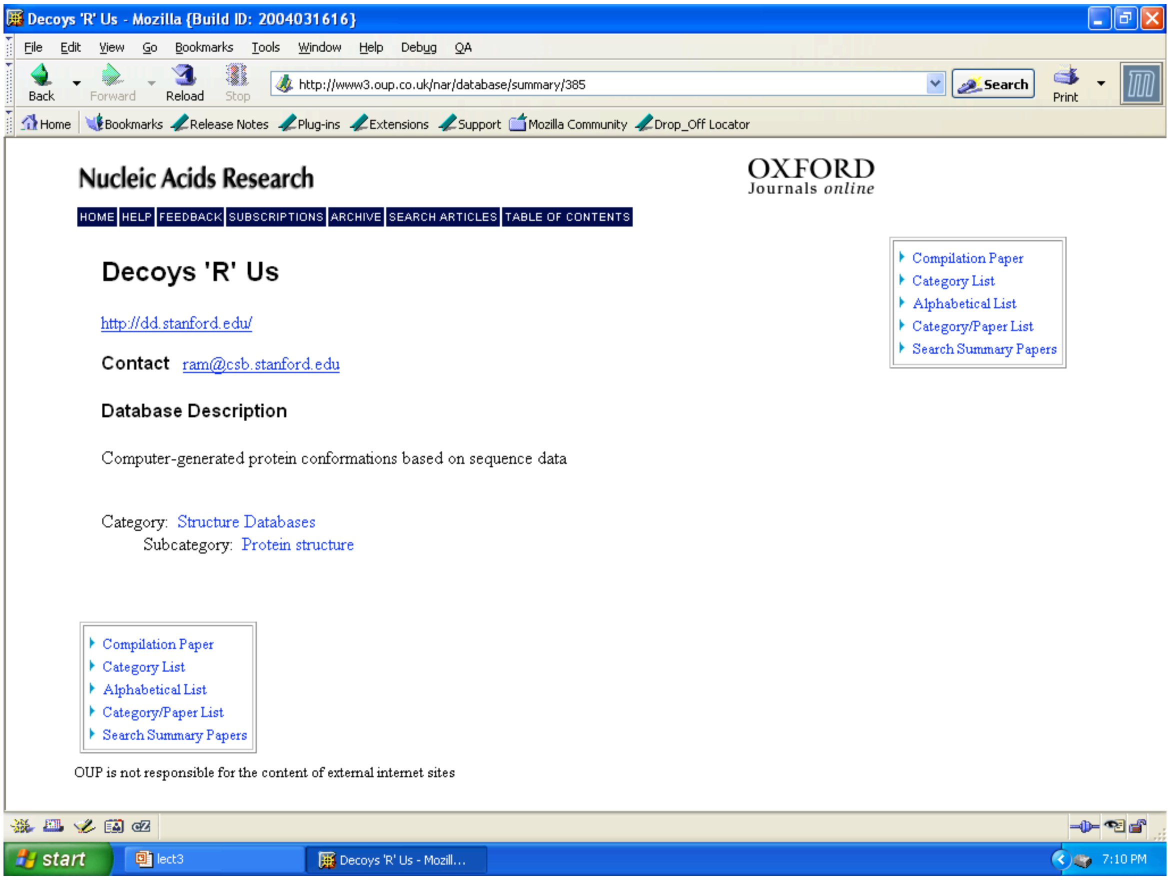


Parameter Estimation



Parameter Estimation





Decoys 'R' Us

http://dd.stanford.edu/

Contact ram@csb.stanford.edu

Database Description

Computer-generated protein conformations based on sequence data

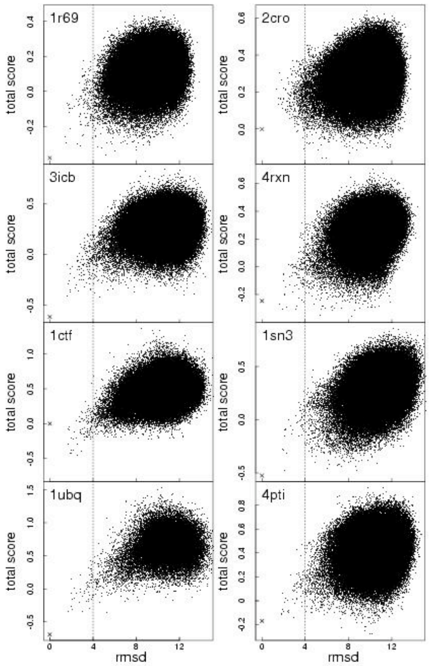
Category: Structure Databases
Subcategory: Protein structure

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

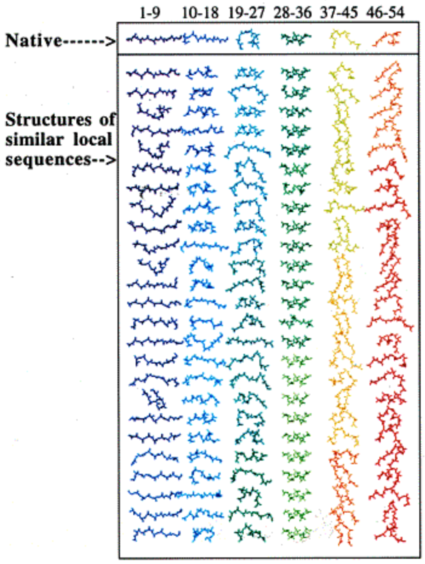
- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

OUP is not responsible for the content of external internet sites

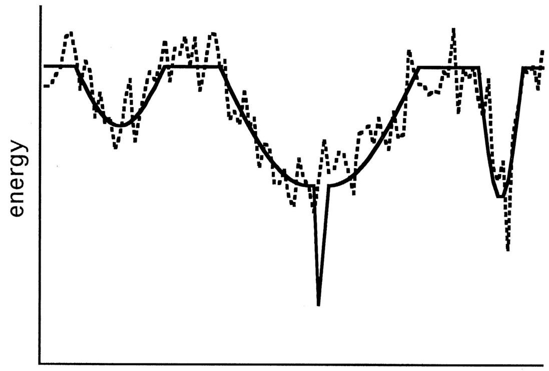
Validation Data Set

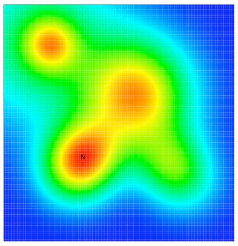


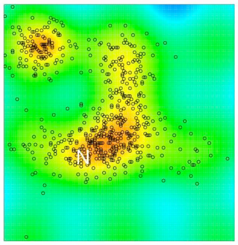
Fragment Selection



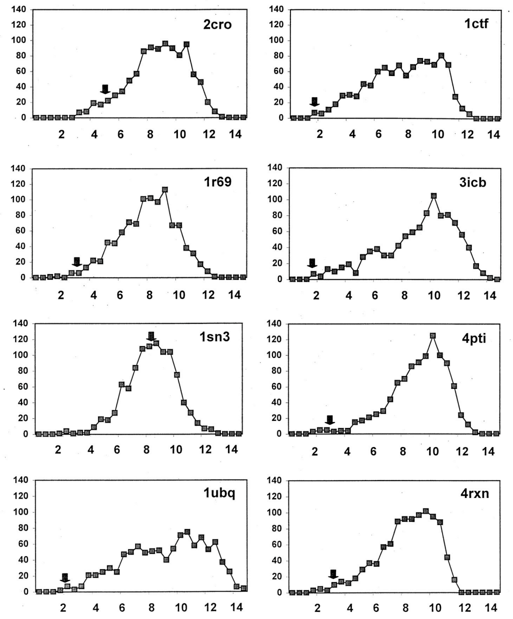
3D Clustering







3D Clustering

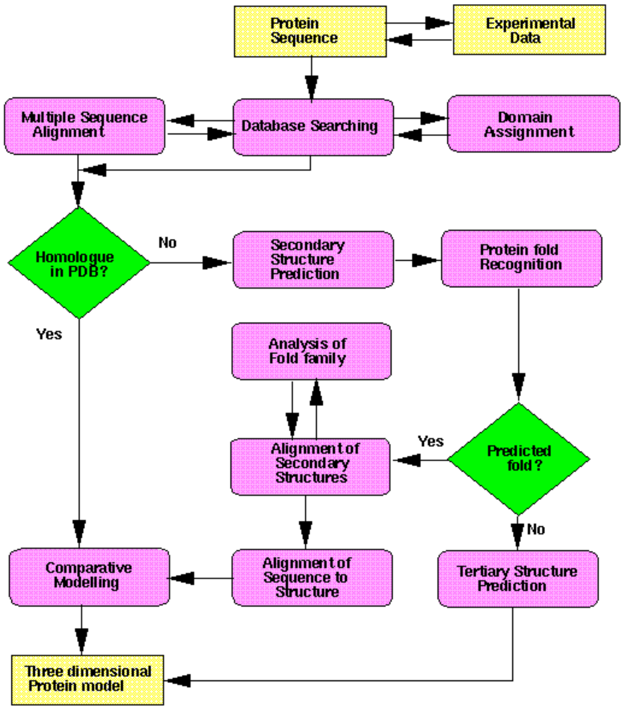


Assessing Structure Prediction

- CASP (Critical Assessment of Protein Structure Prediction)
  - Competitions measuring current state of the art in protein structure prediction.
  - Researchers predict structure of actual protein sequences.
  - Compare with laboratory determination of structure.
  - Held in 1994, 1996, 1998, 2000, 2002, 2004.

- CAFASP (Critical Assessment of Fully Automated Protein Structure Prediction).

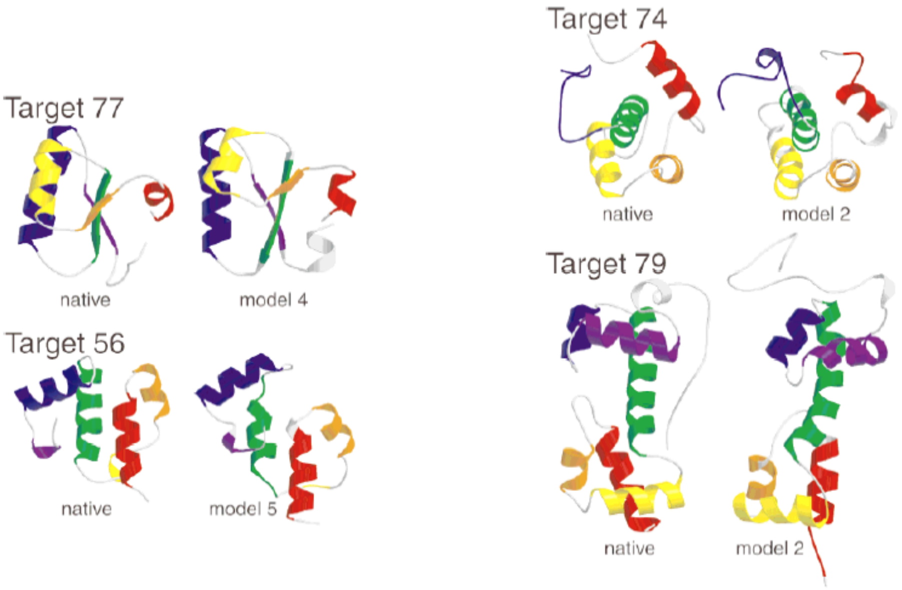
Protein Structure Prediction


Protein Sequence
Experimental Data
Multiple Sequence Alignment
Database Searching
Domain Assignment
Homologue in PDB?
No
Yes
Secondary Structure Prediction
Protein fold Recognition
Analysis of Fold family
Alignment of Secondary Structures
Yes
Predicted fold ?
No
Comparative Modelling
Alignment of Sequence to Structure
Tertiary Structure Prediction
Three dimensional Protein model

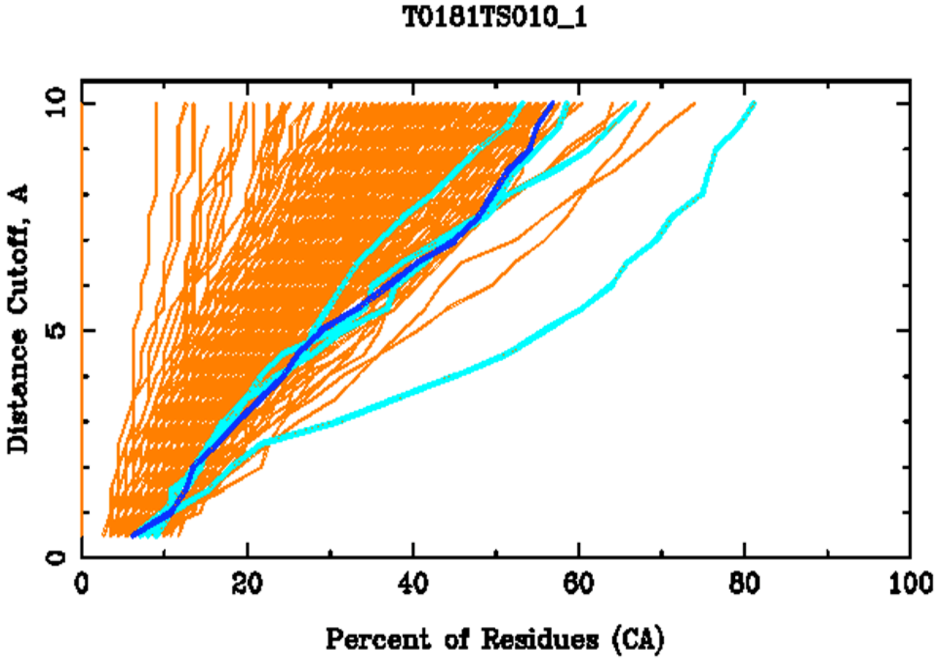
CASP3 Protocol

- Construct a multiple sequence alignment from ϕ-blast.
- Edit the multiple sequence alignment.
- Identify the ab initio targets from the sequence.
- Search the literature for biological and functional information.
- Generate 1200 structures, each the result of 100,000 cycles.
- Analyze the top 50 or so structures by an all-atom scoring function (also using clustering data).
- Rank the top 5 structures according to protein-like appearance and/or expectations from the literature.

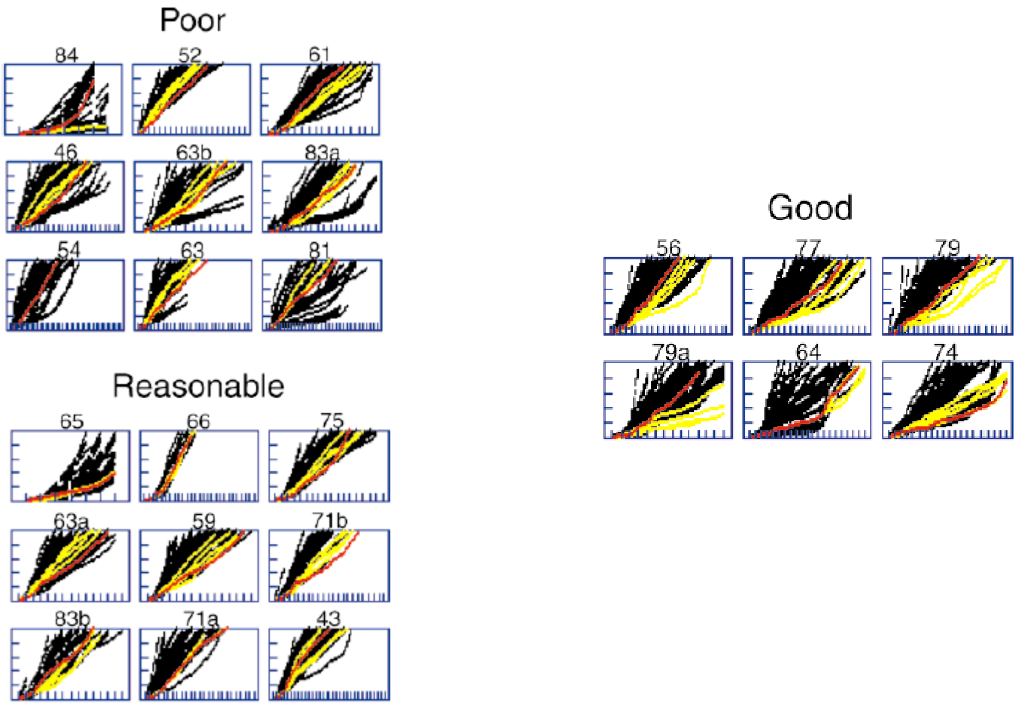
CASP3 Predictions



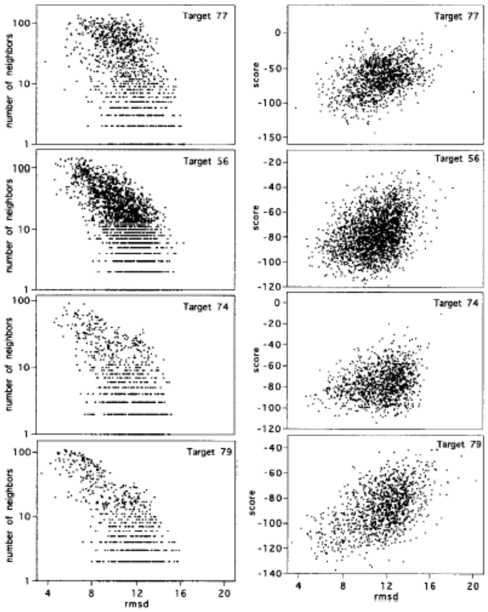
Hubbard Plot



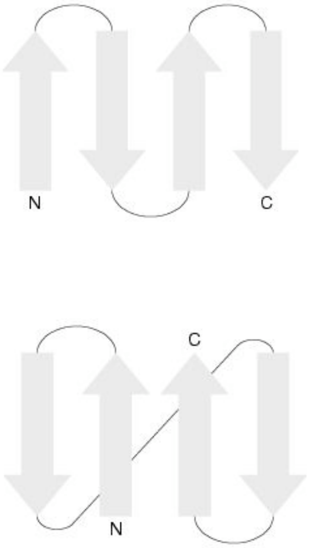
CASP3 Results

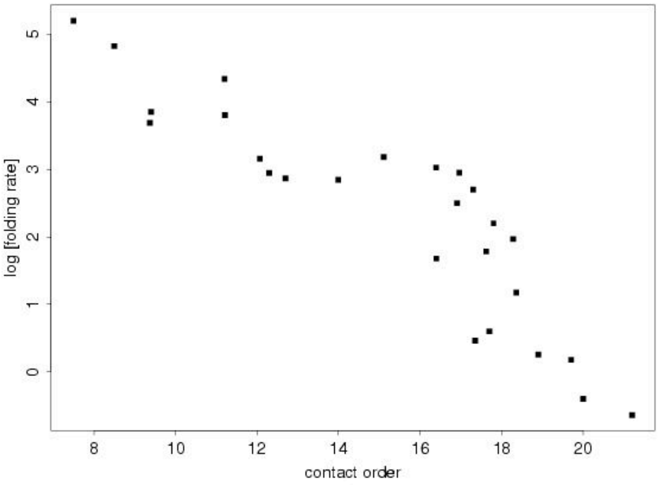


3D Clustering in CASP3

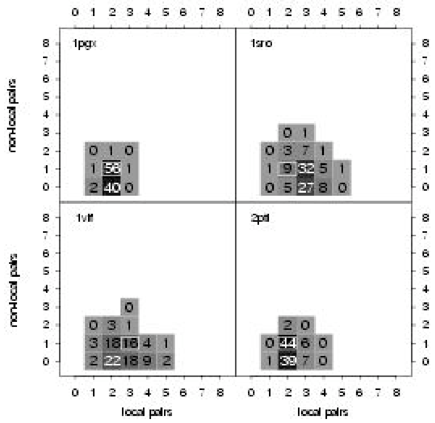


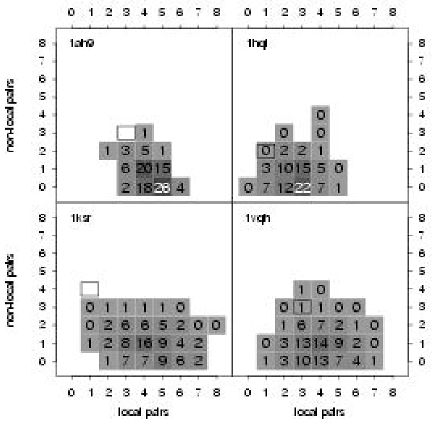
Contact Order



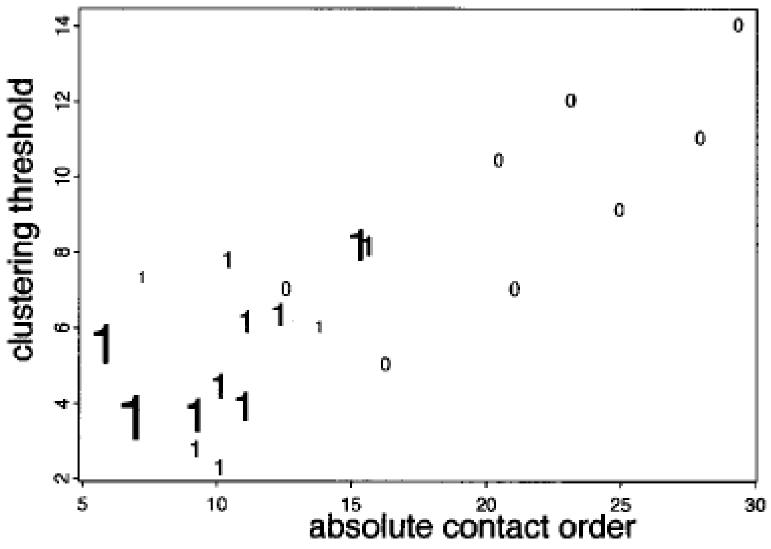


Contact Order

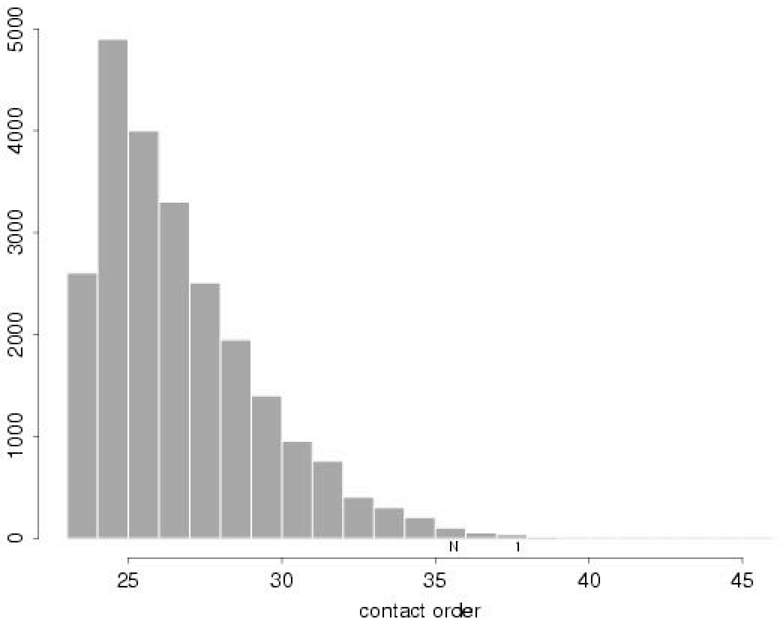




Clustering and Contact Order



Decoy Enrichment in CASP4

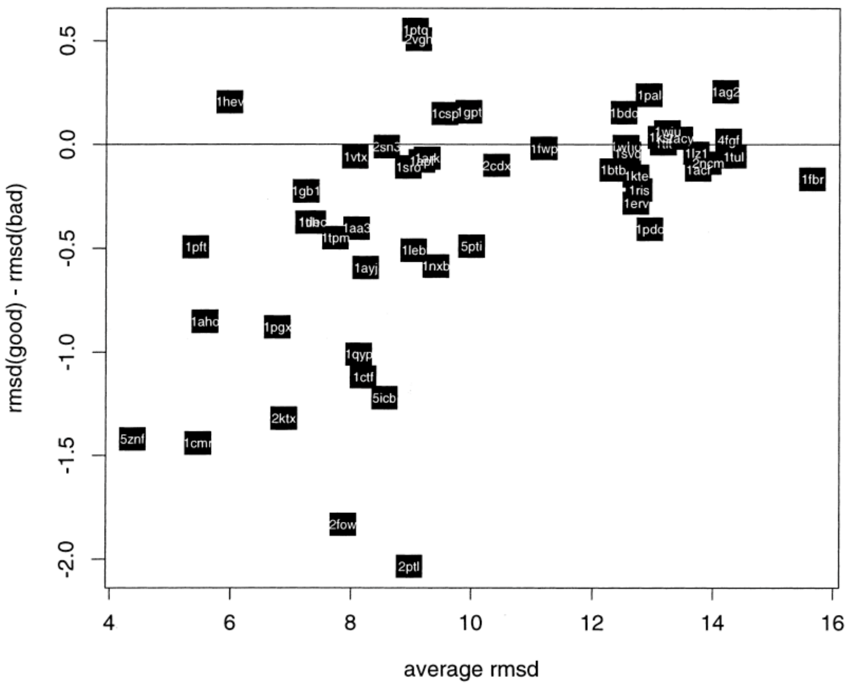


A Filter for Bad β-Sheets

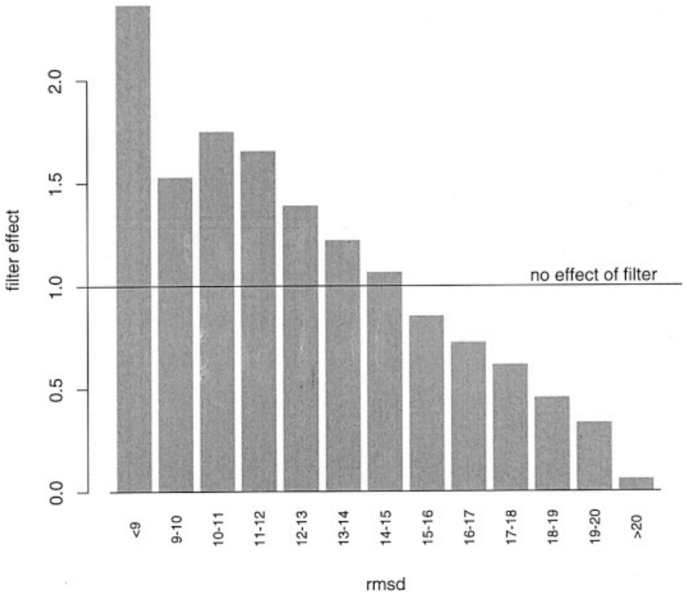
Many decoys do not have proper sheets. Filtering those out seems to enhance the rmsd distribution in the decoy set. Bad features we see in decoys include:

- No strands,
- Single strands,
- Too many neighbours,
- Single strand in sheets,
- Bad dot-product,
- False handedness,
- False sheet type (barrel),
- …

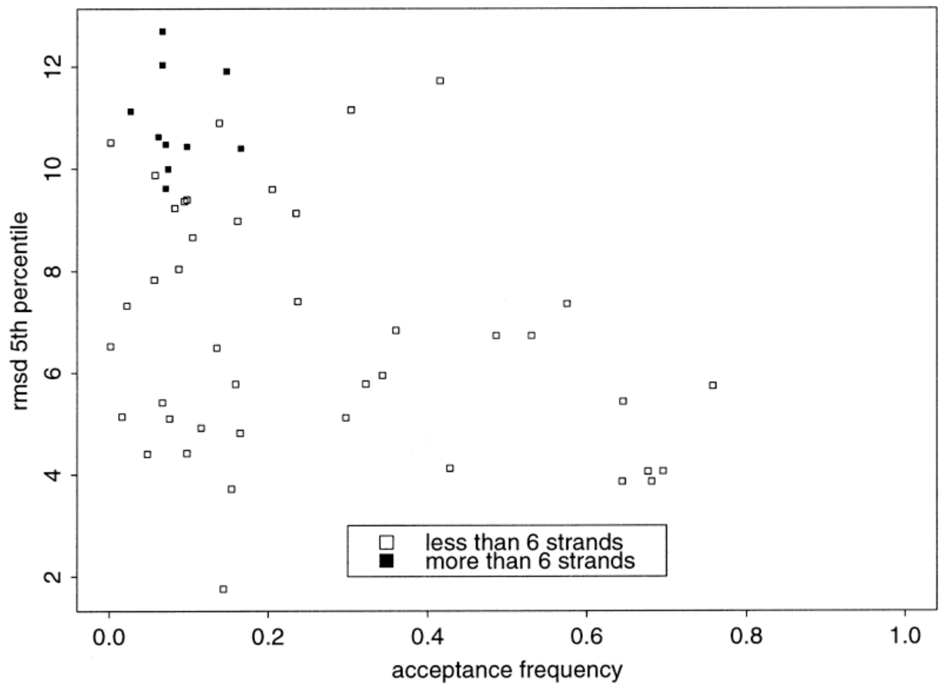
A Filter for Bad β-Sheets



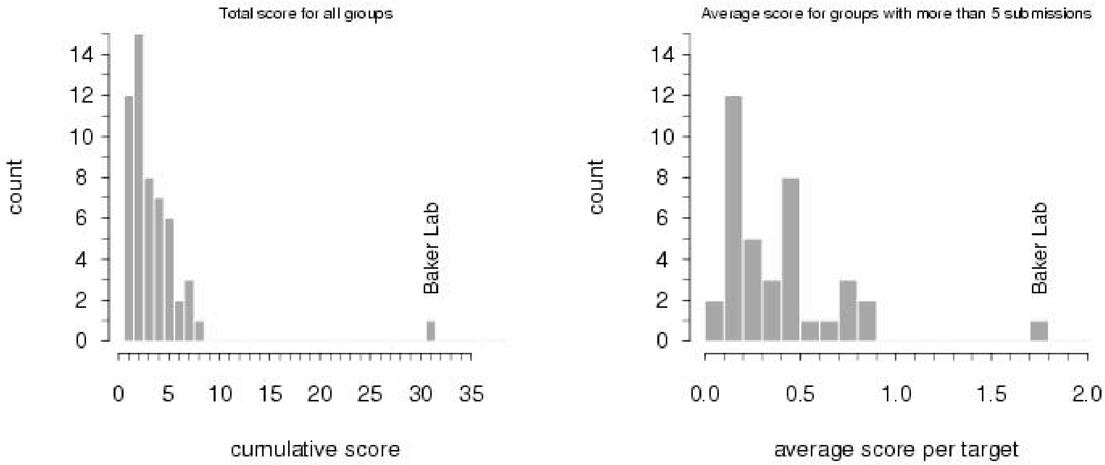
A Filter for Bad β-Sheets



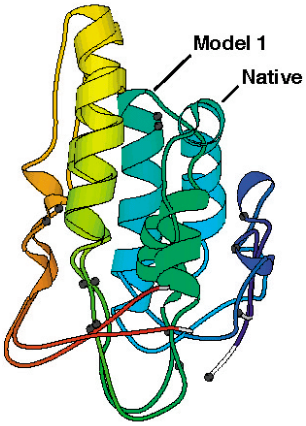
A Filter for Bad β-Sheets

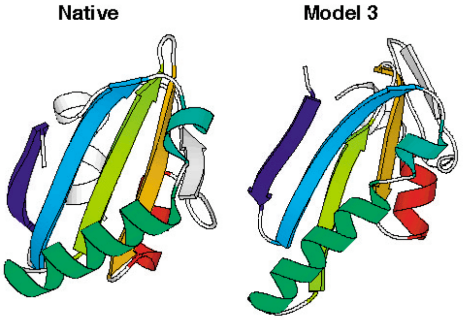


CASP 4

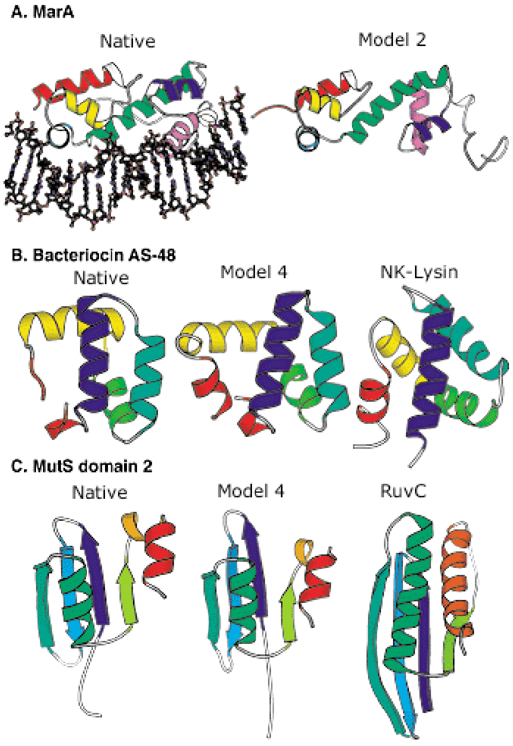


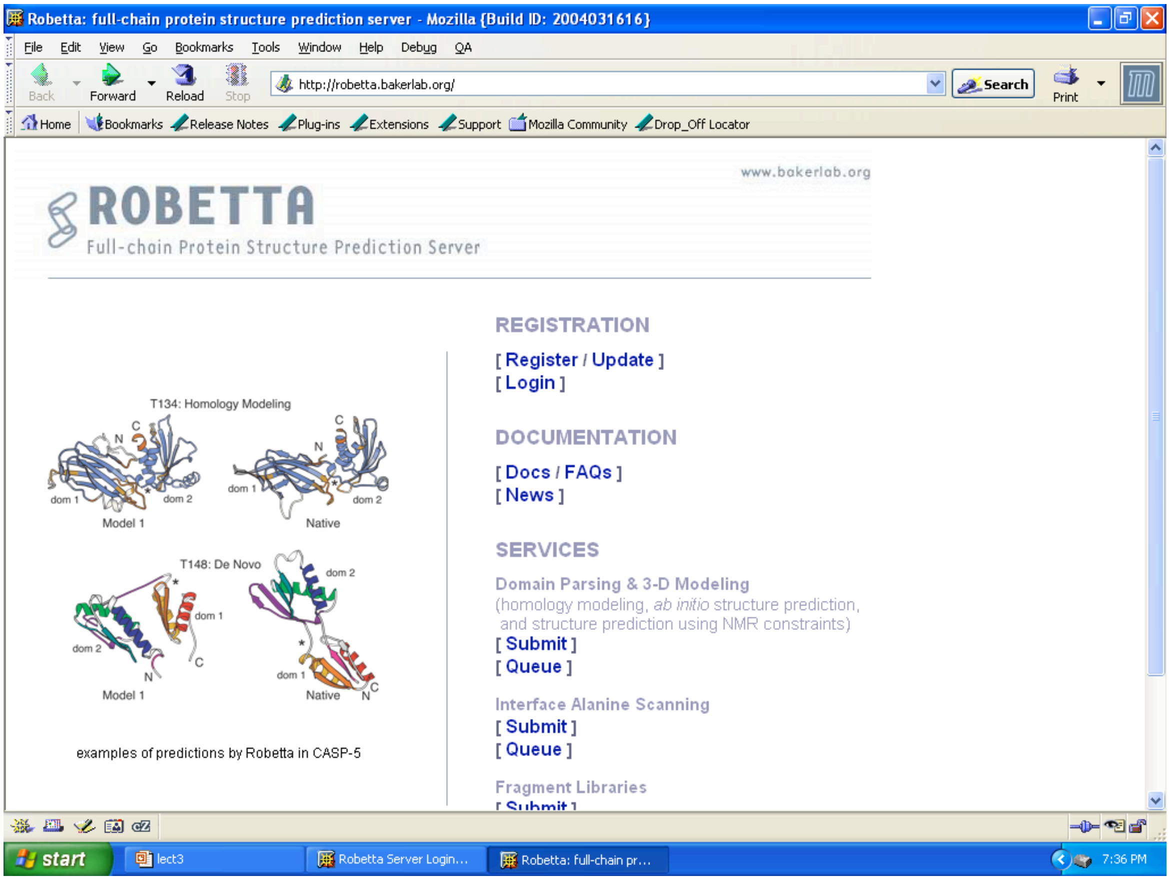
Rosetta in CASP4





CASP 4





www.bakerlab.org

ROBETTA

Full-chain Protein Structure Prediction Server

examples of predictions by Robetta in CASP-5

REGISTRATION

[Register / Update]
[Login]

DOCUMENTATION

[Docs / FAQs]
[News]

SERVICES

Domain Parsing & 3-D Modeling
(homology modeling, *ab initio* structure prediction, and structure prediction using NMR constraints)
[Submit]
[Queue]

Interface Alanine Scanning
[Submit]
[Queue]

Fragment Libraries
[Submit]

Applications and Other Uses of Rosetta

- Other uses of Rosetta:
  - Homology modeling.
  - Rosetta NMR.
  - Protein interactions (docking).

- Applications of Rosetta:
  - Functional annotation of genes.
  - Novel protein design.