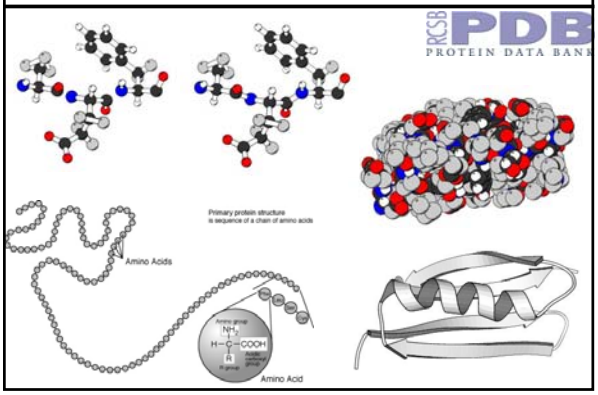
Protein Folding and Structure Prediction

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Proteins



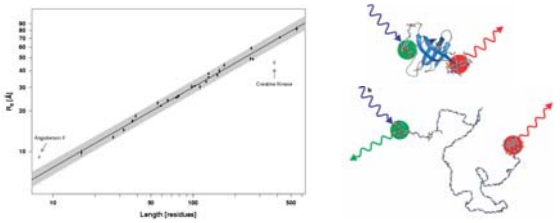
Protein Folding vs Structure Prediction

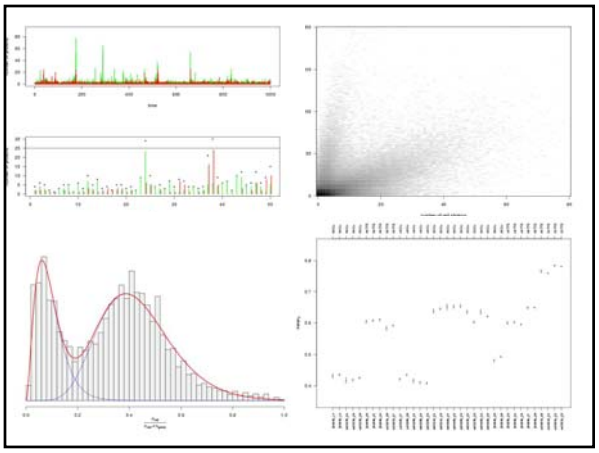
- Protein folding is concerned with the process of the protein taking its three dimensional shape.
- Protein structure prediction is solely concerned with the 3D structure of the protein, using theoretical and empirical means to get to the end result.

This presentation is bit about both. The role of statistics in protein folding and structure prediction is usually to support or discredit some hypothesis based on physical principles, and to build stochastic models.

Some Issues in Protein Folding

- Do chemically denatured proteins behave as random coils?
- What determines the folding rate of a protein?
- Are amino acids in proteins conserved because of folding kinetics?





Flavors of Structure Prediction

- Homology modeling,
- Fold recognition (threading),
- Ab initio (de novo, new folds) methods.

ROSETTA is mainly an ab initio structure prediction algorithm, although various parts of it can be used for other purposes as well (such as homology modeling).

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.

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

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

Sequence dependent:
- hydrophobic burial
- residue pair interaction

Sequence independent:
- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

The Sequence Dependent Term

$$P(aa_1, \ldots, aa_n|X) =$$

$$\prod_i P(aa_i|X) \times$$

$$\prod_{i<j} \frac{P(aa_i, aa_j|X)}{P(aa_i|X)P(aa_j|X)} \times$$

$$\prod_{i<j<k} \frac{P(aa_i, aa_j, aa_k|X)P(aa_i|X)P(aa_j|X)P(aa_k|X)}{P(aa_i, aa_j|X)P(aa_i, aa_k|X)P(aa_j, aa_k|X)} \times$$

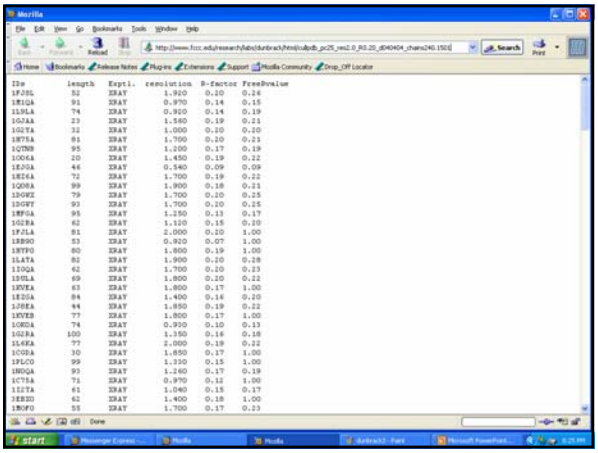
$$\ldots$$

The Sequence Dependent Term

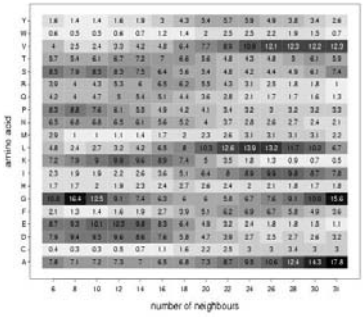
$$P(\text{sequence}|\text{structure}) \approx P_{env} \times P_{pair}$$

$$P_{env} = \prod_i P(aa_i|E_i)$$

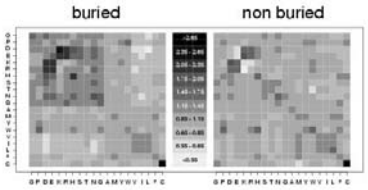
$$P_{pair} = \prod_{i<j} \frac{P(aa_i, aa_j|E_i, E_j, r_{ij})}{P(aa_i|E_i, r_{ij})P(aa_j|E_j, r_{ij})}$$



Hydrophobic Burial

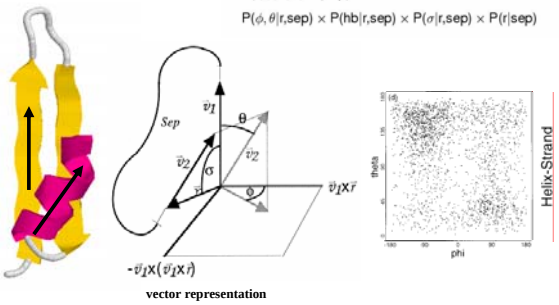


Residue Pair Interaction

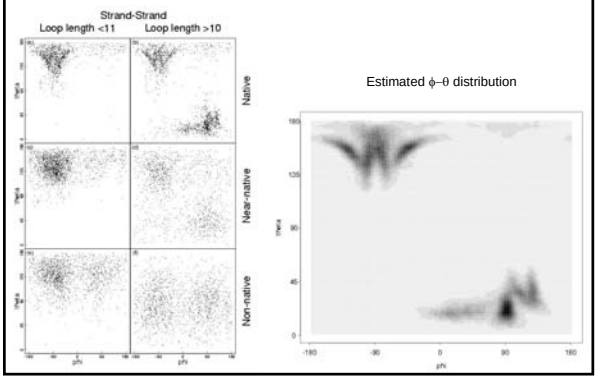


The Sequence Independent Term

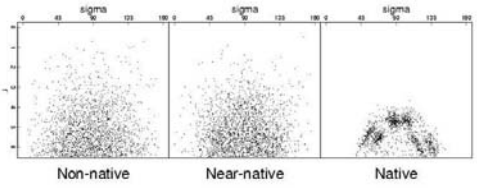
$$P(r,\phi,\theta,\sigma,\text{hb}|\text{sep}) \approx$$
$$P(\phi,\theta|r,\text{sep}) \times P(\text{hb}|r,\text{sep}) \times P(\sigma|r,\text{sep}) \times P(r|\text{sep})$$



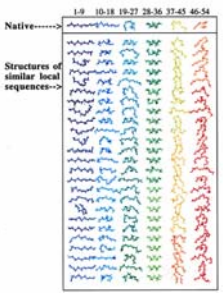
Strand Packing – Helps!

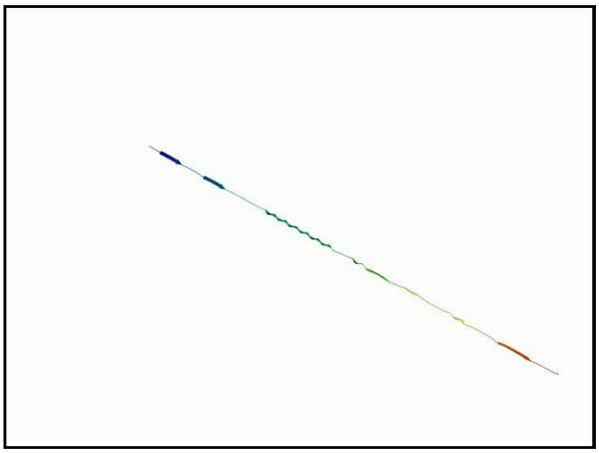


Sheer Angles – Help not!

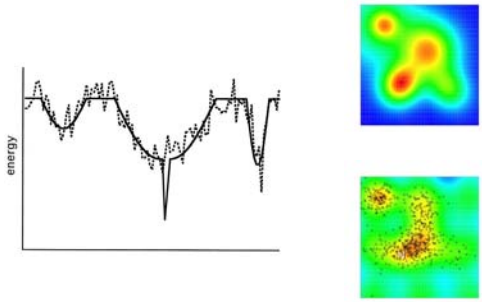


Fragment Selection

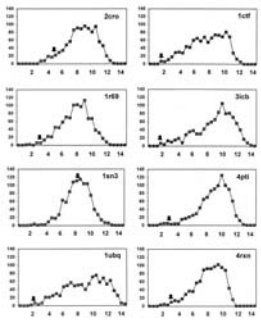




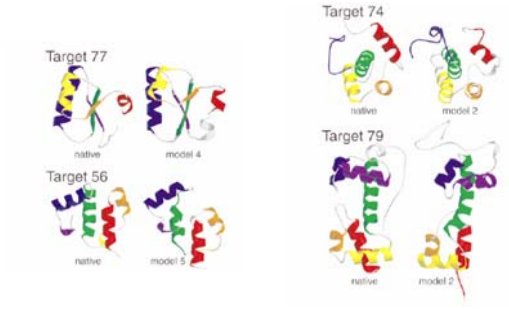
3D Clustering



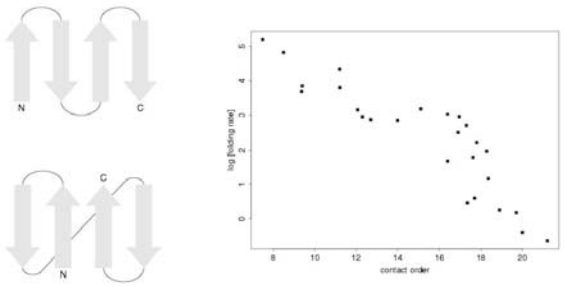
3D Clustering



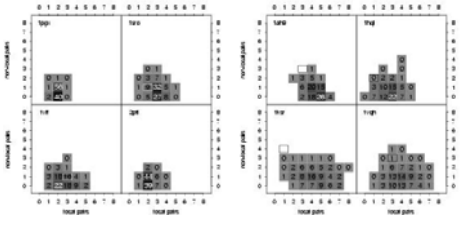
CASP3 Predictions



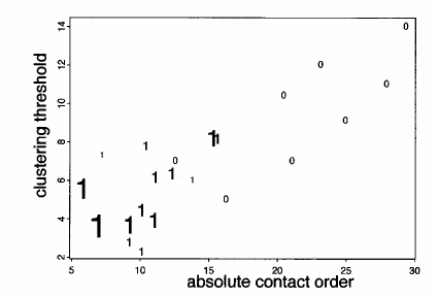
Contact Order



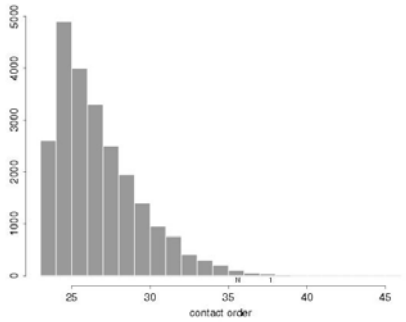
Contact Order



Clustering and Contact Order



Decoy Enrichment in CASP4

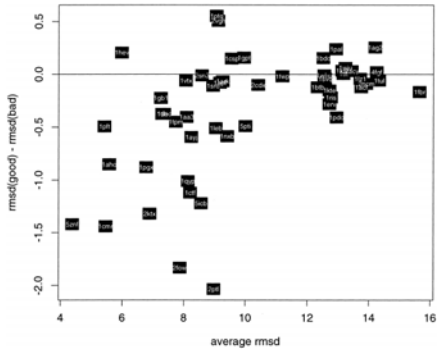


A Filter for Bad β-Sheets

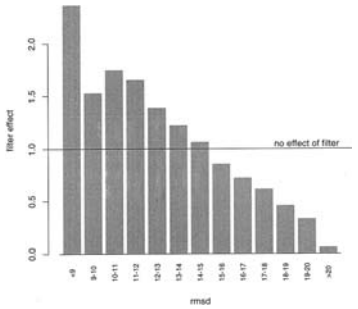
Many decoys do not have proper sheets. Filtering those out seems to enhance the rmsd distribution in the decoy set. Bad features we see in decoys include:

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

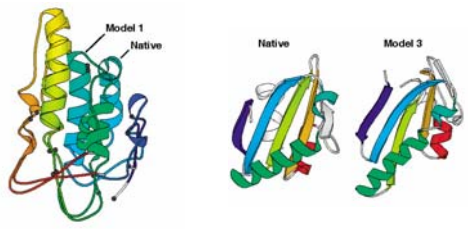
A Filter for Bad β-Sheets

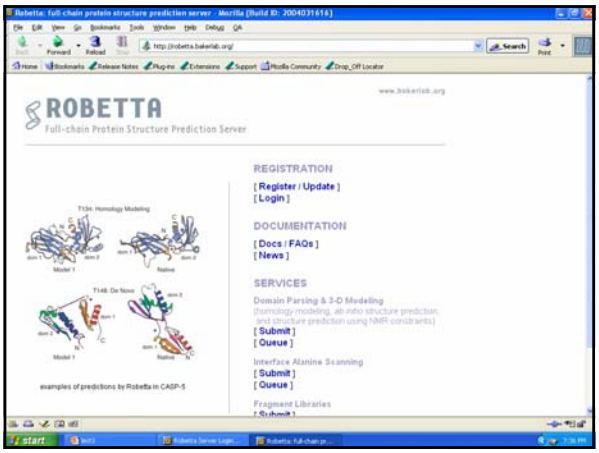


A Filter for Bad β-Sheets



Rosetta in CASP4

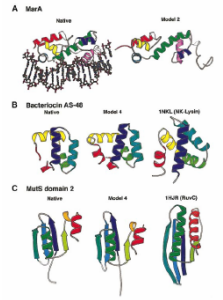




Applications and Other Uses of Rosetta

- Other uses of Rosetta:
  - Homology modeling.
  - Rosetta NMR.
  - Protein interactions (docking).
- Applications of Rosetta:
  - Functional annotation of genes.
  - Novel protein design.

Functional Inference



Collaborators

Collaborators = People who I troubled way more than I should have.

David Baker	University of Washington
Kevin Plaxco	UC Santa Barbara
Richard Bonneau	Institute for Systems Biology
Chris Bystroff	Rensselaer Polytechnic Institute
Dylan Chivian	University of Washington
Charles Kooperberg	Fred Hutchinson Cancer Research Center
Carol Rohl	UC Santa Cruz
Kim Simons	Harvard University
Charlie Strauss	Los Alamos National Laboratory
Jerry Tsai	Texas A&M

Rosetta Developers

Johns Hopkins Biostatistics