Protein Structure Prediction using ROSETTA

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Protein Folding vs Structure Prediction

- Protein folding is concerned with the process of the protein taking its three dimensional shape. The role of statistics is usually to support or discredit some hypothesis based on physical principles.
- 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 about the latter.

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 Humble Beginnings

- Kim Simons and David Baker tackle ab initio structure prediction (1995/96).
- A bit later, Charles Kooperberg and Ingo Ruczinski join the project.
- Two publications appear:
  - Simons et al (1997): Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, JMB 268, pp 209-25.
  - Simons et al (1999): Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, Proteins 34, pp 82-95.
- With the help of Richard Bonneau and Chris Bystroff, Rosetta is used for the first time on unknown targets in CASP3 (1998).

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

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

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

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

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

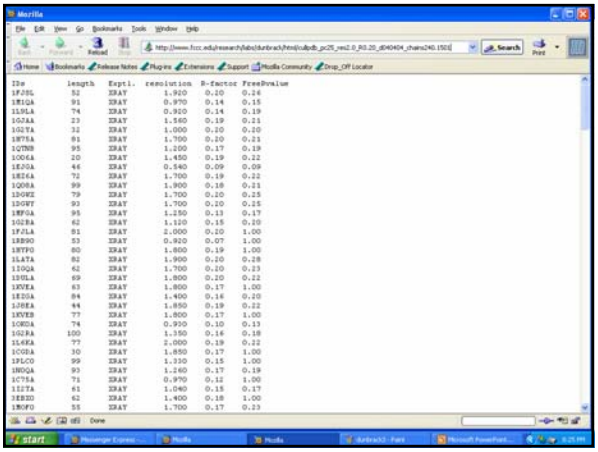
$$\ldots$$

The Sequence Dependent Term

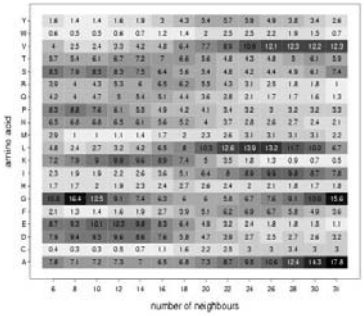
$$P(\text{sequence}|\text{structure}) \approx P_{env} \times P_{pair}$$

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

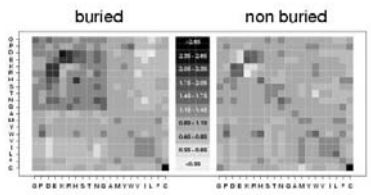
$$P_{pair} = \prod_{i<j} \frac{P(aa_i, aa_j|E_i, E_j, r_{ij})}{P(aa_i|E_i, r_{ij})P(aa_j|E_j, r_{ij})}$$



Hydrophobic Burial

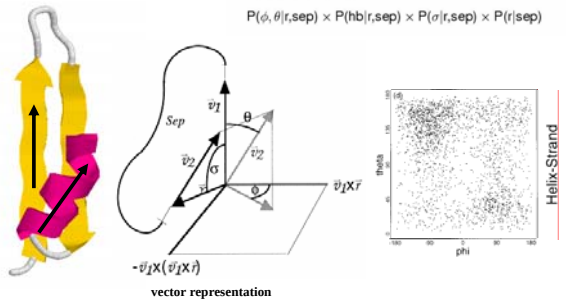


Residue Pair Interaction



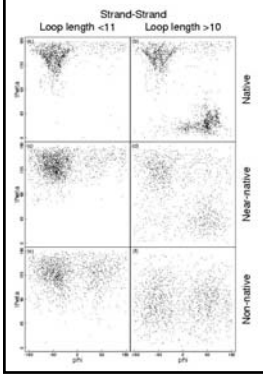
The Sequence Independent Term

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

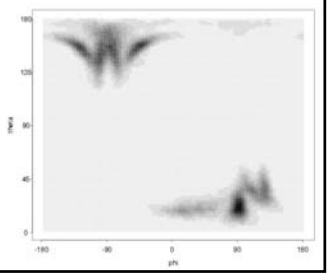


vector representation

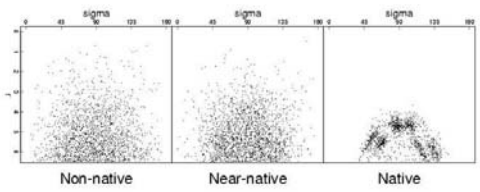
Strand Packing – Helps!



Estimated ϕ–θ distribution



Sheer Angles – Help not!



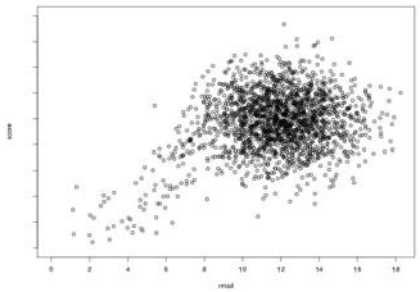
The Model

$$P(\text{structure}) = P_A^{w_A} P_B^{w_B} P_C^{w_C}, \quad w_X > 0.$$

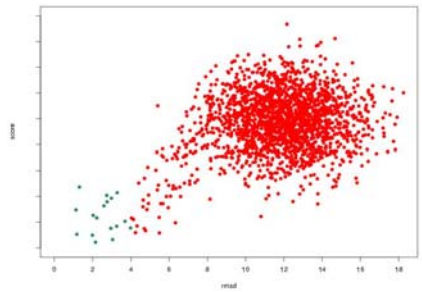
$$-\log P(\text{structure}|\text{sequence}) \propto$$
$$-\log P(\text{sequence}|\text{structure}) - \log P(\text{structure})$$

$$g(\text{rmsd}) = w_{\text{protein}} + w_{\text{HS}} \log P_{\text{HS}} + w_{\text{SS}} \log P_{\text{SS}} + w_{\text{vdw}} \text{VdW} +$$
$$w_{\text{sheet}} \log P_{\text{sheet}} + w_{\text{seq}} (\log P_{\text{env}} + \log P_{\text{pair}})$$

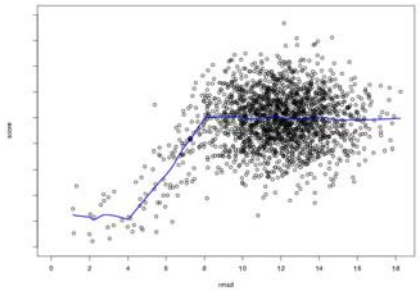
Parameter Estimation



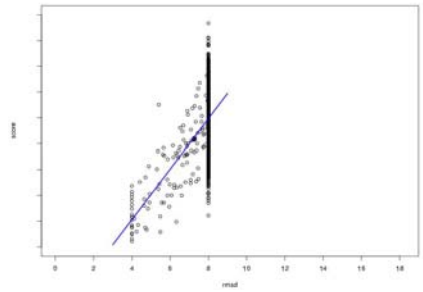
Parameter Estimation



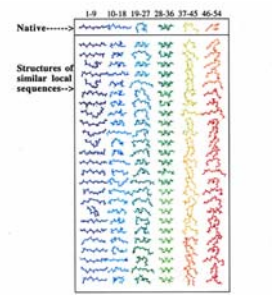
Parameter Estimation

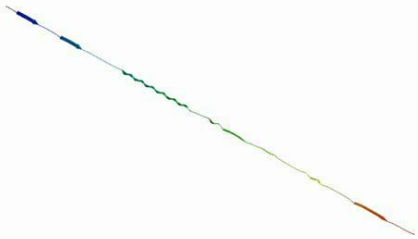


Parameter Estimation

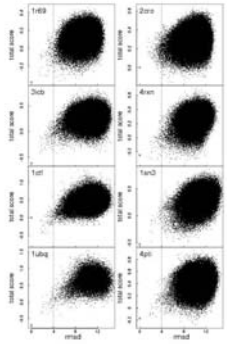


Fragment Selection

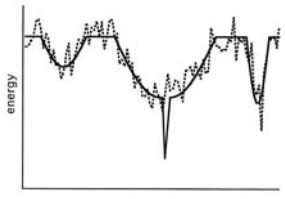


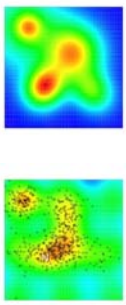


Validation Data Set

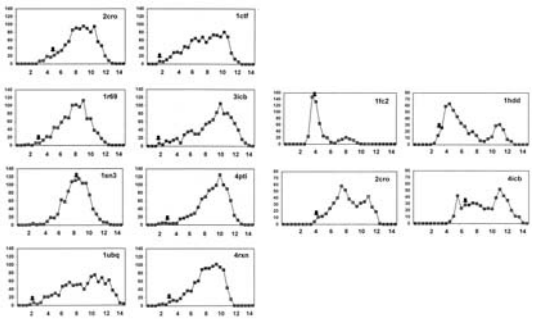


3D Clustering

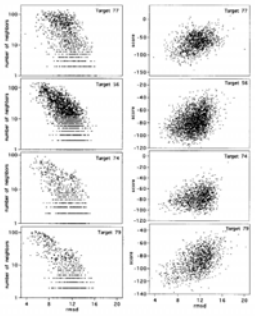




3D Clustering



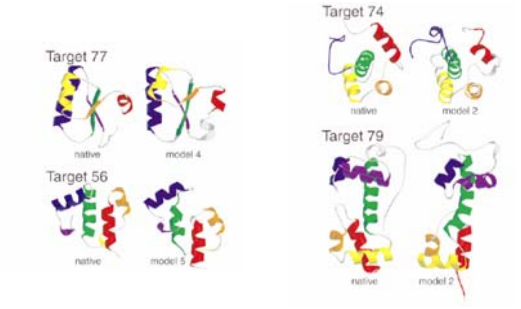
3D Clustering in CASP3



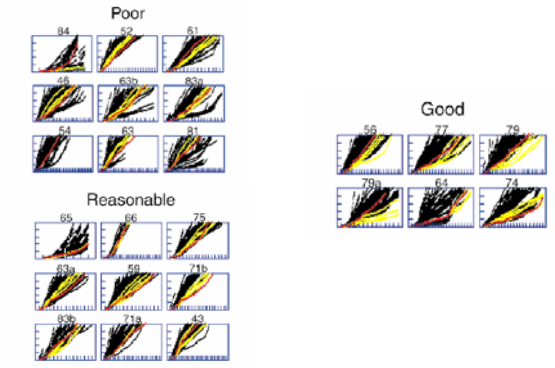
CASP3 Protocol

- 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.

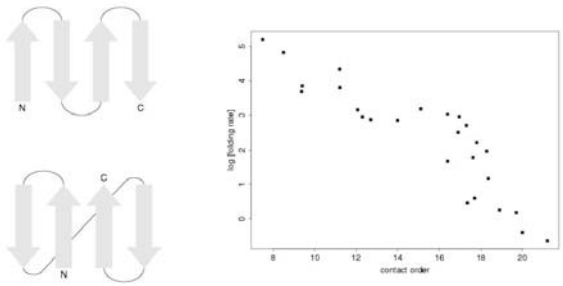
CASP3 Predictions



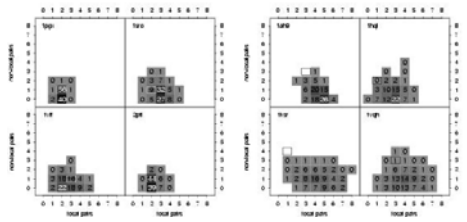
CASP3 Results



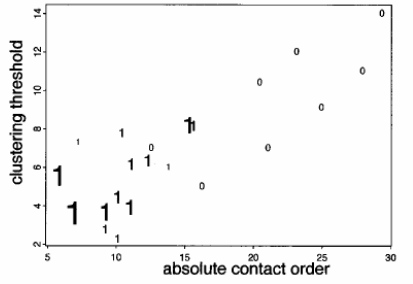
Contact Order



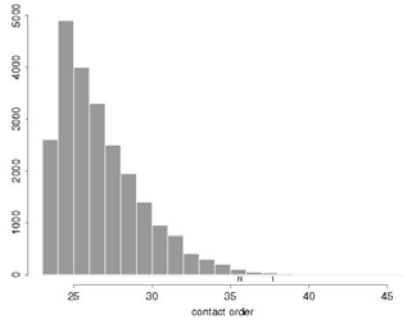
Contact Order



Clustering and Contact Order



Decoy Enrichment in CASP4

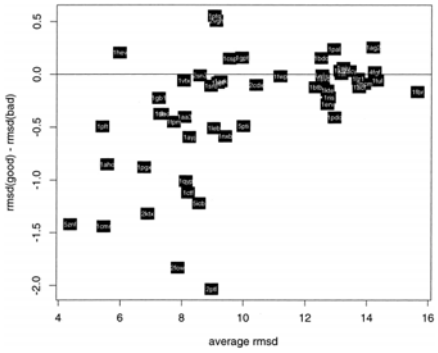


A Filter for Bad β-Sheets

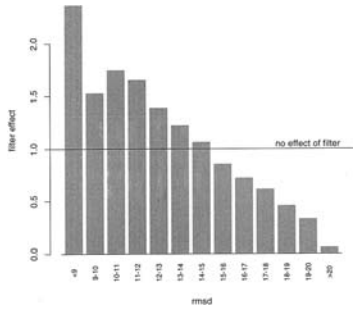
Many decoys do not have proper sheets. Filtering those out seems to enhance the rmsd distribution in the decoy set. Bad features we see in decoys include:

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

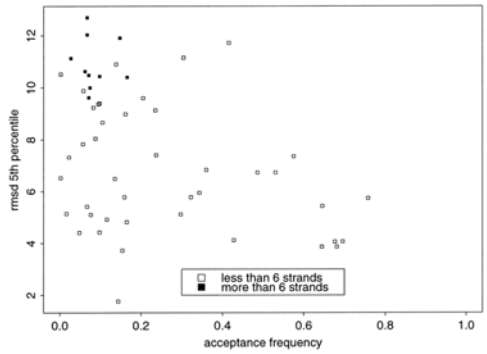
A Filter for Bad β-Sheets



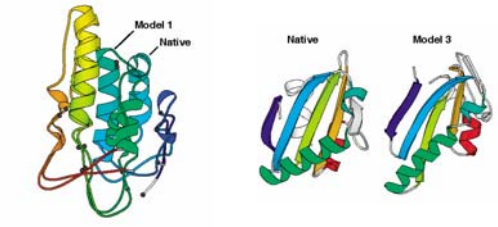
A Filter for Bad β-Sheets

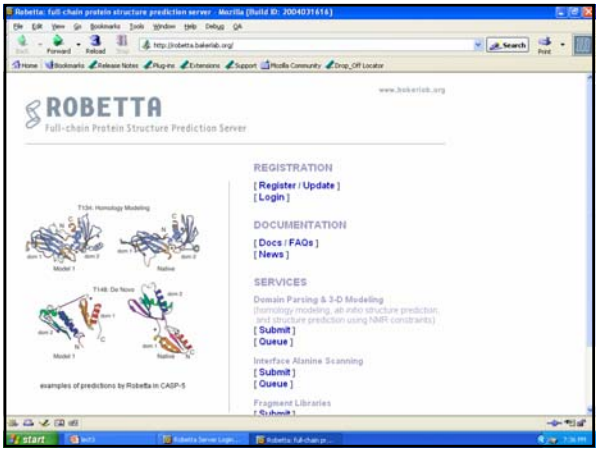


A Filter for Bad β-Sheets



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.

Collaborators

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

David Baker	University of Washington
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