

Protein Folding and Structure Prediction

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Protein Folding and Structure Prediction

- Protein **fold**ing is concerned with the process of the protein taking its three dimensional shape. The role of statistics is often 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 a bit about both.

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: | Sequence independent: |
| • hydrophobic burial | • helix-strand packing |
| • residue pair interaction | • strand-strand packing |
| | • sheet configurations |
| | • vdW interactions |

The Sequence Dependent Term

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

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

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

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

...

The Sequence Dependent Term

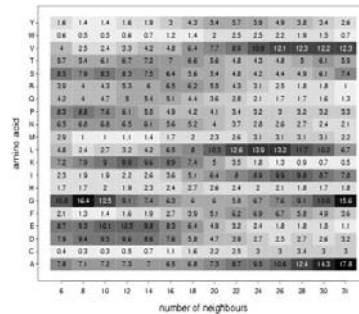
$$P(\text{sequence} | \text{structure}) \approx P_{\text{env}} \times P_{\text{pair}}$$

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

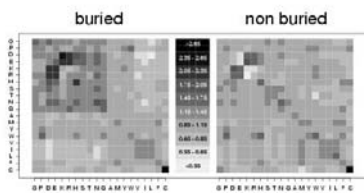
$$P_{\text{pair}} = \prod_{i < j} \frac{P(aa_i, aa_j | E_i, E_j, r_{ij})}{P(aa_i | E_i)P(aa_j | E_j, r_{ij})}$$

ID	length	ExpL	resolution	B-factor	FreeRoiLue
1F29G	52	2.8AT	1.920	0.20	0.24
1R1QA	91	2.8AT	0.970	0.14	0.15
1L6LA	74	2.8AT	0.920	0.14	0.19
1G2AA	23	2.8AT	1.580	0.19	0.21
1G7TA	32	2.8AT	1.000	0.20	0.20
1R7XA	81	2.8AT	1.700	0.20	0.21
1G7NO	95	2.8AT	1.200	0.17	0.19
1G7GA	100	2.8AT	1.450	0.19	0.22
1E2DA	46	2.8AT	0.540	0.09	0.09
1E2EA	70	2.8AT	1.700	0.19	0.22
1G0RA	99	2.8AT	1.900	0.18	0.21
1D0FE	79	2.8AT	1.700	0.20	0.23
1D0FF	93	2.8AT	1.700	0.20	0.25
1R7DA	95	2.8AT	1.250	0.13	0.17
1G7BA	62	2.8AT	1.120	0.15	0.20
1F2LA	91	2.8AT	2.000	0.20	1.00
1R9PO	53	2.8AT	0.920	0.07	1.00
1R7PD	80	2.8AT	1.800	0.19	1.00
1L1TA	82	2.8AT	1.900	0.20	0.28
1G0LA	62	2.8AT	1.700	0.20	0.23
1R7LA	69	2.8AT	1.800	0.20	0.22
1R7EA	83	2.8AT	1.800	0.17	1.00
1E2DA	84	2.8AT	1.400	0.14	0.20
1R7EA	84	2.8AT	1.800	0.19	0.22
1R7EB	77	2.8AT	1.800	0.17	1.00
1G0GA	74	2.8AT	0.970	0.10	0.13
1G7BA	74	2.8AT	1.150	0.14	0.18
1L6KA	77	2.8AT	2.000	0.19	0.22
1G0RA	100	2.8AT	1.800	0.17	1.00
1F2OD	99	2.8AT	1.130	0.15	1.00
1R0GA	99	2.8AT	1.240	0.17	0.19
1C7EA	71	2.8AT	0.970	0.12	1.00
1L1TA	61	2.8AT	1.040	0.15	0.17
1E2DD	62	2.8AT	1.400	0.18	1.00
1R0FD	55	2.8AT	1.700	0.17	0.23

Hydrophobic Burial



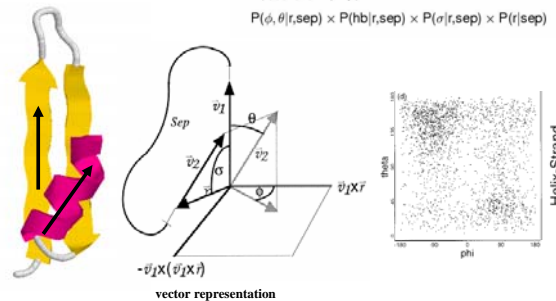
Residue Pair Interaction



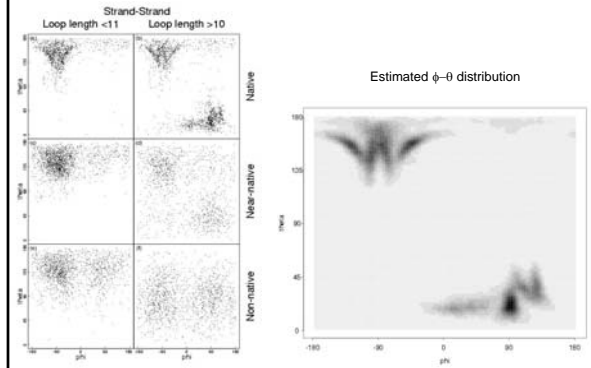
The Sequence Independent Term

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

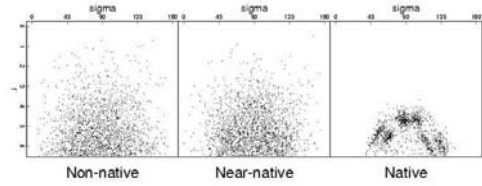
$$P(\phi, \theta | r, \text{sep}) \times P(hb | r, \text{sep}) \times P(\sigma | r, \text{sep}) \times P(r | \text{sep})$$



Strand Packing – Helps!

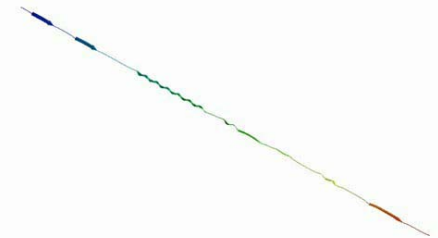
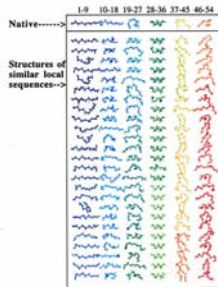


Shear Angles – Help not!

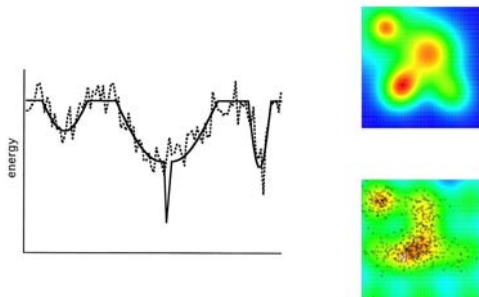


Simons et al (1999): Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, Proteins 34, pp 82-95.

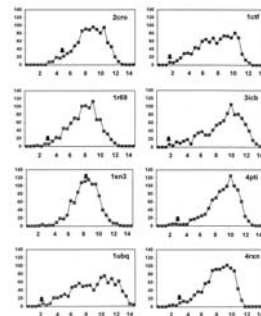
Fragment Selection

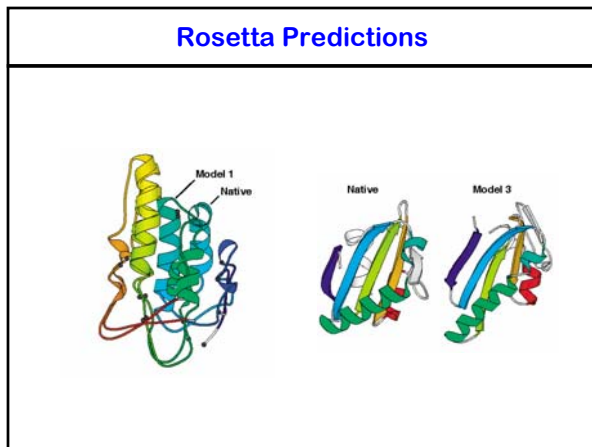
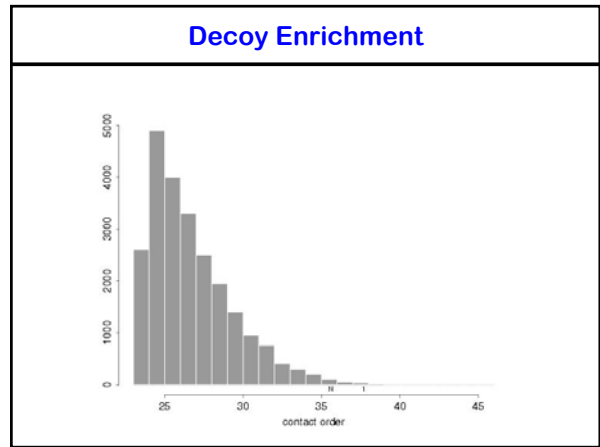
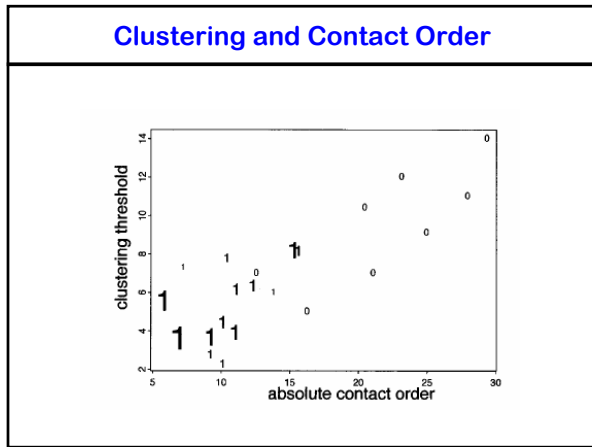
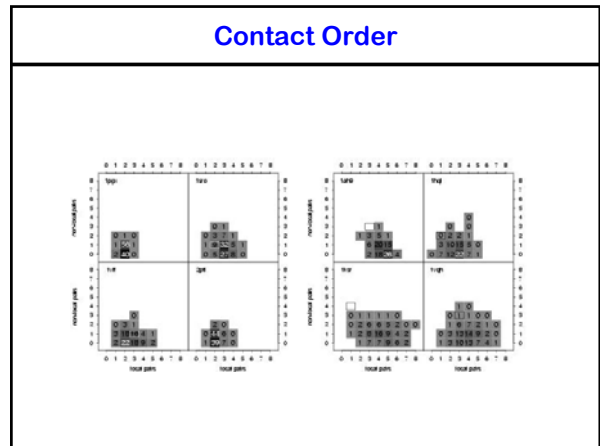
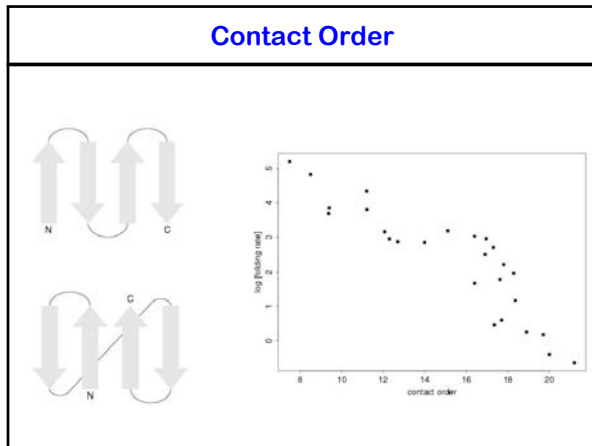


3D Clustering



3D Clustering





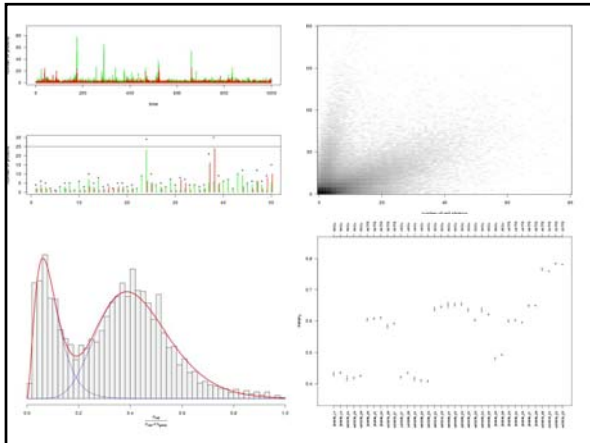
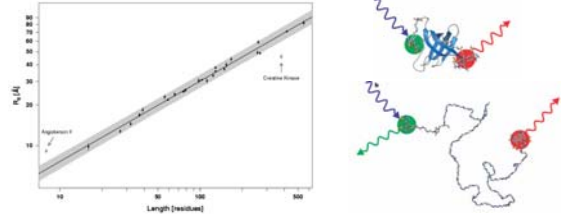
The screenshot shows the Rosetta Full-chain Protein Structure Prediction Server website. The page includes navigation links, registration options, documentation, and services. The website is titled 'ROBETTA Full-chain Protein Structure Prediction Server' and features a navigation menu with links for Home, Downloads, Plugins, Extensions, Support, Media Community, and Drop Off Location. The main content area includes sections for REGISTRATION (Register / Update, Login), DOCUMENTATION (Docs / FAQs, News), and SERVICES (Domain Parsing & 3-D Modeling, Interface, Alanine Scanning, Fragment Libraries). There are also images of protein structures and examples of predictions by Rosetta in CASP-5.

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.

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?



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