Computergestützte Strukturbiologie
(Strukturelle Bioinformatik)

New fold prediction

Sommersemester 2009

Peter Güntert

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modeling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

Scheme of protein structure predicition

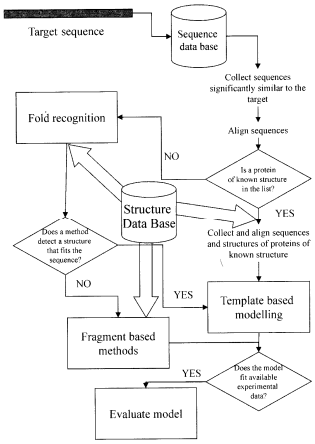


Figure 4.1 A guide to protein-structure prediction. The first step is always a search in the protein sequence database. Comparative modeling should be used when a protein of known structure sharing sequence similarity with the protein under examination is present in the database. If this is not so, fold-recognition methods should be applied and, should they fail, the user should resort to new fold or fragment-based methods. Note the central role played by the structure database in all these heuristic methods.

CASP: Fragment-based predictions

Figure 6.2 Some examples of fragment-based predictions submitted to CASP experiments.

Fragment-based approaches

- **Rosetta (David Baker)**
- **Fragfold (David Jones)**

Degenerate sequence-to-structure relationship

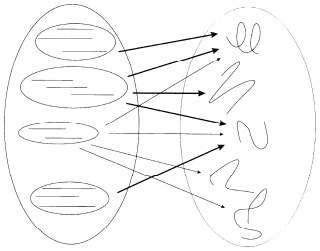


Figure 6.1 The local sequence-to-structure relationship is degenerate. For some recurring local structures, however, a correlation can be identified. In the figure, the left part depicts the space of sequence fragments. Each line represents a sequence and two similar sequences are closer to each other. Some groups of sequences will show preference for a subset of local structures (indicated by the thicker lines in the figure) while others will be less specific.

Steps of fragment-based structure prediction

- Split sequence into fragments
- For each fragment, search the database of known structures for regions with a similar sequence ("neighbors")
- Use an optimization technique to find the best combination of fragments

Fragment search

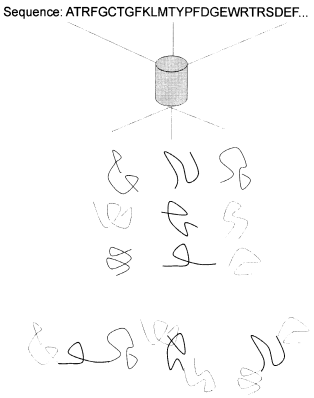


Figure 6.3 Schematic explanation of the first steps of the Rosetta method. The query sequence is split in fragments nine amino acids long. Each fragment sequence is used to search for similar fragments among the sequences of proteins of known structure. Next, the fragments are joined.

Distance between target and template fragment sequences in Rosetta

$$dist = \sum_{i=1}^{9} \sum_{aa=1}^{20} \left| S(aa,i) - X(aa,i) \right|$$

- $S(aa,i)$ and $X(aa,i)$ are the frequencies of the amino acid in position i = 1,…,9 of the target and template nine-residue sequences or alignments.
- The 25 closest "neighbors" from the database of known 3D protein structures are chosen.

ROSETTA: Distance between fragments

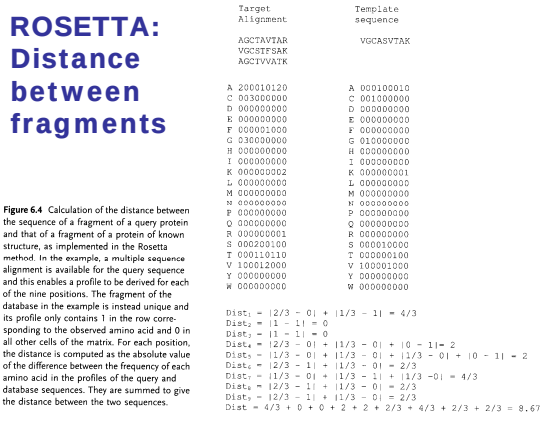
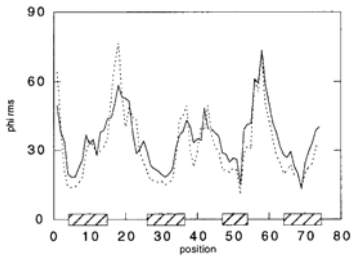


Figure 6.4 Calculation of the distance between the sequence of a fragment of a query protein and that of a fragment of a protein of known structure, as implemented in the Rosetta method. In the example, a multiple sequence alignment is available for the query sequence and this enables a profile to be derived for each of the nine positions. The fragment of the database in the example is instead unique and its profile only contains 1 in the row corresponding to the observed amino acid and 0 in all other cells of the matrix. For each position, the distance is computed as the absolute value of the difference between the frequency of each amino acid in the profiles of the query and database sequences. They are summed to give the distance between the two sequences.

Structural variability and similarity to true structure for fragments



Correlation between structural variability and similarity to the true structure in nearest neighbor sets. Variability in phi and rmsd from the native structure for entire calbindin sequence. Each position is represented by the segment with the lowest variability. The four helices in the native structure are indicated by hatched bars.

Simulated annealing in Rosetta

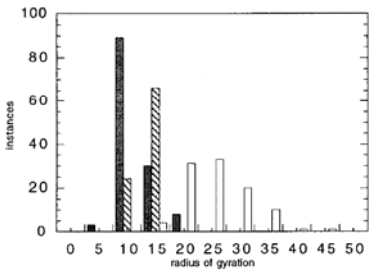
- Simplified model: main-chain heavy atoms, and C^{β}
- Torsion angles as degrees of freedom
- For each 9-residue sequence fragment, find 25 nearest sequence neighbors
- Start from extended chain
- In each Monte Carlo step, substitute the dihedral angles of a randomly chosen neighbor at a randomly chosen position for those of the current position
- Conformations are initially evaluated using Bayesian probabilities. In subsequent cycles, knowledge based potentials are used.

Radius of gyration

For a rigid body consisting of *n* particles with mass m_i located at distance r_i from the center of mass, the radius of gyration is defined by

$$R_G = \sqrt{\sum_{i=1}^{n} m_i r_i^2 \Big/ \sum_{i=1}^{n} m_i}$$

Radii of gyration of simulated and native structures



Comparison of the radii of gyrations of simulated and native structures. 100 structures were generated for chains of 100 residues by splicing together protein fragments using either no scoring function (open bars), or the square of the radius of the gyration as the scoring function (hatched bars). Histograms were computed using 5 Å bins. The distribution of radii of gyrations for the small (50 to 150 residue) proteins in the pdbselect 25 set is shown for comparison (filled bars).

PROTEINS: Structure, Function, and Genetics 53:457–468 (2003)

Rosetta Predictions in CASP5: Successes, Failures, and Prospects for Complete Automation

Philip Bradley[1], Dylan Chivian[1], Jens Meiler[2], Kira M.S. Misura[1], Carol A. Rohl[1], William R. Schief[1], William J. Wedemeyer[2], Ora Schueler-Furman, Paul Murphy, Jack Schonbrun, Charles E.M. Strauss, and David Baker[2]*

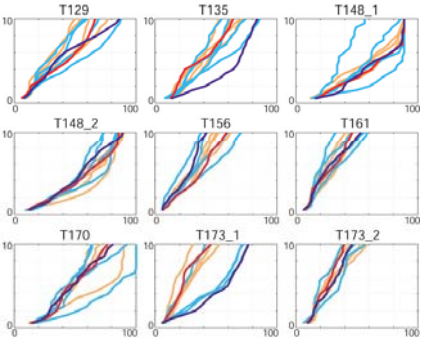
Department of Biochemistry, University of Washington, Seattle, Washington

ABSTRACT We describe predictions of the structures of CASP5 targets using Rosetta. The Rosetta fragment insertion protocol was used to generate models for entire target domains without detectable sequence similarity to a protein of known structure and to build long loop insertions (and N-and C-terminal extensions) in cases where a structural template was available. Encouraging results were obtained both for the de novo predictions and for the long loop insertions; we describe here the successes as well as the failures in the context of current efforts to improve the Rosetta method. In particular, de novo predictions failed for large proteins that were incorrectly parsed into domains and for topologically complex (high contact order) proteins with swapping of segments between domains. However, for the remaining targets, at least one of the five submitted models had a long fragment with significant similarity to the native structure. A fully automated version of the CASP5 protocol produced results that were comparable to the human-assisted predictions for most of the targets, suggesting that automated genomic-scale, de novo protein structure prediction may soon be worthwhile. For the three targets where the human-assisted predictions were significantly closer to the native structure, we identify the steps that remain to be automated. Proteins 2003;53:457–468.

ROSETTA Flowchart

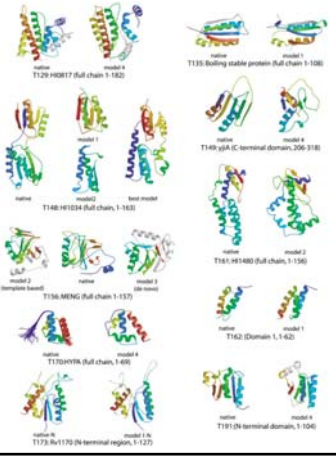
Flowchart of general Rosetta protocol. Starting with obtaining the target sequence, steps for target identification, decoy generation, and selection are outlined for both the template-based approach (used for targets with homologous structures available in the PDB) and for the fragment insertion approach (used for new fold and difficult fold recognition targets).

Global distance test



Global distance test (GDT) plots for selected targets comparing the CASP5 Rosetta submissions with predictions made with a fully automated version of the same protocol. Cyan (models 2–5) and dark blue (model 1) represent the CASP5 submissions, orange (models 2–5) and red (model 1) represent models made with a fully automated version of the CASP5 protocol (see Materials and Methods). The *y* axis represents a C^{α} RMSD cutoff under which to fit the model to the native structure, and the *x* axis represents the percentage of the model that will fit below that cutoff value.

ROSETTA results in CASP5



Ribbon diagrams of predictions made by using the fragment insertion approach. The native structure and best submitted model are shown colored from the N-terminus (blue) to C-terminus (red). For T148, the best generated model is also shown, and for T156, both template-based and fragment insertion based models are shown. For targets T173, T135, T156, and T191, colored regions deviate from the native structure by <4 Å, and gray regions deviate by >4 Å. For targets T129 and T156, colored regions deviate from the native structure by <6 Å C^{α} RMSD, whereas the gray regions deviate by >6 Å.

Rosetta: CASP5 Targets predicted with fragment insertion

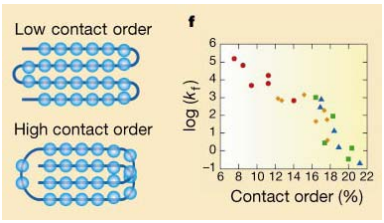
TABLE I. Summary of Results for CASP5 Targets Predicted With Fragment Insertion by the Rosetta Algorithm

Name[a]	class[b]	cr[c]	α/β[d] [%]	Length	Number of amino acids with an RMSD below 4 Å/6 Å[e] Human[f]	Standard[g]	Best[h]
129	nf	30.1	64/0	170	108/153	87/116	111/159
149_2	nf	34.6	23/35	116	52/71	44/62	76/92
161	nf	33.7	53/11	154	45/83	57/79	55/95
162_3	nf	24.6	36/38	168	58/79	—	68/95
181	nf	25.1	30/18	111	35/59	52/65	65/103
146_1	frfnf	31.4	28/25	107	28/51	—	42/54
146_2	frfnf	29.2	23/26	89	45/60	—	70/76
146_3	frfnf	21.9	0/10	56	27/31	—	26/39
146_4	frfnf	9.2	19/0	47	23/30	—	33/40
170	frfnf	16.3	60/0	69	64/67	60/64	66/68
172_2	frfnf	24.7	54/0	101	52/62	—	90/101
173	frfnf	55.1	35/15	287	127/149	69/84	127/149
186_3	frfnf	5.2	0/33	36	28/32	—	—
187_1	frfnf	42.7	42/19	187	57/87	—	76/114
135	frfs	31.7	34/30	106	83/98	54/64	94/105
148_1	frfs	23	38/32	71	62/64	57/62	65/66
148_2	frfs	23.1	41/27	91	73/74	75/77	80/90
162_1	frfs	13.1	76/0	56	56/56	—	56/56
162_2	frfs	16.3	0/25	51	33/43	—	38/49
187_2	frfs	38	38/14	227	51/85	—	85/120
191_1	frfs	28.1	45/21	139	89/100	85/98	102/105
174_1	frfs	47.2	38/26	197	54/64	—	52/67
174_2	frfs	34.6	36/25	155	44/47	—	47/62
156	frfs	46.4	18/32	156	59/89	71/88	81/107

[a] CASP identification number.
[b] Assessor classification (nf, new fold; frfnf, fold recognition/new fold; frfs, fold recognition analog; frfs, fold recognition homologue).
[c] Contact order.
[d] Fraction of amino acids in α-helix or β-strand conformation.
[e] The number of residues (Cα atoms) of the model superimposable (using a variant of MaxSub[36] which uses RMSD as the threshold) on the native structure within a 4 Å RMSD cutoff (left) and within a 6 Å cutoff (right).
[f] Best Rosetta model submitted for CASP5.
[g] Best fully automated prediction using standard CASP5 protocol.
[h] Best ROSETTA model in decoy population before filtering.
[i] The best submission for T187 was a comparative model based on template 1 vpe with 57 and 116 residues aligned within 4 Å and 6 Å, respectively.
[j] The best submission for T156 was a comparative model based on template 1 dik with 78 and 107 residues aligned within 4 Å and 6 Å, respectively.

Contact order

The relative contact order is the average separation along the sequence of residues in physical contact in a folded protein, divided by the length of the protein.



The contact order is strongly correlated with the folding rate of a protein.

Toward High-Resolution de Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, David Baker*

The prediction of protein structure from amino acid sequence is a grand challenge of computational molecular biology. By using a combination of improved low- and high-resolution conformational sampling methods, improved atomically detailed potential functions that capture the jigsaw puzzle–like packing of protein cores, and high-performance computing, high-resolution structure prediction (<1.5 angstroms) can be achieved for small protein domains (<85 residues). The primary bottleneck to consistent high-resolution prediction appears to be conformational sampling.

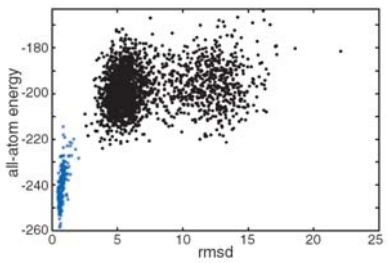
Science 309, 1868–1871 (2005)

Prediction results

Table 1. Benchmark proteins and results. Protein Data Bank (PDB) (18) or Structural Classification of Proteins (SCOP) (19) ID is given in column 1 (10). Protein length, fraction alpha helix, and fraction beta strand are given in columns 2 to 4. C_α-RMSD values for the model with the lowest all-atom energy in rounds 1 and 2 are given in columns 5 and 6, respectively (20). RMSD values calculated over all heavy atoms in the protein core (21) are given in parentheses. Column 7 reports the best C_α-RMSD of the centers of the largest five clusters when the low-energy models from round 1 are clustered.

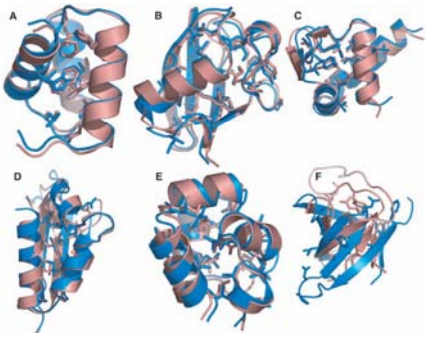
ID	L	%α	%β	Round 1	Round 2	Cluster	Protein name
1b72A	49	69	0	0.8 (0.8)	1.1 (0.9)	1.0	Hox-B1 homeobox protein
1bfrA	59	5	40	11.1 (9.0)	10.8 (8.5)	10.9	Fyn tyrosine kinase
1tif_	59	22	37	5.3 (2.3)	4.1 (2.8)	3.8	IF3-N
2reb_2	60	61	20	1.2 (0.9)	2.1 (1.6)	1.3	RecA
1r69_	61	63	0	2.1 (2.4)	1.2 (1.5)	1.7	434 repressor
1csp_	67	4	53	5.1 (4.5)	4.7 (4.2)	5.1	Cold-shock protein
1dZA_	69	46	33	2.6 (2.3)	2.6 (2.2)	1.9	RNA binding protein A
1n0uA4	69	43	24	9.9 (8.3)	10.2 (8.1)	2.7	Elongation factor 2
1mla_2	70	34	37	8.4 (7.3)	8.7 (8.1)	7.2	Malonyl-CoA ACP transacylase
1af7_	72	72	0	10.1 (7.9)	10.4 (8.1)	1.7	Cher domain 1
1ogwA_	72	26	33	2.7 (2.3)	1.0 (1.0)	2.6	Ubiquitin
1dcjA_	73	31	27	3.2 (2.2)	2.5 (2.4)	2.0	Yhhp
1dtjA_	74	39	27	1.0 (0.8)	1.2 (0.9)	1.8	KH domain of Nova-2
1o2fB_	77	38	27	10.1 (8.7)	N/A	10.3	Glucose-permease IIBC
1mkyA3	81	32	24	3.2 (3.6)	6.3 (6.1)	3.7	Enga
1tig_	88	35	35	4.1 (4.2)	3.5 (3.4)	2.4	IF3-C

Free-energy landscape for barstar



Free-energy landscape for the small protein barstar (PDB code 1a19). Rosetta all-atom-energy (y axis) is plotted against C^α-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). The free-energy function includes the entropic contribution to the solvation free energy but not the configurational entropy.

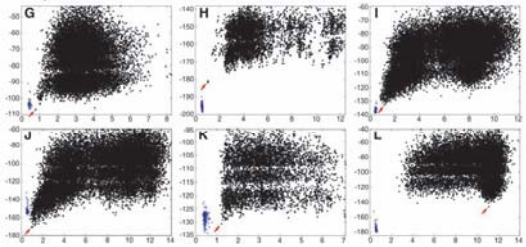
High-resolution de novo structure predictions



Superposition of low-energy models (blue) with experimental structures (red) showing core side chains.

A: Hox-B1
B: Ubiquitin
C: RecA
D: KH domain of Nova-2
E: 434 repressor
F: Fyn tyrosine kinase

Energy vs. accuracy



Plots of C^{α}-RMSD (x axis) against all atom energy (y axis) for refined natives (blue points) and the de novo models (black points). Red arrows indicate the lowest energy de novo models.

Protein design

- **Inverse protein folding problem**
- **Design the sequence of a protein that will fold into a given 3D structure.**
- **Structure can be that of an existing protein ("sequence redesign") or a completely new fold, not yet observed.**

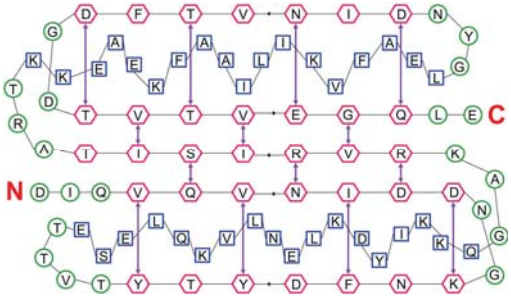
Design of a Novel Globular Protein Fold with Atomic-Level Accuracy

Brian Kuhlman,[1*†] Gautam Dantas,[1*] Gregory C. Ireton,[4] Gabriele Varani,[1,2] Barry L. Stoddard,[4] David Baker[1,3‡]

A major challenge of computational protein design is the creation of novel proteins with arbitrarily chosen three-dimensional structures. Here, we used a general computational strategy that iterates between sequence design and structure prediction to design a 93-residue α/β protein called Top7 with a novel sequence and topology. Top7 was found experimentally to be folded and extremely stable, and the x-ray crystal structure of Top7 is similar (root mean square deviation equals 1.2 angstroms) to the design model. The ability to design a new protein fold makes possible the exploration of the large regions of the protein universe not yet observed in nature.

Science 302, 1364–1368 (2002)

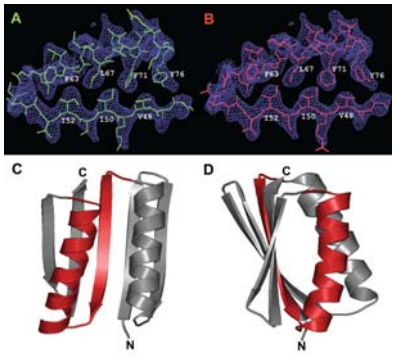
Designed globular protein fold



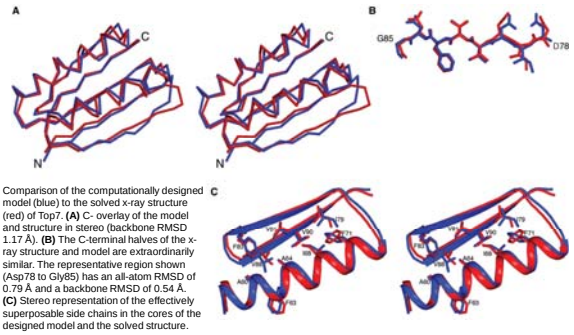
A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

Top7 structure

Schematic representation of Top7 in unbiased SAD density. (A and B) Stick representations of residues 46 to 76 from the computationally designed Top7 (left, green) and from the 2.5 Å x-ray structure (right, red) are shown in unbiased density (blue). The map was generated from SAD phasing from a single SeMet-substituted variant of Top7, followed by density modification. (C and D) Ribbon diagrams of Top7 with residues 46 to 76 highlighted in red. The two diagrams are related by a 90° rotation around the vertical axis.



Designed and X-ray structure of Top7



Comparison of the computationally designed model (blue) to the solved x-ray structure (red) of Top7. **(A)** C- overlay of the model and structure in stereo (backbone RMSD 1.17 Å). **(B)** The C-terminal halves of the x-ray structure and model are extraordinarily similar. The representative region shown (Asp78 to Gly85) has an all-atom RMSD of 0.79 Å and a backbone RMSD of 0.54 Å. **(C)** Stereo representation of the effectively superposable side chains in the cores of the designed model and the solved structure.

Literatur

- Anna Tramontano: *Protein Structure Prediction, Wiley-VCH*, 2006.
- K. T. Simons et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268, 209-225 (1997) [Rosetta algorithm]
- P. Bradley et al. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Proteins* 53, 457-468 (2003) [Rosetta applications]
- D. T. Jones et al. Prediction of novel and analogous folds using fragment assembly and fold recognition *Proteins* Suppl 7, 143-151 (2005) [Fragfold]
- P. Bradley et al. *Science* 309, 1868-1871 (2005) [High-resolution de novo structure prediction]
- B. Kuhlman et al. Design of a novel globular protein fold with atomic-level precision. *Science* 302, 1364-1368 (2003) [Successful protein design]