Protein Structure Prediction

Roland Schulz

September 28, 2007

1 Introduction

The structure of an ordered protein is essential for the understanding of its function. Even though the number of experimental available proteins is exponentially increasing over the last years, there is a large number of proteins with unknown fold and without an obvious homology with any protein which has been resolved. Refolding experiments show that the protein sequence defines a unique native fold which is, in most cases, the free energy minimum. In theory, this free energy minimum can be computed from quantum mechanics and thus predict the structure from the sequence. In practice, ab initio and molecular dynamics (MD) methods are too slow or too inaccurate. Thus, the best de novo prediction methods use mainly statistical information from known structures [Rohl et al., 2004b].

The CASP competition shows the progress of the different prediction methods in the last decade. The most accurate prediction method so far is the template, or homology modeling approach, which predicts the structure by a comparison to a similar sequence. For two thirds of sequences, a similar sequence can be found and thus the structure can be predicted by homology modeling with good precision for those with less than 300 residues. Even though it is likely that most of the remaining third also has known folds, homology modeling does not work for sequence difference larger than 20 percent and thus de novo methods are needed for those [Zhou et al., 2007]. A group of methods, called "threading", were developed which used structural information from many less closely related structures. However, they are less accurate than modern fragment methods and are thus only used as a component in some software [Moult, 2005].

The idea of the fragment assembly strategy is that a local sequence has a high probability for one or few specific local structures and that the whole structure depends on the interplay of the most likely local structures and non-local interactions between them. Rosetta [Baker and Sali, 2001, Rohl

et al., 2004b], identified by CASP as the most accurate de novo prediction programs [Jauch et al., 2007], uses this strategy. It computes the most likely local structures from a databank of known structures, combines them, approximates the non-local interactions with a scoring function and minimizes it with a Monte Carlo simulated annealing search. As a result, Rosetta is able to predict low- to moderate-accuracy models with a 3-6ÅCα root mean square deviation (RMSD). This prediction has been shown to be useful to gain biological insight. The models often have a correct global topology and correctly identified secondary structure. As well, the functional residues often cluster to an active site [Rohl et al., 2004b]. Rosetta was also able to predict the first close to atomic-level structure only from the sequence, which can be seen in fig. 1. The method is fast enough to be used in large scale prediction of hundreds of protein families [Bonneau et al., 2002]. However, the method does not give information about misfolding or folding pathways as molecular dynamics related approaches like Folding@Home [Pande, 2006].

2 Methods

Rosetta, a knowledge based prediction method, uses the Bayes statical theorem to compute the structure from the knowledge of the structure of short fragments. The Bayes theory states

$$P(structure|sequence) = P(structure)\frac{P(sequence|structure)}{P(sequence)}$$

The right-hand side properties can be computed from known structures. To predict the full structure from the so computed probabilities, several steps are required. A fragment library has to be built, the fragment structure has to be assigned and a scoring function has to be minimized. The require methods will now be described [Rohl et al., 2004b].

2.1 Fragment Library

The software uses 3 and 9 residue long fragments. For all overlapping fragments in the target sequence, the 200 most likely angles for 3 and 9 long fragments are computed from X-ray resolved structures. The matching fragments are found in the protein data bank (PDB) by a PSIBLAST search. They are ranked by minimized steric overlap, favorable torsion angles and secondary structure compatible with a secondary structure prediction by Psipred, SAM-T99 and JUFO.

2.2 Scoring Function

Two different scoring functions are available. One is more coarse-grained and thus faster to compute, but it is not as accurate as the other. The second function is all-atomic and thus more accurate, but not as fast to compute. The coarse-grained function only depends on the torsion angles of the backbone with the side chains described by a centroid located at the center of mass. The all-atomic description also depends on the rotamer of the side-chain. Both functions consist of many individual terms of which the full description is too long for this paper. They are summarized in the Table I and II from [Rohl et al., 2004b]. The references cited in the table, explaining and deriving these terms, are 7: [Bowers et al., 2000], 12: [Rohl et al., 2004a], 14: [Jr and Cohen, 1997], 15: [Kuhlman and Baker, 2000], 16: [Simons et al., 1997], 17: [Simons et al., 1999], 18: [Lazaridis and Karplus, 1999], 19: [Kortemme et al., 2003], 20: [Wedemeyer and Baker, 2003]. All terms in Table 1 use probabilities computed from the fragment library, except vdw and rg, which are geometrical formulas. All terms including vdw and rg can be easily computed only from the torsion angles. For the all-atomic function the LJ and solv terms are geometric and the ref term depends on a value per amino acid. The remaining terms are again computed from the probabilities for the fragments. The most important difference to commonly used all-atomic MD force fields is that hydrogen bonds are also computed from the geometric dependent probabilities, instead of using electrostatic calculations with partial charges. The optimal side-chain rotamers for the all-atomic function are computed as an independent, separate step during the dihedral angles minimization. Thus, the following description of the backbone optimization is also valid for the all-atomic function.

2.3 Fragment Insertion by Monte Carlo

The torsion angles from the fragments in the library are assigned to the sequence by a Monte Carlo procedure. The Monte Carlo procedure is a method to minimize any function, which can be evaluated for every possible state. It only requires a starting state and a set of possible moves. It chooses randomly a possible move and accepts it with the Metropolis-Hasting acceptance probability $P = \exp(\frac{\Delta E}{kT})$. Thus, every move with decreasing energy is accepted and also some with increasing energy are accepted, which is necessary to escape local minima. The temperature T is changed during the minimization (called simulated annealing). The starting state is arbitrarily selected as the fully extended configuration and the scoring function can be easily computed for any possible combination of dihedral angles.

2.4 Fragment Assignment and Local Moves

The most basic move is the fragment assignment. A fragment along the sequence is randomly selected and its dihedral angles are overwritten with those in the library. The model in the library is chosen with a probability according to the rank of the possible fragment models.

A fragment assignment is a global move. The whole protein structure is effected by the net rotation and translation of the backbone to each side caused by the fragment assignment. This net effect is in general non zero. The advantage of a global move is that it can change the overall structure faster than local moves. However, the acceptance probability is small because global rearrangement destroys the already formed local contacts and thus can increase the energy significantly.

Three different local moves are used and, on average, have much higher acceptance probability also after the protein is already partly minimized. The first is a small or shear motion of random dihedral angles with negligible global effect. The second is a fragment insertion which is explicitly selected to have only a local effect because it has a neglectable net rotation and translation (gunn method) or a negligible MSD change for the rest of the protein (called frag). The third move is a fragment insertion with a compensating change of neighboring dihedral angles (called crank and wobble). In Fig 2 from [Rohl et al., 2004b] one can see the crank move and in Fig 3, from the same reference, one can see the average acceptance rate and effectiveness of the different moves.

3 Discussion

The effectiveness of the protein structure prediction in general and the free modeling prediction in particular can be best judged by the CASP results. Groups participating in CASP submit their prediction for soon to be released proteins and assessors analyze those by numerical methods like GDT_TS [Bystroff and Baker, 1998] and visual inspection. The articles from the assessor groups are thus the best source for a comparison of the methods and progress in the field. One can clearly see a progress since the start of CASP. In CASP1 most of the new fold predictions were almost random [Moult, 2005]. A comparison of two consecutive CASP is difficult because the progress made in two years time is not so large and the small number of targets may result in varring difficulty of the targets.

The most recent CASP is CASP7. The assessment for the free modeling targets (the "new fold" category was renamed) showed that Rosetta was the

most accurate [Jauch et al., 2007], which can be partly credited to an extensive all-atom refinement made possible by the large computing power of a distributed computing network based on BIONC [Das et al., 2007]. Rosetta was also successfully used for the homology modeling target as a refinement step. Tasser [Zhou et al., 2007], a newer software also using the fragment approach, is very interesting because it was able to predict the targets with similar accuracy while needing far less computing time. The free modeling method has predicted some structures with very high accuracy, sometimes even exceeding the accuracy of template/homology based approaches. However, on average, the template based approach is still more accurate, especially if all evolutionary information is considered. Even though the fragment approach shows progress and is the best known free modeling prediction method, the best strategies for all the individual steps of the method (scoring function, fragment selection, fragment assembly and minimization) are yet unknown. As well, the method is computationally expensive and, as yet, no free modeling method predicts the correct fold for the majority of the targets.

The most important future research will be regarding better refinement and I am particularly interested in comparisons to MD simulation. Further improvement in the all-atomic refinement both for template and de novo based approaches would allow the use of those structures as starting structures for MD simulations and would enable docking and enzymatic calculations. This would make it an important tool, together with the still expensive experimental methods, with which to solve structures. Better refinement will require more accurate scoring functions and more efficient minimization methods. Additionally, the de novo approach will need improvement in the percentage of correct folds and has to be extended to longer sequences to make it a generally useful tool. The latter is limited at the moment because the possibilities to assemble a structure out of fixed size fragments increases exponentially with the sequence length [Zhang and Skolnick, 2004]

MD Umbrella sampling between the different predicted models could compute the free energy differences of these models. This would allow the comparison of MD and Rosetta energy function and analysis of in which cases they are similar and in which cases they are different. This could give a better understanding of the error of both energy functions and could possibly reveal whether kinetically not accessible lower minima than the native state exist. Better understanding of the error in the MD energy function would allow refinement of already close models using MD simulations. It is believed that this is currently not possible mainly because of the error in the MD force field [Baker and Sali, 2001]. In case the energies do not differ too much in the future, the different predicted models could be used as sampling start

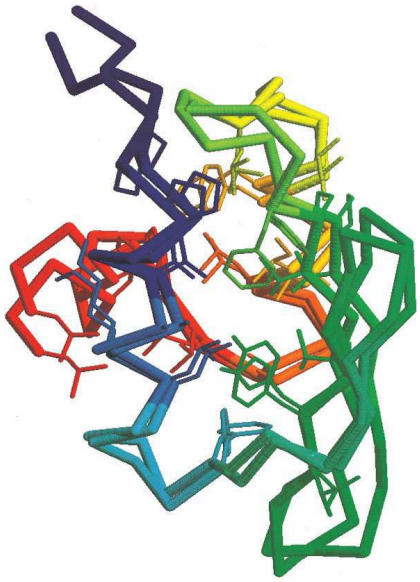


Figure 1: Close to atomic-level structure prediction from CASP6 [Bradley et al., 2005]

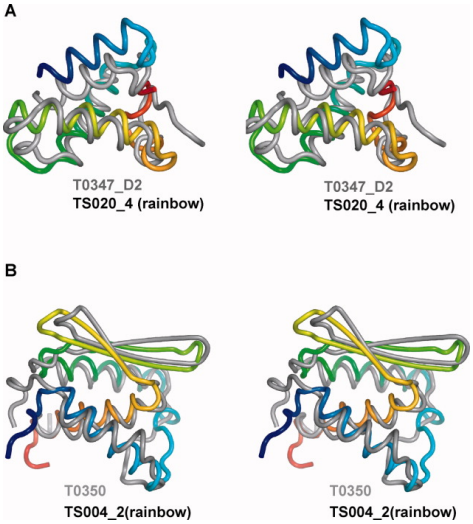


Figure 2: Well predicted structures in CASP7 [Jauch et al., 2007]. TS020 is the Baker group and TS004 is done with ROBETTA

points in theories using metastable states [Noé et al., 2007].

4 Comment

My main source of information was the very good review of Rosetta [Rohl et al., 2004b]. All but the references describing the scoring function terms in Tables I and II were directly used for the paper at the cited places. I never worked or read details about de-novo prediction before. The only useful prior knowledge was in MD and Monte Carlo. The paper is written by myself and I did not use anything but the cited references.

References

D. Baker and A. Sali. Protein Structure Prediction and Structural Genomics. *Science*, 294(5540):93–96, 2001. doi: 10.1126/science.1065659. URL `http://www.sciencemag.org/cgi/content/abstract/294/5540/93`.

R. Bonneau, C. Strauss, C. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, and D. Baker. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol*, 322(1):65–78, 2002.

P. Bowers, C. Strauss, and D. Baker. De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR*, 18(4):311–318, 2000.

P. Bradley, L. Malmstrom, B. Qian, J. Schonbrun, D. Chivian, D. Kim, J. Meiler, K. Misura, and D. Baker. Free modeling with Rosetta in CASP6. *Proteins*, 61(Suppl 7):128–134, 2005.

C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol*, 281(3):565–77, 1998.

R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, D. E. Kim, W. H. Sheffler, L. Malmstrm, A. M. Wollacott, C. Wang, I. Andre, and D. Baker. Structure prediction for casp7 targets using extensive all-atom refinement with rosetta@home. *Proteins: Structure, Function, and Bioinformatics*, 9999(9999):NA, 2007. doi: 10.1002/prot.21636. URL `http://dx.doi.org/10.1002/prot.21636`.

R. Jauch, H. C. Yeo, P. R. Kolatkar, and N. D. Clarke. Assessment of casp7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 9999(9999):NA, 2007. URL `http://dx.doi.org/10.1002/prot.21771`.

R. D. Jr and F. Cohen. Bayesian statistical analysis of the backbone-dependent rotamer preferences of protein sidechains. *Protein Science*, 6: 1661–1681, 1997.

T. Kortemme, A. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol*, 326:1239–1259, 2003.

B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19): 10383–10388, 2000.

T. Lazaridis and M. Karplus. RESEARCH ARTICLES Effective Energy Function for Proteins in Solution. *PROTEINS: Structure, Function, and Genetics*, 35:133–152, 1999.

J. Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–289, 2005. URL http://www.sciencedirect.com/science/article/B6VS6-4GBD6KT-1/2/08167b756ffa066f3d63f71cbbf97908.

F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *The Journal of Chemical Physics*, 126(15):155102, 2007. doi: 10.1063/1.2714539. URL http://link.aip.org/link/?JCP/126/155102/1.

V. Pande. Rosetta and folding, 2006. URL http://forum.folding-community.org/viewtopic.php?p=125338#125338.

C. Rohl, C. Strauss, D. Chivian, and D. Baker. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins Structure Function and Bioinformatics*, 55(3):656–677, 2004a.

C. Rohl, C. Strauss, K. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods Enzymol*, 383:66–93, 2004b.

K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 268(1):209–225, 1997.

K. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins Structure Function and Genetics*, 34(1):82–95, 1999.

W. Wedemeyer and D. Baker. Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. *Proteins Structure Function and Genetics*, 53(2):262–272, 2003.

Y. Zhang and J. Skolnick. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys J*, 87(4):2647–2655, Oct 2004. doi: 10.1529/biophysj.104.045385. URL http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=15454459&dopt=Citation.

H. Zhou, S. B. Pandit, S. Y. Lee, J. Borreguero, H. Chen, L. Wroblewska, and J. Skolnick. Analysis of tasser-based casp7 protein structure prediction results. *Proteins: Structure, Function, and Bioinformatics*, 9999(9999):NA, 2007. URL http://dx.doi.org/10.1002/prot.21649.

TABLE I
COMPONENTS OF ROSETTA ENERGY FUNCTION[a]

Name	Description (putative physical origin)	Functional form	Parameters (values)
env[b]	Residue environment (solvation)	$\sum_i -\ln[P(\mathrm{aa}_i\|\mathrm{nb}_i)]$	i = residue index aa = amino acid type nb = number of neighboring residues[c] (0, 1, 2... 30, >30)
pair[b]	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i, \mathrm{aa}_j\|s_{ij}d_{ij})}{P(\mathrm{aa}_i\|s_{ij}d_{ij})P(\mathrm{aa}_i\|s_{ij}d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = centroid–centroid distance (10–12, 7.5–10, 5–7.5, <5 Å) s = sequence separation (>8 residues)
SS[d]	Strand pairing (hydrogen bonding)	Scheme A : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$ Scheme B : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_{d\sigma}$ where $\mathrm{SS}_{\phi,\theta} = \sum_m \sum_{n>m} -\ln[P(\phi_{mn}, \theta_{mn}\|d_{mn}, \mathrm{sp}_{mn}, s_{mn})]$ $\mathrm{SS}_{hb} = \sum_m \sum_{n>m} -\ln[P(\mathrm{hb}_{mn}\|d_{mn}, s_{mn})]$ $\mathrm{SS}_d = \sum_m \sum_{n>m} -\ln[P(d_{mn}\|s_{mn})]$ $\mathrm{SS}_{d\sigma} = \sum_m \sum_{n>m} -\ln[\mathrm{P}(d_{mn}\sigma_{mn}\|\rho_m, \rho_n)]$	m, n = strand dimer indices; dimer is two consecutive strand residues V = vector between first N atom and last C atom of dimer m = unit vector between $\hat{V}_m$ and $\hat{V}_n$ midpoints x = unit vector along carbon–oxygen bond of first dimer residue y = unit vector along oxygen–carbon bond of second dimer residue ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) hb = dimer twist, $\sum_{k=m,n} 0.5(\lvert\hat{m}\cdot\hat{x}_k\rvert + \lvert\hat{m}\cdot\hat{y}_k\rvert)$ (< 0.33, 0.33–0.66, 0.66–1.0, 1.0–1.33, 1.33–1.6, 1.6–1.8, 1.8–2.0) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 6.5 Å) σ = angle between $\hat{V}_m$ and $\hat{M}$ (18° bins) sp = sequence separation between dimer-containing strands (< 2, 2–10, > 10 residues) s = sequence separation between dimers (>5 or >10) ρ = mean angle between vectors $\hat{m}$, $\hat{x}$ and $\hat{m}$, $\hat{y}$ (180° bins)
sheet[e]	Strand arrangement into sheets	$-\ln[P(n_{\mathrm{sheets}} n_{\mathrm{lonestrands}}\|n_{\mathrm{strands}})]$	n_{sheets} = number of sheets $n_{\mathrm{lone\ strands}}$ = number of unpaired strands n_{strands} = total number of strands
HS	Helix–strand packing	$\sum_m \sum_n -\ln[P(\phi_{mn}, \psi_{mn}\|sp_{mn}d_{mn})]$	m = strand dimer index; dimer is two consecutive strand residues n = helix dimer index; dimer is central two residues of four consecutive helical residues $\hat{V}$ = vector between first N atom and last C atom of dimer ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) sp = sequence separation between dimer-containing helix and strand (binned < 2, 2–10, >10 residues) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 12 Å)
rg	Radius of gyration (vdw attraction; solvation)	$\sqrt{\langle d_{ij}^2 \rangle}$	i, j = residue indices d = distance between residue centroids
cbeta	Cβ density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\mathrm{compact}}(\mathrm{nb}_{i,sh})}{P_{\mathrm{random}}(\mathrm{nb}_{i,sh})}\right]$	i = residue index sh = shell radius (6, 12 Å) nb = number of neighboring residues within shell[f] P_{compact} = probability in compact structures assembled from fragments P_{random} = probability in structures assembled randomly from fragments
vdw[g]	Steric repulsion	$\sum_i \sum_{j>i} \frac{\left(r_{ij}^2 - d_{ij}^2\right)^2}{r_{ij}};\ d_{ij} < r_{ij}$	i, j = residue (or centroid) indices d = interatomic distance r = summed van der Waals radii[h]

[a] All terms originally described in Refs. 16 and 17.
[b] Binned function values are linearly interpolated, yielding analytic derivatives.

(continued)

TABLE I *(continued)*

[c] Neighbors within a 10-Å radius. Residue position defined by Cβ coordinates (Cα for glycine).
[d] Interactions between dimers within the same strand are neglected. Favorable interactions are limited to preserve pairwise strand interactions, that is, dimer m can interact favorably with dimers from at most one strand on each side, with the most favorable dimer interaction ($SS_{\phi,\theta} + SS_{hb} + SS_d$) determining the identity of the interacting strand. $SS_{d\sigma}$ is exempt from the requirement of pairwise strand interactions. SS_{hb} is evaluated only for m, n pairs for which $SS_{\phi,\theta}$ is favorable. $SS_{d\sigma}$ is evaluated only for m, n pairs for which $SS_{\phi,\theta}$ and $S\tilde{S}_{hb}$ are favorable. A bonus is awarded for each favorable dimer interaction for which $|m - n| > 11$ and strand separation is more than eight residues.
[e] A sheet is composed of all strands with dimer pairs <5.5 Å apart, allowing each strand having at most one neighboring strand on each side. Discrimination between alternate strand pairings is determined according the most favorable dimer interaction. Probability distributions fitted to $c(n_{\text{strands}}) - 0.9n_{\text{sheets}} - 2.7n_{\text{lone strands}}$ where $c(n_{\text{strands}}) = (0.07, 0.41, 0.43, 0.60, 0.61, 0.85, 0.86, 1.12)$.
[f] Residue position defined by Cβ coordinates (Cα for glycine).
[g] Not evaluated for atom (centroid) pairs whose interatomic distance depends on the torsion angles of a single residue.
[h] Radii determined from (1) 25th closest distance seen for atom pair in pdbselect25 structures, (2) the fifth closest distance observed in X-ray structures with better than 1.3-Å resolution and <40% sequence identity, or (3) X-ray structures of <2-Å resolution, excluding i, $i + 1$ contacts (centroid radii only).

TABLE II
COMPONENTS OF ROSETTA ALL-ATOM ENERGY FUNCTION[a]

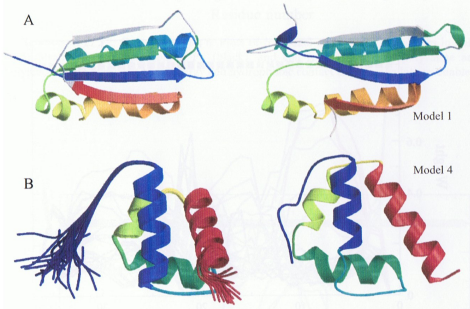
Name	Description (physical origin)	Functional form	Parameters	Ref.
rama	Ramachandran torsion preferences	$\sum_i -\ln[P(\phi_i, \psi_i \mid \text{aa}_i, \text{ss}_i)]$	i = residue index ϕ,ψ = backbone torsion angles (36° bins) aa = amino acid type ss = secondary structure type[b]	7, 12
LJ[c]	Lennard–Jones interactions	$\sum_i \sum_{j>i} \begin{cases} \left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right] e_{ij}, & \text{if } \frac{d_{ij}}{r_{ij}} > 0.6 \\ \left[-8759.2\left(\frac{d_{ij}}{r_{ij}}\right) + 5672.0\right] e_{ij}, & \text{else} \end{cases}$	i, j = residue indices d = interatomic distance e = geometric mean of atom well depths[d] r = summed van der Waals radii[e]	15
hb[f]	Hydrogen bonding	$\sum_i \sum_j (-\ln[P(d_{ij} \mid h_j \text{ss}_{ij})] - \ln[P(\cos\theta_{ij} \mid d_{ij} h_j \text{ss}_{ij})] - \ln[P(\cos\psi_{ij} \mid d_{ij} h_j \text{ss}_{ij})])$	i = donor residue index j = acceptor residue index d = acceptor–proton interatomic distance h = hybridization (sp^2, sp^3) ss = secondary structure type[g] θ = proton–acceptor–acceptor base bond angle ψ = donor–proton–acceptor bond angle	19–21
solv	Solvation	$\sum_i \left[\Delta G_i^{\text{ref}} - \sum_j \left(\frac{2\Delta G_i^{\text{free}}}{4\pi^{3/2}\lambda_i r_{ij}^2} e^{-d_{ij}^2} V_j + \frac{2\Delta G_i^{\text{free}}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i\right)\right]$	i,j = atom indices d = distance between atoms r = summed van der Waal radii[e] λ = correlation length[h] V = atomic volume[h] $\Delta G^{\text{ref}}, \Delta G^{\text{free}}$ = energy of a fully solvated atom[h]	15, 18

(continued)

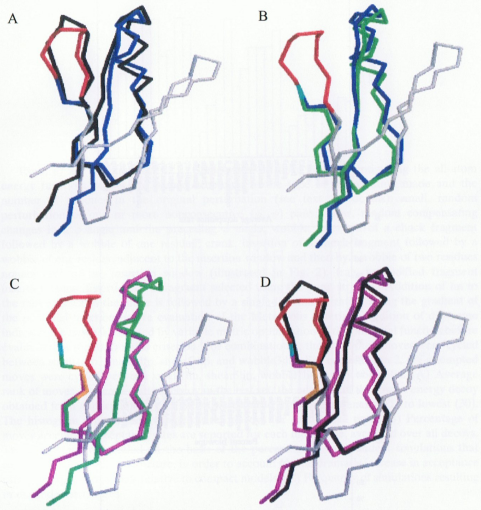
TABLE II *(continued)*

pair	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i,\ \mathrm{aa}_j\|d_{ij})}{P(\mathrm{aa}_i\|d_{ij})P(\mathrm{aa}_i\|d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = distance between residues[i]	15
dun	Rotamer self-energy	$\sum_i -\ln\left[\frac{P(\mathrm{rot}_i\|\phi_i,\psi_i)P(\mathrm{aa}_i\|\phi_i,\psi_i)}{P(\mathrm{aa}_i)}\right]$	i, j = residue indices rot = Dunbrack backbone-dependent rotamer aa = amino acid type ϕ, ψ = backbone torsion angles	14, 15
ref	Unfolded state reference energy	$\sum_{\mathrm{aa}} n_{\mathrm{aa}}$	aa = amino acid type n = number of residues	15

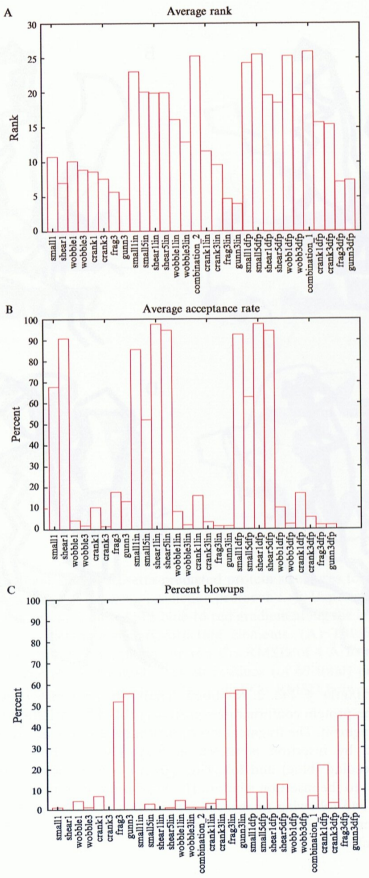
[a] All binned function values are linearly interpolated, yielding analytic derivatives, except as noted.
[b] Three-state secondary structure type as assigned by DSSP.[22]
[c] Not evaluated for atom pairs whose interatomic distance depends on the torsion angles of a single residue.
[d] Well depths taken from CHARMm19 parameter set.[23]
[e] Radii determined from fitting atom distances in protein X-ray structures to the 6–12 Lennard–Jones potential using CHARMm19 well depths.
[f] Evaluated only for donor acceptor pairs for which $1.4 \le d \le 3.0$ and $90^\circ \le \psi, \theta \le 180^\circ$. Side-chain hydrogen bonds in involving atoms forming main-chain hydrogen bonds are not evaluated. Individual probability distributions are fitted to eighth-order polynomials and analytically differentiated.
[g] Secondary structure types for hydrogen bonds are assigned as helical ($j - i = 4$, main chain); strand ($|j - i| > 4$, main chain), or other.
[h] Values taken from Lazaridis and Karplus.[18]
[i] Residue position defined by Cβ coordinates (Cα of glycine).



ROHL *ET AL.*, CHAPTER 4, FIG. 1. Rosetta-predicted protein structures for CASP 5 targets. *Right:* Models predicted using the *de novo* prediction protocol. *Left:* Experimental structure of each protein. Protein chains are colored in a blue-to-red gradient along the length of the chain to highlight correctly predicted secondary structure elements. (A) T0135. The predicted model has 54 residues (of 106 total) predicted at a Cα RMSD of 4 Å to the experimental structure. (B) T0171. The predicted model has 60 residues (of 69 total) predicted at a Cα RMSD of 4 Å to the experimental structure. The global Cα RMSD between the prediction and the experimental structure is 4.2 Å.



ROHL *ET AL.*, CHAPTER 4, FIG. 2. Modified "crank" fragment insertion into 1 dan. (A) Superposition of the protein conformations preceding (black) and following (blue) insertion of a nine-residue fragment. The fragment insertion window is shown in red. The portion of the chain unperturbed by insertion is shown in gray. (B) Superposition of the protein conformations preceding (blue) and following (green) optimization of angles at a wobble site (cyan) adjacent to the insertion window. (C) Superposition of the protein conformations preceding (green) and following (magenta) optimization of angles at a second wobble site (orange) nonadjacent to the insertion window. (D) Superposition of the original (black) and final (magenta) conformations.



ROHL *ET AL.*, CHAPTER 4, FIG. 3. Comparison of move types in optimizing the all-atom energy function. Moves are named according to the type of perturbation made and the number of residues in the original perturbation (see text for details): small, random perturbation of one or more nonconsecutive (ϕ, ψ) pairs; shear, random compensating changes in a ϕ angle and the preceding ψ angle; wobble, insertion of a chuck fragment followed by a wobble of one residue; crank, insertion of a chuck fragment followed by a wobble of one residue adjacent to the insertion window and then by a wobble of two residues nonadjacent to the insertion window (illustrated in Fig. 2); frag, unmodified fragment insertion; gunn, insertion of a fragment selected using the gunn strategy. Addition of lin to the move indicates the move is followed by a single-line minimization along the gradient of the potential function before evaluation of the Metropolis criterion. Addition of dfp name indicates the move is followed by variable metric optimization of the potential function before evaluation of the Metropolis criterion. For combination 1, the attempted moves were cycled between small1dfp, small5dfp, shear5dfp, and wobble3dfp. For combination 2, the attempted moves were cycled between small1lin, shear5lin, wobble1lin, and wobble3lin. (A) Average rank of moves. For each starting decoy in the test set, the energies of the lowest energy decoy obtained from application of each move were sorted from highest energy (1) to lowest (30). The histogram reports the average overall decoys for each move type. (B) Percentage of moves accepted. Acceptance rates are reported for each move type, averaged over all decoys. The percentage was scaled on the basis of the percentage of independent simulations that resulted in an expanded structure, in order to account for the dramatic increase in acceptance rate into expanded models relative to compact models. (C) Frequency of simulations resulting in expanded structures.