Supporting Information

Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-based Force Field

Dong Xu and Yang Zhang

Table S1. Architectures of the four back-propagation Neural Networks for feature prediction.

	Window size	# nodes in input layer	# nodes in hidden layer 1	# nodes in hidden layer 2	# nodes in output layer
Solvent	21	21*23=483	60	60	1
Phi	21	21*23=483	50	50	1
Psi	17	17*23=391	80	80	1
Beta-turn	25	25*23=575	150	150	1

Figure S1. Illustration of backbone hydrogen bonds. (A) Single H-bond between acceptor residue *i* and donor residue *j*. (B) One virtual H-bond between residue *i* and *i+3* and one real H-bond between residue *i* and *i+4* in a helix. (C) Two H-bonds between residues (*i-1*, *j*) and (*j*, *i+1*) in a parallel β-sheet determine the β-pair (*i*, *j*). (D) Two H-bonds between residues (*i*, *j+1*) and (*j-1*, *i*) in a parallel β-sheet determine the β-pair (*i*, *j*). (E) Two H-bonds between residues (*i*, *j*) and (*j*, *i*) in an antiparallel β-sheet determine the β-pair (*i*, *j*). (F) Two H-bonds between residues (*i-1*, *j+1*) and (*j-1*, *i+1*) in an antiparallel β-sheet determine the β-pair (*i*, *j*).

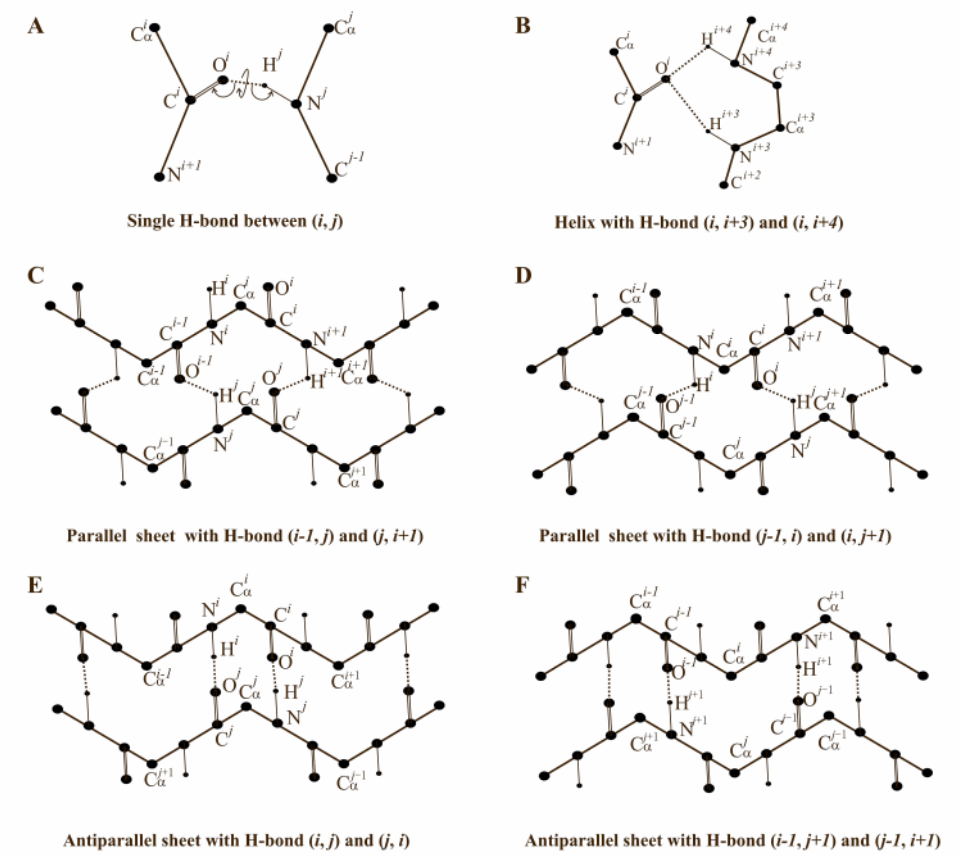

Figure S2. An illustrative example of solvent accessibility estimation and prediction for protein chain 1bgfA. The blue curve is SA calculated by EDTSurf from full-atomic structure, the red curve is that estimated from the reduced model by the pair-wise residue contacts, and the green curve is that predicted by Neural Network from the target sequence.

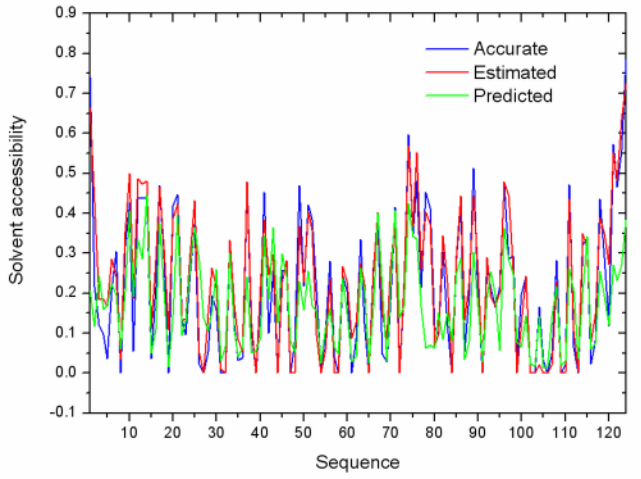

Figure S3. Four Ramachandran plots in terms of different residue types and different secondary structure types. The value of each point is the negative logarithm of the number of torsion-angle pairs in that point. (A) Aspartic acid, α-helix. (B) Aspartic acid, β-sheet. (C) Aspartic acid, coil. (D) Arginine, α-helix.

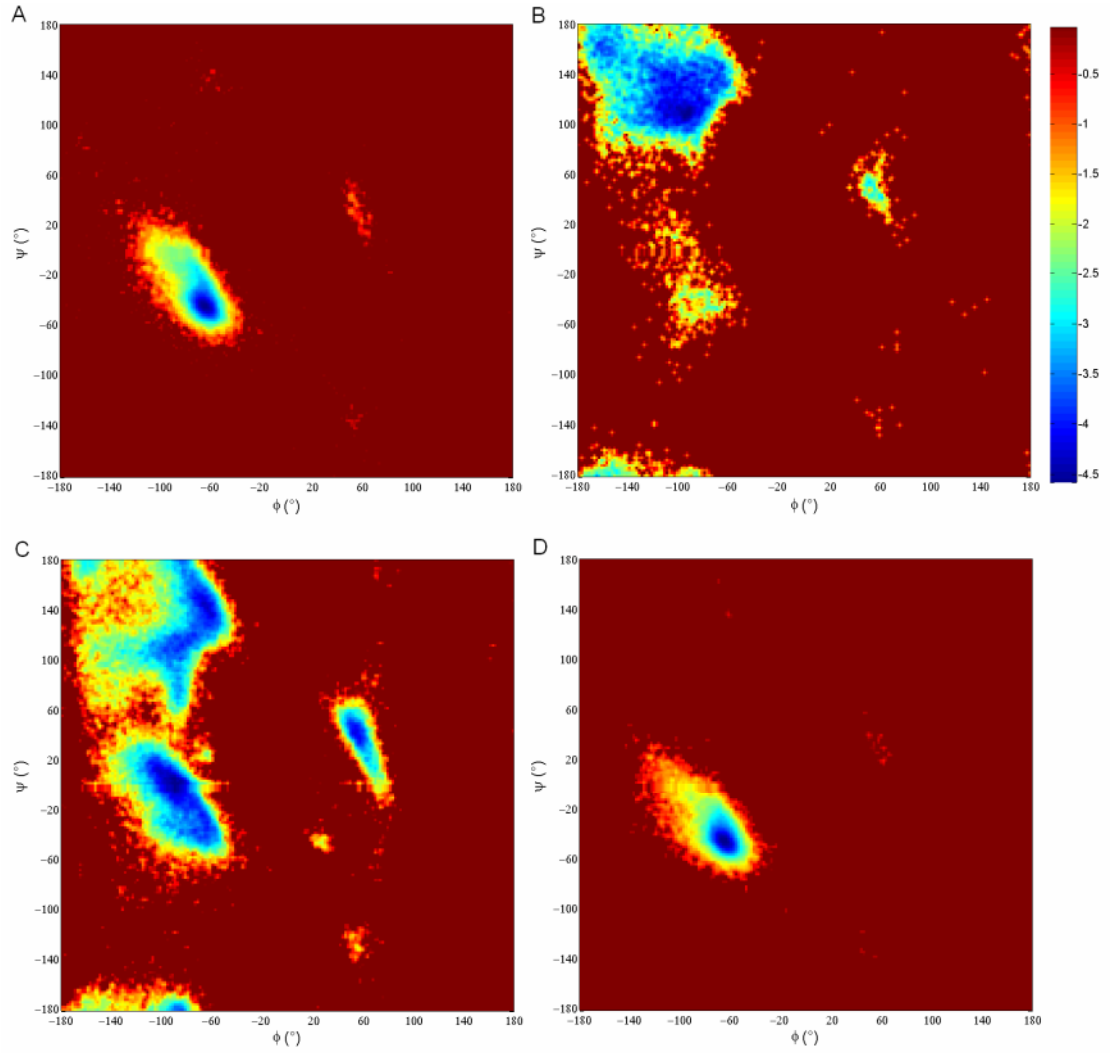

Figure S4. Average and minimum radii of gyration in terms of protein lengths.

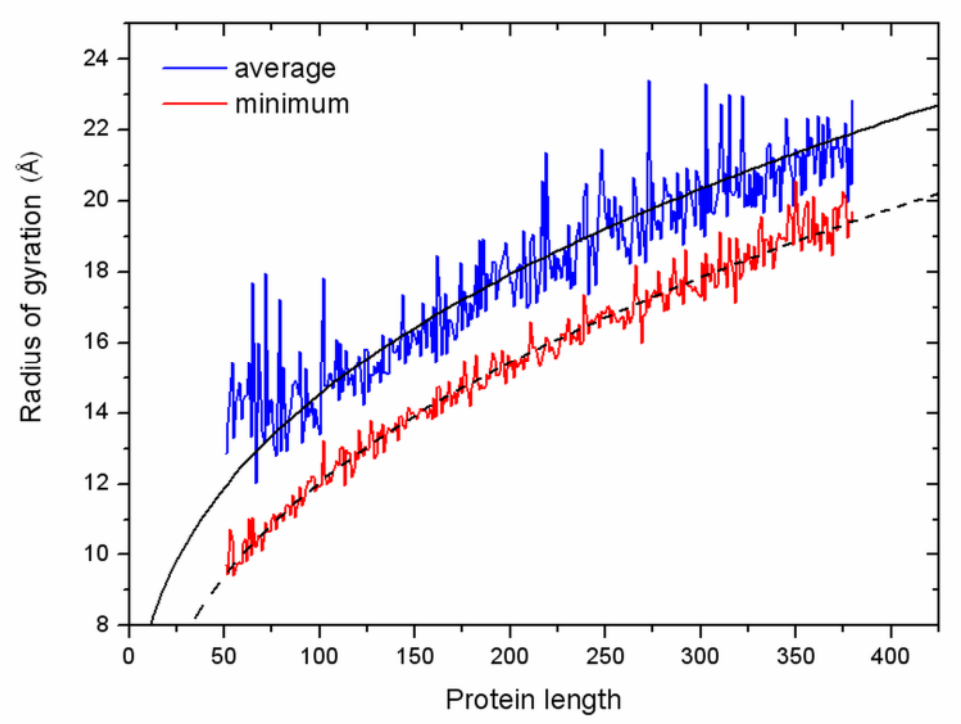

Figure S5. Energy between every pair of α-helices in terms of their torsion angle and distance.

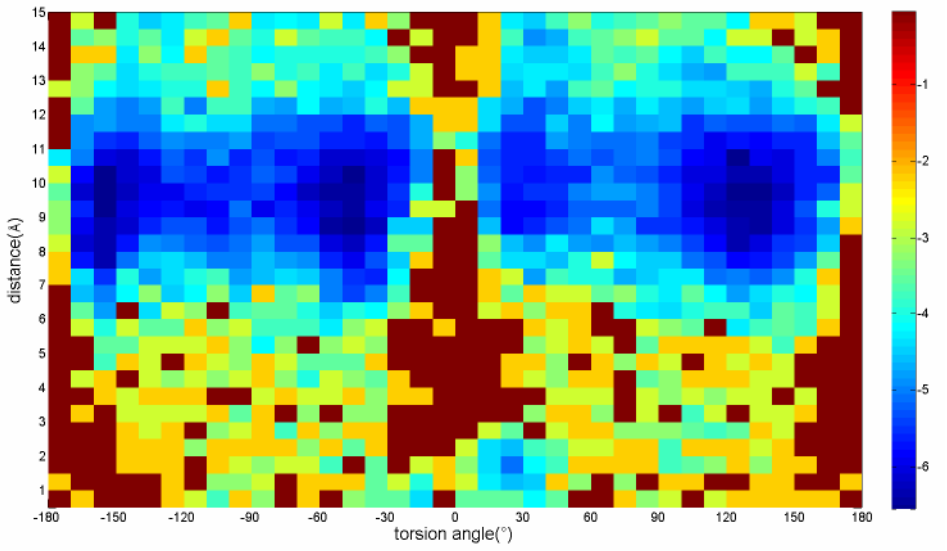

Figure S6. Energy between every pair of β-pairing residues. (A) In parallel β-sheets. (B) In antiparallel β-sheets.

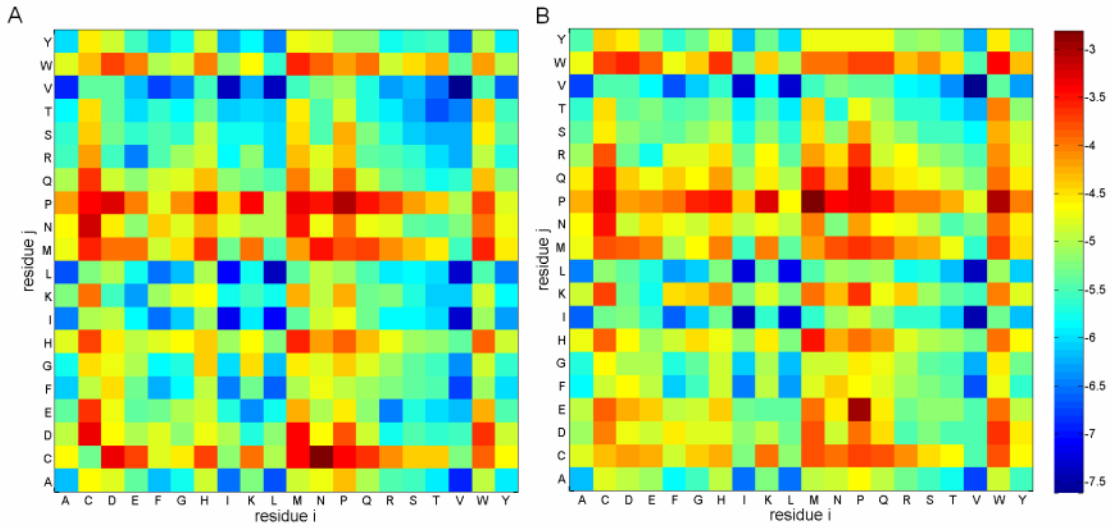

Figure S7. Probabilities of fragment substitution movement in terms of Monte Carlo cycles.

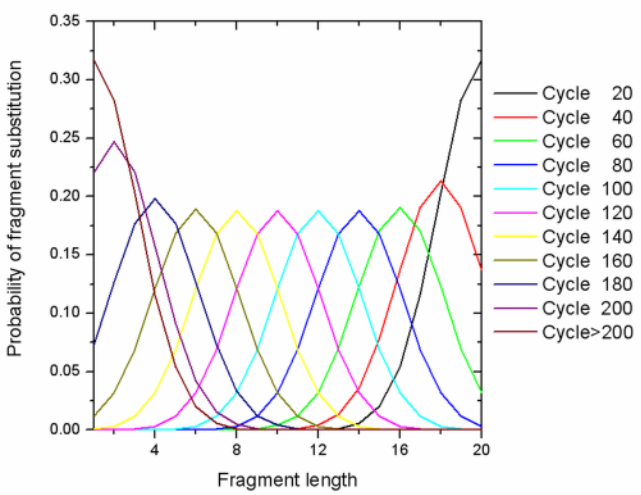

Figure S8. QUARK local movement analysis. (A) Acceptance rate of the highest temperature replica for different movements. (B) Acceptance rate of the lowest temperature replica for different movements. (C) Acceptance rate of replicas in the first cycle for different movements. (D) Acceptance rate of replicas in the first 200 cycles for different movements.

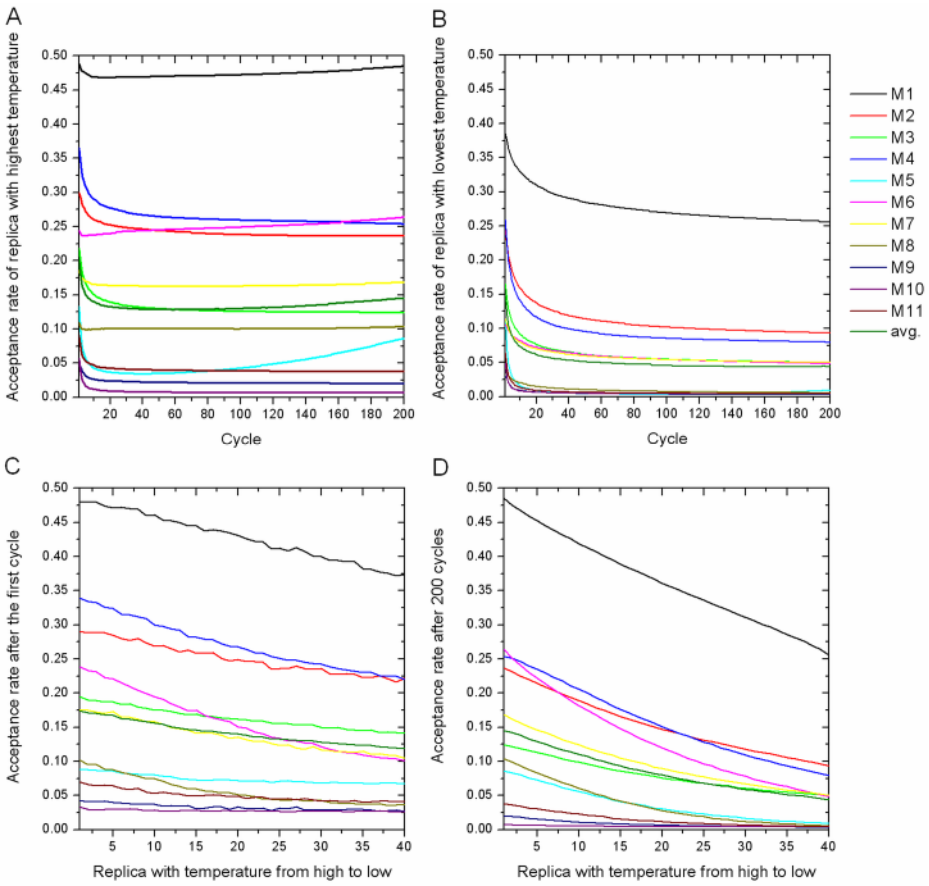

Table S2. Proportions of 11 movements during the Monte Carlo simulation.

Movements		Descriptions	Percentages
Residue-level	M1	Bond-length perturbation	5%
	M2	Bond-angle perturbation	5%
	M3	Torsion-angle perturbation	21%
	M4	Torsion-angle pair selected from look-up table	10%
Segment-level	M5	Fragment substitution	34%
	M6	LMProt	5%
	M7	Rotation	5%
	M8	Shifting	5%
Topology-level	M9	Helix packing	2%
	M10	Beta pairing	5%
	M11	Beta-turn formation	3%

Figure S9. QUARK global movement analysis. (A) Average acceptance rate of replica swaps. (B) Energy trajectories of the first and last five replicas.

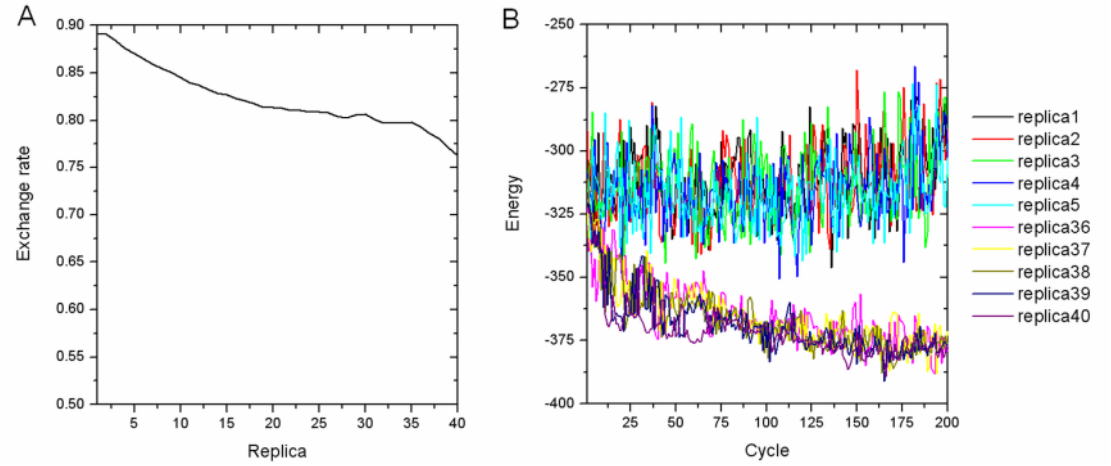

Figure S10. Temperature distribution of 40 replicas for proteins of 70 and 150 amino acids.