Supplementary Online Materials for:

Design of a Novel Globular Protein Fold with Atomic Level Accuracy

Brian Kuhlman*, Gautam Dantas*, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker

* These authors contributed equally to this work

Energies and sequence for Top7 before and after alternating cycles of backbone and sequence optimization.

```
before DIEITVRINNNGEDYDYKKTATTLSEINAHFEELEKHLKEENGEKITISVKLRNEKEAYW
after  DIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKDYIKKQGAKRVRISITARTKKEAEK

before VAAKIKEQALRAGVETIQIDKQSDTMTATLGKQ
after  FAAILIKVFAELGYNDINVTFDGDTVTVEGQLE
```

Table S1. Energies for Top7 before and after iterative cycles of backbone and sequence optimization (kcal / mole). Expected Lennard-Jones energies are derived from the average Lennard-Jones energy for each of the twenty amino acids for different degrees of burial.

	Top7 before relaxation	Final Top7 model
Lennard-Jones (LJ) attractive	-370	-385

Lennard-Jones (LJ) repulsive	28	8.6
Hydrogen bonding	-89	-80
Solvation energy	188	175
Total energy	-324	-386
LJ attractive – expected LJ attractive (avg. per buried residue)	0.3	-0.3
LJ repulsive – expected LJ repulsive (avg. per buried residue)	0.2	-0.2

EXPERIMENTAL METHODS

Protein Expression and Purification

Synthetic genes which place the computationally selected protein sequences under the control of the T7 promoter, with a C terminal 6X His tag, and a codon usage optimal for *Escherichia coli* (*E. coli*) were obtained from BlueHeron Biotechnologies. The gene constructs were cloned in plasmid pet29b(+) (Novagen) and expressed in the BL21(DE3)pLysS strain of *E. Coli*. Cells were grown in LB media at 37°C to an OD_{600} of 0.6, induced with 1mM isopropyl-thio-β-D-galactosidase (IPTG), and cells were harvested after another 5 hours of growth at 37°C. Harvested cells were lysed by three freeze-thaw cycles, and soluble protein collected after centrifugation of cellular debris. Soluble protein was purified on a Ni^{+} affinity column (Pharmacia Biotech) followed by10^4-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein was further purified on a QFF anion exchange column (Pharmacia) with a 30mM to 500mM NaCl gradient in 25mM TRIS-HCl, pH 8.0, followed by a final 10^4-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein identity and purity was determined by SDS-PAGE and ESI-MALDI Mass Spectroscopy. Protein concentrations were determined by

UV absorbance at 280nm with extinction coefficients calculated using the ExPASy Protparam tool (http://us.expasy.org/tools/protparam.html).

The following modifications were made to the above procedure for Top7 crystallography. A Lys^{37} to Met^{37} point mutant of Top7 (Top7_K37M) was generated using the Single Quikchange Mutagenesis kit (Stratagene). Selenomethionine containing Top7_K37M was expressed in minimal media from the *E. coli* strain BL21(DE3) adapted for growth with methionine pathway inhibition (*1*). Cells were grown in minimal media at 37°C to an OD_{600} of 0.8 and the following amino acids were added to inhibit the methionine biosynthetic pathway: 100 mg/L lysine, threonine, phenylalanine; 75 mg/L selenomethionine; 50mg/L leucine, isoleucine, valine. Following a 15-minute incubation at 37°C, IPTG was added to induce expression and the cultures were harvested after 5 hours of growth at 37°C. Purification was performed as described.

^{15}N-labelled Top7 was prepared by expression in M9 minimal media with ^{15}N-labelled NH_4Cl. Purification was performed as described for unlabelled protein.

Circular Dichroism (CD)

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260-200nm) at varying protein concentrations (15-25μM), guanidinium hydrochloride (Gu-HCl) concentrations (0-8.3M), and temperatures (0-98°C) were collected in a 1mm pathlength cuvette. Gu-HCl induced protein denaturation was followed by the change in ellipticity at 220nm in a 1cm pathlength cuvette, using a Microlab titrator (Hamilton) for

denaturant mixing. Temperature was maintained at 25°C with a Peltier device. All CD data were converted to mean residue ellipticity. Temperature induced protein denaturation was followed by the change in ellipticity at 220nm in a 2mm pathlength cuvette. To obtain a value for DG_U^{H2O}, chemical denaturation curves were fit by nonlinear least squares analysis using the linear extrapolation model as applied by Santoro and Bolen. To obtain a value for $DC°_p$, thermal denaturation curves were fit using the Gibbs-Hemholtz equation in the form:

$$f = f_f + \frac{(f_u - f_f)}{1 + e^{\frac{-\Delta G°}{R \cdot T}}}$$

$$-\Delta G° = \Delta H°(1 - \frac{T}{T_m}) + \Delta C°_p\{T - T_m - T \cdot \ln(\frac{T}{T_m})\}$$

where f is CD signal, f_f and f_u are the estimated CD signal for the folded and unfolded states, respectively, R is the gas constant, T is temperature, T_m is the temperature where 50% of the protein is folded, $DG°$ is the change in the Gibbs free energy for the unfolding reaction, $DH°$ is the change in enthalpy, and $DC°_p$ is the change in heat capacity.

Nuclear Magnetic Resonance

The 2D NOESY spectrum of ~1mM Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz and 200ms mixing time using Watergate suppression. The 2D HSQC spectrum of ~1mM 15N-labelled Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz using the fast HSQC scheme of Mori et al. (*2*)

Crystallization

Selenomethionyl substituted Top7_K37M was crystallized in hanging drops (1 μl of protein solution at 25 mg/ml with 1 μl of well solution). The well solutions ranged from 15 - 20% PEG 3350 and 250 mM ammonium formate pH 6.6. The protein crystals grew within a day and were between 50-200 μm on a side. They were initially transferred to a cryo-solution of well solution at 25% PEG 3350 plus 25 % (v/v) glycerol in 4 steps of increasing glycerol and flash frozen in liquid nitrogen. With this treatment the crystals diffracted in a trigonal space group ($P3_221$) with unit cell dimensions a = 35.9 Å, b = 35.9 Å, c = 140.6 Å. A single wavelength (0.9793 Å) anomalous dispersion (SAD) (*3*) data set was collected to 2.5 Å resolution on beam-line 8.2.1 at the ALS (Advanced Light Source, Lawrence Berkeley Laboratory, Berkeley) using a four panel ADSC CCD area detector. Multiwavelength data collection (MAD phasing) was not possible due to significant radiation decay. Data were processed and scaled using HKL2000 (*4*).

Structure Determination

The structure of Top7_K35M was solved by molecular replacement with the program EPMR (*5*), and by direct rebuilding into an unbiased SAD electron density map and residual difference Fourier maps. For molecular replacement, 19 surface large surface residues such as Lys, Arg, and Glx were truncated to Ala in the search model. The correlation coefficient for the initial MR search, using data to 4.0 Å resolution, was 0.52, vs. background of 0.36. For SAD phasing, the position of SeMet 37 was determined from an anomalous difference Patterson map. The initial phasing power and figure of merit for SAD phasing was 1.99 and 0.24 prior to density modification. An interpretable

electron density map was obtained after density modification with solvent flipping with a solvent content of 43 % (CNS). An initial model was built using XtalView (*6*) and O (*7*). The model was refined with CNS using the mlhl target (maximum likelihood, Hendrickson-Lattman coefficients) with 5% of the data excluded for the calculation of the cross-validating free R (*8*). 88% of all the built residues are in the most favorable regions of Ramachandran space and 12% are in the allowed regions (*9*). Statistics from phasing and refinement are shown in Table S2. The structure has been deposited in the PDB with the accession code 1QYS. Examples of the experimental electron density map were generated with XtalView and Raster 3D (*10*). Ribbon diagrams were generated with SwissPDB Viewer (*11*).

Table S2. Crystal Structure Statistics

DATA COLLECTION	
Resolution	50-2.5Å
Space Group	$P3_221$ [primitive trigonal]
Unit Cell Dimensions	35.9 Å, 35.9 Å, 140.6 Å
Wavelength	0.9793
Asymmetric Unit	Monomer
V_m	2.1 $Å^3$/dalton
Total Reflections	144,933
Unique Reflections	6,989
Completeness / (2.59-2.5)	99.1 % / (100.0%)
R_{merge} / (2.59-2.5)	4.5 / (34.4)
I / σ / (2.59-2.5)	37.8 (5.0)
PHASING	
Phasing Power	1.99

Figure of Merit (before/after DM)	0.24 (0.85)
REFINEMENT	
R_{work}	0.268
R_{free}	0.293
Number of atoms	693
Number of waters	7
Residues in most-favored regions	75 (88.2%)
Residues in additional allowed regions	7 (8.2 %)
Residues in generously allowed regions	3 (3.5%)
Residues in disallowed regions	0 (0.0%)
r.m.s.d bond lengh	0.0076
r.m.s.d. bond angles	1.35
Mean B value, mainchain	61.30 $Å^2$
Mean B value, sidechain	66.67 $Å^2$

Energy Function

The energy of a protein was computed as a linear sum of the following 11 energy terms.

$$E_{protein} = W_{rot}E_{rot} + W_{aa|phi,psi}E_{aa|phi,psi} + W_{rama}E_{rama} + W_{atr}E_{atr} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{bb\_hbond}E_{bb\_hbond} + W_{sc\_hbond}E_{sc\_hbond} + W_{sc\_bb\_hbond}E_{sc\_bb\_hbond} + W_{pair}E_{pair} - E_{ref}$$

The weights (W) for each term are given in a table at the end of this section. To calculate the solvation energy (E_{solv}) and the Lennard-Jones energies (E_{atr} and E_{rep}) the various atoms of the 20 amino acids were binned into types (Table S3).

Table S3: Definitions for atom types used in the energy functions

Atom Type Number	Atom type description
1	carbonyl carbon in sidechain of Asn and Gln, and guanidyl carbon in Arg
2	carboxyl carbon in Asp and Glu
3	aliphatic carbon with one hydrogen
4	aliphatic carbon with two hydrogens
5	aliphatic carbon with three hydrogens
6	aromatic ring carbon
7	nitrogen in Trp sidechain
8	nitrogen in His sidechain
9	nitrogen in Asn and Gln sidechain
10	nitrogen in Lys sidechain
11	nitrogen in Arg sidechain
12	nitrogen in Pro backbone
13	hydroxyl oxygen
14	carbonyl oxygen in Asn and Gln sidechains
15	carboxyl oxygen in Asp and Glu
16	sulfur in Cys and Met
17	backbone nitrogen
18	backbone alpha carbon
19	backbone carbonyl carbon
20	backbone oxygen
21	polar hydrogen
22	nonpolar hydrogen
23	aromatic hydrogen
24	backbone HN

Lennard-Jones Potential (E_{atr} and E_{rep})

A standard 12-6 Lennard-Jones potential is used except there is cutoff distance below which the potential is extrapolated linearly. Favorable energies are placed in E_{atr} and unfavorable energies are placed in E_{rep}.

$$E_{atr} = \sum_{i}^{natom} \sum_{j>i}^{natom} \left[\left(\frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}}{d_{ij}} \right)^{6} \right] e_{ij} \quad if \quad \frac{r_{ij}}{d_{ij}} < 1.12$$

$$E_{rep} = \sum_{i}^{natom} \sum_{j>i}^{natom} \left[\left(\frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}}{d_{ij}} \right)^{6} \right] e_{ij} \quad if \quad 1.33 > \frac{r_{ij}}{d_{ij}} > 1.12 \quad + \quad \sum_{i}^{natom} \sum_{j>i}^{natom} y_{\text{int } ercept} - d_{ij} * slope \quad if \quad \frac{r_{ij}}{d_{ij}} > 1.33$$

$$slope = -12 e_{ij} \left(1.33^{13} - 1.33^{7} \right) * \left(1 / r_{ij} \right)$$

$$y_{\text{int } ercept} = -slope * \left(\frac{r_{ij}}{1.33} \right) + e_{ij} \left(1.33^{12} - 2(1.33)^{6} \right)$$

$$r_{ij} = r_i + r_j$$

$$e_{ij} = \sqrt{e_i e_j}$$

Table S4. Well depths and radii used for the Lennard-Jones calculations. The well depths are those used in the CHARMM19 parameter set (*12*). The radii were determined by fitting the Lennard-Jones potential to the distribution of distances observed between the atom types in the PDB.

Atom Type	Radii(*r*)	well depth (*e*)
1	2.00	0.1200
2	2.00	0.1200
3	2.00	0.0486
4	2.00	0.1142
5	2.00	0.1811
6	2.00	0.1200
7	1.75[1]	0.2384
8	1.75[1]	0.2384
9	1.75[1]	0.2384
10	1.75[1]	0.2384
11	1.75[1]	0.2384
12	1.75[1]	0.2384
13	1.55[1,2]	0.1591
14	1.55[2]	0.1591
15	1.55[2]	0.2100
16	1.90	0.1600
17	1.75	0.2384
18	2.00	0.0486
19	2.00	0.1400
20	1.55	0.1591
21	1.00[3]	0.0500
22	1.20	0.0500

23	1.20	0.0500
24	1.00[3]	0.0500

[1]These atom types are hydrogen bond donors and when paired with atom types that are hydrogen bond acceptors(13,14,15), r_{ij} is set to 2.95, the optimal distance for hydrogen bonding. This is to prevent the repulsive portion of the Lennard-Jones term from disfavoring hydrogen bonds.

[2]These atom types are hydrogen bond acceptors and when paired with atom types that are hydrogen bond donors (7,8,9,10,11,12,13) r_{ij} is set to 2.95.

[3]These are polar hydrogens and when paired with hydrogen bond acceptors (13,14,15), r_{ij} is set to 1.95.

Lazaridis-Karplus solvation model (E_{solv})

An implicit solvation model developed by Lazaridis and Karplus is used to evaluate the solvation energy of a protein (*13*).

$$E_{solv} = \sum_{i}^{natom} \sum_{j>i}^{natom} \left\{ \frac{-2\Delta G_i^{free}}{4p\sqrt{p}\, l_i r_{ij}^2} \exp(-d_{ij}^2) V_j - \frac{2\Delta G_j^{free}}{4p\sqrt{p}\, l_j r_{ij}^2} \exp(-d_{ji}^2) V_i \right\}$$

d_{ij} and r_{ij} are the same as in E_{atr}, ΔG^{free} is related to the solvation energy of the fully solvated atom, λ_i is a correlation length, and V is atomic volume. The values for the parameters are taken from Lazaridis and Karplus, except some of the ΔG^{free} values have been perturbed to better reproduce the relative frequencies amino acids are placed in the core versus the surface during design experiments (Table S5). We have left out the

intrinsic solvation energy of each atom because the sum of these values is a constant for each amino acid and can be incorporated into the reference energies.

Table S5. Parameters for the Lazaridis-Karplus solvation model.

Atom Type	DG^{free}	V	l
1	0.00	14.7	3.5
2	-1.40	8.3	3.5
3	-0.25	23.7	3.5
4	0.52	22.4	3.5
5	1.50	30.0	3.5
6	0.08	18.4	3.5
7	-8.9	4.4	3.5
8	-4.0	4.4	3.5
9	-7.8	11.2	3.5
10	-20.0	11.2	6.0
11	-11.0	11.2	6.0
12	-1.55	0.0	3.5
13	-6.77	10.8	3.5
14	-7.8	10.8	3.5
15	-10.0	10.8	6.0
16	-4.1	14.7	3.5
17	-5.0	4.4	3.5
18	1.00	23.7	3.5
19	1.00	14.7	3.5
20	-5.00	10.8	3.5
21	0.00	0.0	3.5
22	0.00	0.0	3.5
23	0.00	0.0	3.5
24	0.00	0.0	3.5

Rotamer Self-energy (E_{rot})

$$E_{rot} = \sum_{i}^{nres} -ln(P(rot(i) \mid phi(i), psi(i))$$

E_{rot} represents the internal energy of a rotamer and was derived from Protein Data Bank statistics by observing the probability of a particular rotamer and amino acid for a given phi angle and psi angle. These probabilities were taken directly from Dunbrack and Cohen (*14*). During the final design simulations we also considered rotamers with chi angles perturbed from the most commonly observed chi angles (+/- 0.5 standard deviation). These sub-rotamers were penalized by assuming a gaussian distribution about the mean using tabulated variances from Dunbrack and Cohen.

Amino acid preferences for particular regions of phi, psi space ($E_{aa|phi,psi}$)

A non-redundant set of PDB files were used to determine the probabilities for observing each of the 20 amino acids within 10°x10° bins in phi,psi space, P(aa,|phi,psi). The energy was calculated by taking the negative log of the probabilities.

Amino acid dependent torsion potential for phi and psi (E_{rama})

For each of the 20 amino-acid types in each of three secondary structure types (helix, strand, and other as defined by DSSP), the frequency of (phi,psi) pairs was determined for 10°x10° bins. Probabilities were calculated using added pseudocounts, and the potential calculated by taking the log of the interpolated probabilities.

Residue pair potential (E_{pair})

$$E_{pair} = \sum_{i}^{nres} \sum_{j>i}^{nres} \frac{P(aa_i, aa_j \mid d_{ij}, env_i, env_j)}{P(aa_i \mid d_{ij}, env_i) P(aa_j \mid d_{ij}, env_j)}$$

E_{pair} is derived from the probability of seeing two amino acids close together in space in the PDB database after accounting for the intrinsic probabilities of these amino acids to be in that environment (*15*). Two classes of environments are considered, buried and exposed, and five distance bins were used, 0-4.5, 4.5-6.0, 6.0-7.5, 7.5-9.0 and 9.0-10.5. This term was only evaluated between polar amino acids. The distances were measured between the action centers on each residue, e.g. the nitrogen on the lysine sidechain.

Orientation-dependent hydrogen bonding term (E_{bb_hbond}, E_{sc_hbond}, $E_{bb\_sc\_hbond}$)

The energy of backbone-backbone, sidechain-backbone and sidechain-sidechain hydrogen bonds were determined using a function derived from the distances and angles observed for naturally occurring hydrogen bonds in the PDB database. This function is described in detail in the supporting material of Kortemme & Baker (*16*). In this study we did not weight the strength of the hydrogen bonds according to their degree of burial. We removed this weight to encourage hydrogen bonds at positions that are partially buried.

Energy of the unfolded state (E_{ref})

$$E_{ref} = \sum_{i}^{nres} W_{ref}(aa(i))$$

To approximate the energy of the unfolded state each amino acid is assigned a empirically determined reference energy.

Setting the weights

The weights on these terms and the 20 reference energies were determined by maximizing the product of $\exp(-E(aa_{obs}))/(?\ \exp(-E(aa_i)))$ over a training set of 30 proteins using a conjugate-gradient-based optimization method, where $E(aa_{obs})$ is the energy of the native amino acid at a position and the partition function in the denominator is over all 20 amino acids at each position. In this process only one residue was changed at a time and all other residues were kept in their native conformation. Subsequently the parameters were refined slightly on the basis of the results of complete redesign calculations on the training-set proteins. The weights used for this study are given in table S6. The reference values are dramatically different than we have used previously because we have removed the intrinsic solvation energy of an atom from the Lazaridis-Karplus solvation energy.

Table S6: Weights used for the energy function.

W_{atr}	0.80	W_{ref}(Ile)	-0.45
W_{rep}	0.65	W_{ref}(Lys)	1.30
W_{sol}	0.65	W_{ref}(Leu)	-1.62
W_{pair}	0.65	W_{ref}(Met)	0.25
W_{bb_hbond}	0.80	W_{ref}(Asn)	0.17
W_{sc_hbond}	1.10	W_{ref}(Pro)	-0.30
$W_{sc\_bb\_hbond}$	1.10	W_{ref}(Gln)	0.14
W_{rot}	0.70	W_{ref}(Arg)	0.60

W_{ref}(Ala)	0.05	W_{ref}(Ser)	0. 14
W_{ref}(Asp)	1.43	W_{ref}(Thr)	0.31
W_{ref}(Glu)	1.44	W_{ref}(Val)	-0.25
W_{ref}(Phe)	-1.20	W_{ref}(Trp)	-1.85
W_{ref}(Gly)	0.37	W_{ref}(Tyr)	-1.08
W_{ref}(His)	-1.92		

References

S1. S. Doublie, *Methods in Enzymology* **276**, 523 (1997).
S2. S. Mori, C. Abeygunawardana, M. O. Johnson, P. C. van Zijl, *J Magn Reson B* **108**, 94 (1995).
S3. W. A. Hendrickson, *Science* **254**, 51 (1991).
S4. Z. Otwinowski, W. Minor, in *Macromolecular crystallography* W. Charles, W. Carter, R. M. Sweet, Eds. (Academic Press, San Diego, 1997), vol. 276, pp. 307.
S5. C. R. Kissinger, D. K. Gehlhaar. (Agouron Pharmaceuticals, La Jolla, CA, 1997).
S6. D. E. McRee, *J. Structural Biology* **125**, 156 (1999).
S7. T. A. Jones, J.-Y. Zou, S. W. Cowan, M. Kjeldgaard, *Acta Cryst A* **47**, 110 (1991).
S8. G. J. Keywegt, T. A. Jones, in *From First Map to Final Model* S. Bailey, R. Hubbard, D. Waller, Eds. (SERC Daresbury Laboratory, Warrington, U.K., 1996) pp. 59
S9. R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, *J. Appl. Cryst.* **26**, 283 (1993).
S10. E. A. Merritt, D. J. Bacon, *Macromolecular Crystallography, Pt B* **277**, 505 (1997).
S11. N. Guex, M. C. Peitsch, *Electrophoresis* **18**, 2714 (1997).
S12. E. Neria, S. Fischer, M. Karplus, *J. Chem. Phys.* **105**, 1902 (1996).
S13. T. Lazaridis, M. Karplus, *Proteins: Struct. Func. Genet.* **35**, 132 (1999).
S14. R. L. Dunbrack, F. E. Cohen, *Protein Sci* **6**, 1661 (1997).
S15. K. T. Simons *et al.*, *34*, 82 (1999).
S16. T. Kortemme, A. V. Morozov, D. Baker, *J Mol Biol* **326**, 1239 (2003).

Supporting Online Material
www.sciencemag.org
Materials and Methods [Experimental and Computational]
Tables S1, S2, S3, S4, S5, S6