Supporting Information

Title: Reconstructing the remote origins of a fold singleton from a flavodoxin-like ancestor

Author list: Saacnicteh Toledo-Patiño (1,2), Manish Chaubey (2), Murray Coles (2) & Birte Höcker (1,2)

Author Affiliations:
1) Department of Biochemistry, University of Bayreuth, 95447 Bayreuth, Germany
2) Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Contents **Page**

Experimental details

Generation of sequence-based profiles and their comparison. Sequence-based profiles were built for all sequences with known structure corresponding to the hemD-like and flavodoxin-like folds contained in the astral database release 2.07 (Chandonia et al., 2019). First, multiple sequence alignments were generated with PSI-BLAST (Altschul et al., 1997) and further employed to build the profiles following the build.pl protocol described elsewhere (Söding 2005). Profile comparisons were carried out with the secondary structure prediction function switched off to perform strictly sequence-based alignments. Standard parameters were employed as described earlier (Söding 2005). Obtained alignments were sorted according to their sequence identity. The best ranked alignment, corresponding to *Thermus thermophilus* LitR and *Pseudomonas aeruginosa* U3S, were employed to define the truncated proteins and the sequence used as template for further modifications.

Cloning of *cU3S* and *cU3S*Δ. The *cU3S* gene fragment was amplified from plasmid pET28a-PA5259-U3Spa (Moynie et al., 2013) by PCR using 5'-ATA TCG **CAT ATG** GAT CCG AAA GTG CTG ATC ATG CGC G-3' (cU3S_fwd) with a NdeI site (in bold) as the 5' primer and 5'-AAT **CTC GAG** GGC GGC GCT CGT-3' (cU3S_rev) with a XhoI site (in bold) as the 3' primer. The cU3SΔ gene presenting the 6 amino acid deletion was amplified via gene assembly in several steps. Two fragments were amplified: fragment A cU3S_fwd as 5' primer and 5'- CGG CCG GAT AGT CGA GTG GCA GGT A-3' as 3' primer and fragment B using 5'-TAC CTG CCA CTC GAC TAT CCG GCC G-3' as 5' primer and cU3S_rev as 3' primer. A and B were mixed in equimolar amounts for a last PCR reaction, using the primers cU3S_fwd and cU3S_rev yielding the shortened *cU3S*Δ construct, which lacks the nucleotides coding for the six β-bridge amino acids. The resulting genes *cU3S* and *cU3S*Δ were cloned into pET21a yielding the constructs pET21a-cU3S and pET21a-cU3SΔ. The constructs were sequenced entirely to exclude any inadvertent PCR mutations.

Heterologous Expression and Purification of proteins. The U3S and U3SΔ proteins, which both carry a His_6-tag at their C-termini, were produced in *E. coli* BL21(DE3) in LB media at 20°C. Protein expression was induced at OD_{600}=0.6 by adding isopropyl-β-thiogalactopyranoside to a final concentration of 1 mM, and growth was allowed for 16 h. For NMR structure determination the expression was carried out in ^{15}N-labeled M9 minimal media. Cells were harvested by centrifugation and disrupted in presence of EDTA-free protease inhibitors (SERVA®) by sonication with a Bandelin DH 3100 (Sonoplus) sonicator (2x3 min, 40% amplitude (0.2 sec pulse and 0.8 sec pause), on ice) and the resulting homogenate centrifuged (18.000 rpm, 1 h, 4°C). The cleared lysate was passed through a 0.2 µM pore-sized filter and loaded onto a Ni^{2+}-HisTrap HP 5 ml-column for affinity chromatography. The proteins were eluted with an imidazole gradient in 50 mM KP buffer at pH 8 and 150 mM KCl. The protein was purified further via a preparative Superdex^{TM} S75 gel filtration in 50 mM KP at pH8 and 300 mM KCl.

Analytical methods. The buffer for biophysical characterization was 50 mM KP at pH 8 and 150 mM KCl. The purity of the proteins was checked by electrophoresis on 15% polyacrylamide gels. The protein concentrations were determined photometrically using molar extinction coefficients calculated from the amino acid sequence. Analytical gel filtration was performed on a calibrated analytical Sephadex^{TM} S75 column with a flow rate of 0.5 ml/min. Circular dichroism spectra were recorded with a J-810 CD-spectrometer (Jasco) in a 1 mm cuvette at room temperature. Temperature-induced unfolding was analyzed by following the far-UV CD signal at 222nm at slowly increasing temperatures (1°C/min). The protein concentrations used were 0.2 mg/ml.

NMR structure elucidation. Labeled cU3SΔ was concentrated to 2.9 mM in 50 mM KP and 150 mM KCl 80% H2O/20% D2O (pH 8). All spectra were recorded at 288 K on Bruker AVIII-600 and AVIII-800 spectrometers. Backbone sequential assignments were completed using standard triple resonance experiments implemented using selective proton flip-back techniques for fast pulsing (Diercks et al., 2005). Aliphatic sidechain assignments were completed by a combination of CCH-COSY and CCH-TOCSY experiments, while aromatic assignments were made by linking aromatic spin systems to the respective CH2 protons in a 2D-NOESY spectrum, combined with a PLUSH-TACSY experiment (Carlomagno et al., 1996). Stereospecific assignments and the resulting rotamer assignments were determined from an HNHB experiment and NOESY cross-peak patterns. Distance data were derived from 3D15N-HSQCNOESY and 3D-NNH-NOESY spectra on a ^{15}N-labeled sample, and 3D13C-HSQC-NOESY and 3D-CCH- and 3D-CNH-NOESY spectra (Diercks et al., 1999) on the ^{15}N,^{13}C-labeled sample. Aromatic contacts were observed in a ^{15}N-filtered 2D-NOESY spectrum. Structural restraints were compiled using a protocol aimed at high local accuracy using expectation NOESY spectra to test local conformational hypotheses (in-house software). Chemical shift similarity searches using TALOS (Cornilescu et al., 1999) were used to generate hypotheses for backbone conformations, while sidechain rotamers were searched exhaustively. Conformations identified in this manner were applied via dihedral restraints, using the TALOS-derived tolerances for backbone and ±30° for sidechains. Further NOE contacts were assigned iteratively using back-calculation of expectation NOESY spectra from preliminary structures. NOESY crosspeaks in the

three-dimensional spectra were converted into distance ranges after rescaling according to corresponding HSQC intensities. Crosspeaks were divided into four classes, which resulted in restraints on upper distances of 2.7, 3.2, 4.0 and 5.0 Å, respectively. Additional classes of 3.6 and 4.5 Å were used for medium and weak backbone contacts often affected by spin-diffusion. Lower distance restraints were included for very weak or absent sequential HN-HN crosspeaks using a minimum distance of 3.5 Å and medium intensity or weaker sequential and intraresidue HN-H# crosspeaks using a minimum distance of 2.7 Å. Allowances for the use of pseudoatoms (using averaging) were added for methyl groups and non-stereospecifically assigned methylene groups. Hydrogen bond restraints were applied for residues in secondary structure where donor–acceptor pairs were consistently identified in preliminary calculations. The restraints were applied via inclusion of pseudo-covalent bonds between heavy atom acceptors and hydrogen donors, with force constants of 14 kcal/Å2 and 8 kcal/rad2 on bond lengths and angles, respectively. Structures were calculated with XPLOR (NIH version 2.9.3; Schwieters et al., 2006; 2003) using a three-stage simulated annealing protocol. The second stage included a conformational database potential, while the third used a relaxed force constant on peptide bond planarity. Sets of 50 structures were calculated and a final set of 20 chosen on the basis of lowest restraint violations. An average structure was calculated and regularized to give a structure representative of the ensemble. Statistics for the final structure set are presented in Supplementary Table S3. A preliminary structure for cU3SΔ was also calculated during validation of the recently published CoMAND method of NMR structure determination (ElGamacy et al., 2019). Briefly, this method uses spectral decomposition of CNH-NOESY spectra to derive local conformational parameters, providing input for de novo folding routines (in this case Rosetta; Das & Baker, 2008). Convergence is monitored by a quantitative R-factor expressing the match between back-calculated CNH-NOESY spectra and the experimental spectra. This structure therefore provides independent confirmation of the cU3SΔ fold, with an RMSD of 1.98 Å to the refined structure presented here (Supplementary Figure S4).

Structural analysis. Structural alignments of cU3S NMR structure towards members of the flavodoxin-like fold were assessed employing DALI (Holm & Laakso, 2016), an online bioinformatic tool that automatically generates pairwise alignments of a query structure against a database, in this case PDB90. Structural alignments of the U3S halves against each other and the cU3S NMR structures were assessed with PDBeFold (Krissinel & Henrick, 2004) online server for secondary structure matching. Default parameters were used.

Altschul, S.F.; Madden, T.L., Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 1997, 25 (17), 3389-3402.

Carlomagno, T.; Maurer, M.; Sattler, M.; schwendinger, M.G.; Glaser, S.J.; Griesinger, C. PLUSH TACSY: Homonuclear planar TACSY with two-band selective shaped pulses applied to C(alpha), C' (beta), C (aromatic) correlations. J Biomol NMR 1996, 8, 161

Chandonia, J.M.; Fox, N.K.; Brenner, S.E. SCOPe: Classification of Large Macromolecular Structures in the Structural Classification of Proteins-Extended Database. Nucleic Acids Res. 2019, 47 (D1), D475–81.

Cornilescu, G.; Delaglio, F.; Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR 1999, 13, 289.

Das, R.; Baker, D. Macromolecular Modeling with Rosetta. Annu. Rev. Biochem. 2008, 77, 363.

Diercks, T.; Daniels, M.; Kaptein, R. Extended flip-back schemes for sensitivity enhancement in multidimensional HSQC-type out-and-back experiments. J Biomol NMR 2005, 33, 243.

Diercks, T.; Coles, M.; Kessler, H. An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. J Biomol NMR 1999, 15, 177.

ElGamacy, M.; Riss, M.; Zhu, H.; Truffault, V.; Coles, M. Mapping Local Conformational Landscapes of Proteins in Solution. Structure 2019, 27, 1.

Holm, L.; Laakso, L. M. DALI Server Update. Nucleic Acids Res. 2016, 44 (W1), W351 –W355.

Krissinel, E.; Henrick, K. Secondary-Structure Matching (SSM), a New Tool for Fast Protein Structure Alignment in Three Dimensions. Acta Crystallogr. D, Biol Crystallogr 2004, 60 (Pt 12 Pt 1): 2256–68.

Moynie, L.; Schnell, R.; McMahon, S.A.; Sandalova, T.; Boulkerou, W.A.; Schmidberger, J.W.; Alphey, M.; Cukier, C.; Duthie, F.; Kopec, J.; Liu, H.; Jacewicz, A.; Hunter, W.N.; Naismith, J.H.; Schneider, G. The AEROPATH Project Targeting Pseudomonas Aeruginosa: Crystallographic Studies for Assessment of Potential Targets in Early-Stage Drug Discovery. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 2013, 69 (Pt 1): 25–34.

Schwieters, C.D.; Kuszewski, J.J.; Clore, G.M. Using Xplor-NIH for NMR molecular structure determination. Progr. NMR Spectroscopy 2006, 48, 47.

Schwieters, C.D.; Kuszewski, J.J.; Tjandra, N.; Clore, G.M. The Xplor-NIH NMR molecular structure determination package. J. Magn. Reson. 2003, 160, 65.

Söding, J. 2005. "Protein Homology Detection by HMM-HMM Comparison." *Bioinformatics* 21 (7): 951–60.

FIGURE S1: Sequence alignment of best scored hemD-like half with flavodoxin-like protein.
C-terminal half of uroporphyrinogen III synthase (U3S) from *Pseudomonas aeruginosa (cU3S, orange)* aligned to B12-binding domain of LitR of *Thermus thermophilus* (green) and the modified cU3SΔ sequence lacking the insert (gray).

```
cU3S        HDPKVLIMRGEG-----GREFLAERLRGQGVQVDYLPLYRRRAPDYPAGELLARVRAERL   55
cU3SΔ       HDPKVLIMRGEG-----GREFLAERLRGQGVQVDYLPL------DYPAGELLARVRAERL   49
LitR        -GPPVLVTTPPGERHEIGAMLAAYHLRRKGVPALYLGP------DTPLPDLRALARRLGA   53
             .* **:    *     *  : * :** :** . **        * *  :* * .*

cU3S        NGLVVSSGQGLQNLYQLAAADEIGRLPLFVPS-PRVAEMARELGAQRVIDCRGASAPALL  114
cU3SΔ       NGLVVSSGQGLQNLYQLAAADEIGRLPLFVPS-PRVAEMARELGAQRVIDCRGASAPALL  108
LitR        GAVVLSA-VLSEPLRALPD-GKDLAPRVFLGGQGAGPEEARRLGAEYME-----DLKGLA  106
            ..:*:*:    : *  *   .:     :*: .     * **.***: :      .  .*

cU3S        AAL    117
cU3SΔ       AAL    111
LitR        EAL    109
             **
```

FIGURE S2: Stability measured by thermal denaturation.
0.2 mg/ml protein in 50mM KP (pH 8.0), 150 mM KCl (cU3S in grey and dashed, and cU3SΔ in orange, solid line) were incubated at increasing temperatures while monitoring the CD signal at 222nm in a 1-mm cuvette.

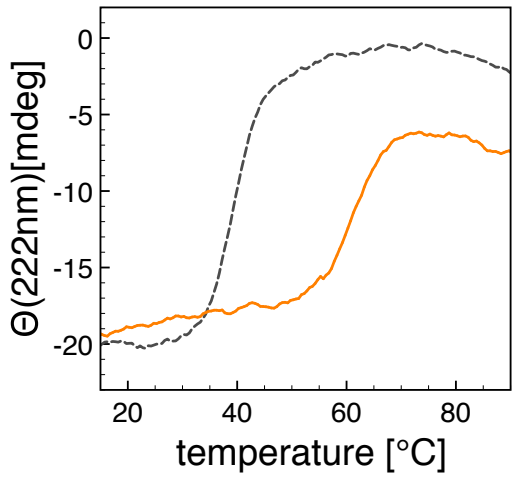
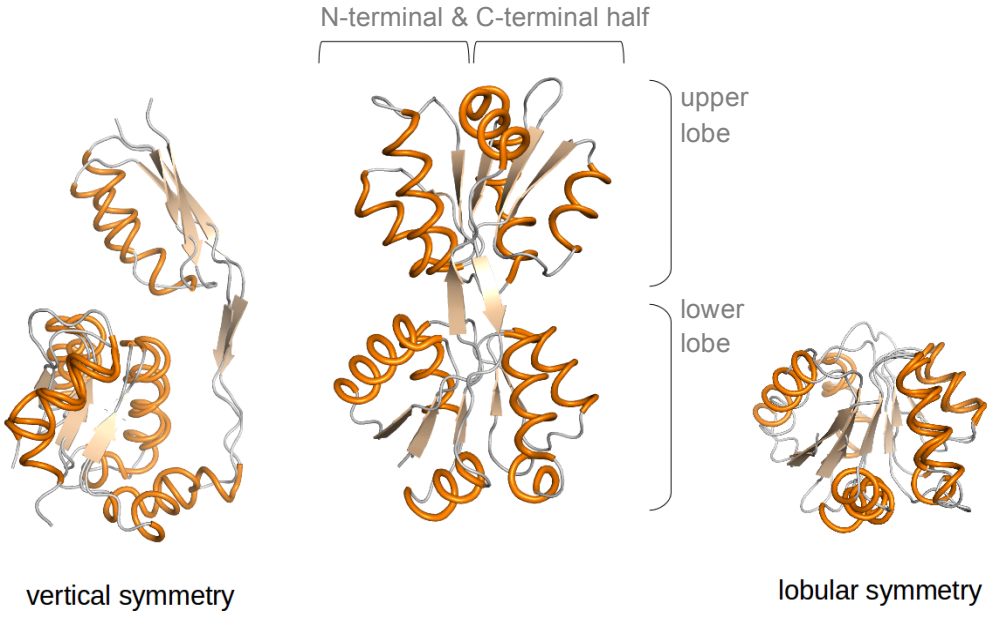


FIGURE S3: HemD-like fold symmetry.
The hemD-like fold has two distinct symmetry axes; both N- and C-terminal halves (left) as well as its lobes (right) can be aligned structurally. The example shown here is *P. aeruginosa* U3S (pdb 4es6). While the HemD-like halves can be superposed with RMSD values up to 2.1 over >100 residues, lobes align to each other only over <75 residues.



N- vs. C-terminal half
superposition = 108 of 124 res
RMSD = 2.4 and Z = 5.8
alignment length = 128 res
Identity = 18%
Similarity = 35%

lower vs. upper lobe
superposition = 67 of 120 res
RMSD = 2.0 and Z = 6
alignment length = 120 res
Identity = 26%
Similarity = 22%

FIGURE S4: NMR structure of cU3SΔ.
Overlay of the refined NMR structure of *cU3SΔ* (in purple) and a model (in orange) calculated during validation of the recently published CoMAND method of NMR structure determination (ElGamacy et al., 2019).

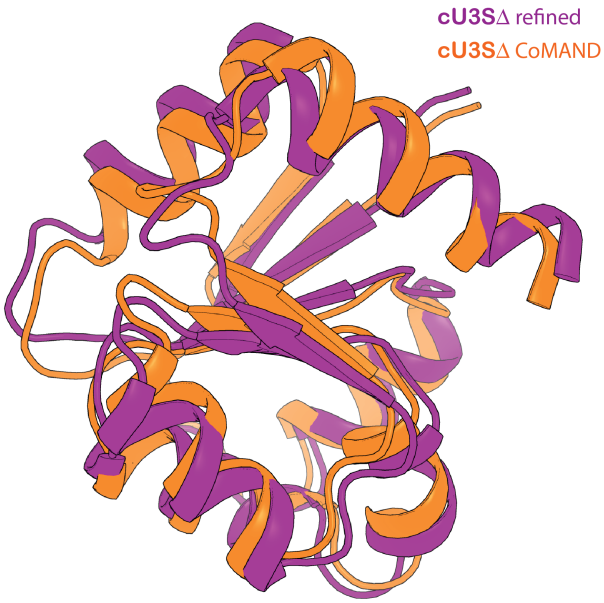
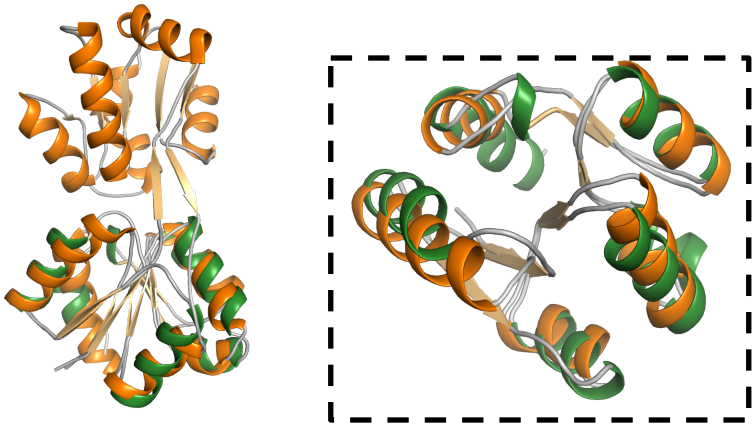
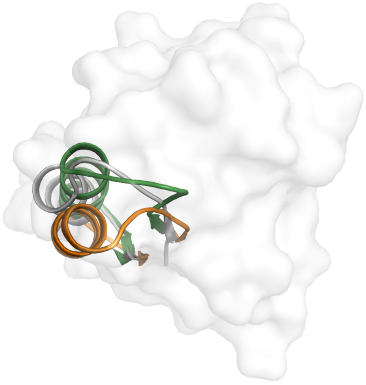


FIGURE S5: Structural analysis of cU3SΔ
(A) *cU3SΔ* NMR structure (green) aligned to the lower lobe of its parental protein U3S from *P. aeruginosa* (orange). (B) Structural alignment of N-terminal αβα-elements corresponding to LitR (green), U3S lower lobe (orange) and cU3SΔ (grey). (C) Remaining regions of the same proteins without the N-terminal αβα-elements. (D) Residue-residue contacts of the αβα-elements of cU3SΔ in its new accepting environment.

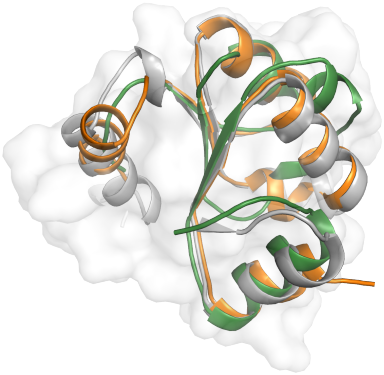
A)



B)



C)



D)

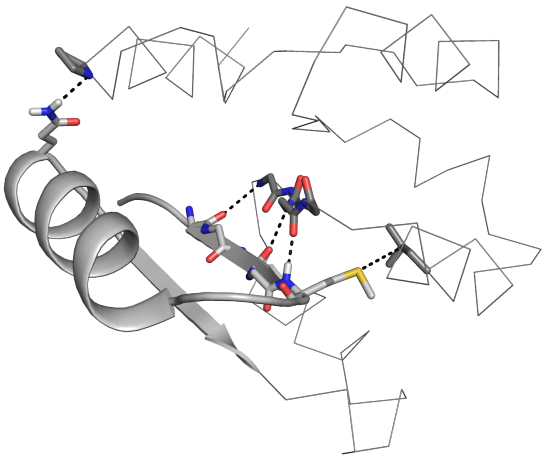


TABLE S1: Scores from profile-profile alignments using U3S full length sequence against pdb70. All N-terminal halves (blue) except one find their C-terminal counterparts (red) and those from different organisms with high probabilities and sequence identities up to 22%.

Query	Target	Probability	Sequence Identity	Aligned columns	Query start res	Query end res	Target start tres	Target end res
human	*Pseudomonas aeruginosa*	98.8	16	120	136	257	6	125
	Thermus thermophilus	98.7	17	118	136	257	33	154
	Pseudomonas syringae	98.7	19	118	136	255	14	131
	Thermus thermophilus HB8	98.6	18	125	129	257	1	129
	human	**98.5**	**14**	**125**	**131**	**258**	**16**	**153**
Thermus thermophilus HB8	*Pseudomonas aeruginosa*	99	18	120	131	254	6	125
	Thermus thermophilus	98.9	19	133	120	254	21	154
	Pseudomonas syringae	98.9	15	118	131	252	14	131
	Thermus thermophilus HB8	**98.8**	**21**	**129**	**124**	**254**	**1**	**129**
	human	98.8	17	121	6	129	155	278
Thermus thermophilus	***Thermus thermophilus***	**98.9**	**18**	**133**	**112**	**246**	**21**	**154**
	Pseudomonas syringae	98.9	15	119	123	245	14	132
	Thermus thermophilus HB8	98.8	21	129	116	246	1	129
	human	98.7	17	117	2	121	159	278
Shewanella amazonensis	*Pseudomonas aeruginosa*	99.1	14	124	115	239	2	125
	Thermus thermophilus	99.1	15	132	103	238	19	153
	Pseudomonas syringae	99	16	121	116	237	11	131
	Thermus thermophilus HB8	98.9	14	123	113	237	2	127
	human	98.7	11	124	113	238	15	151
Pseudomonas syringae	***Pseudomonas syringae***	**98.9**	**22**	**119**	**129**	**247**	**14**	**132**
	human	98.9	19	119	1	119	156	276
	Thermus thermophilus	98.9	14	128	118	248	22	154
	Thermus thermophilus HB8	98.7	15	118	1	119	131	252
Pseudomonas aeruginosa	***Pseudomonas aeruginosa***	**99.2**	**19**	**123**	**125**	**247**	**3**	**125**
	Thermus thermophilus	99.2	13	129	111	247	21	154
	Pseudomonas syringae	99.1	17	122	125	246	11	132
	human	99.1	17	126	1	128	157	284
	Thermus thermophilus HB8	99	19	115	1	116	132	250

TABLE S2: Structural alignment of cU3SΔ against PDB90 using DALI. The NMR structure matched U3S lower lobes up to with 2.2 A and Z scores up to 14. Proteins adopting a flavodoxin-like architecture are also matched with RMS up to 2.7 and Z scores up to 11.2.

No:	Chain	Z	rmsd	lali	nres	%id	Description
1:	4es6-A	14.0	2.2	111	249	70	UROPORPHYRINOGEN-III SYNTHASE;
2:	3re1-B	13.8	2.5	116	257	48	UROPORPHYRINOGEN-III SYNTHETASE;
3:	3mw8-A	11.7	2.5	110	237	28	UROPORPHYRINOGEN-III SYNTHASE;
4:	1wd7-B	11.3	2.7	117	255	14	UROPORPHYRINOGEN III SYNTHASE;
5:	2j48-A	11.2	3.1	109	119	13	TWO-COMPONENT SENSOR KINASE;
6:	1jr2-A	10.9	3.0	111	260	17	UROPORPHYRINOGEN-III SYNTHASE;
7:	4qpj-D	10.7	2.7	112	121	13	PHOSPHOTRANSFERASE;
8:	1xhe-B	10.3	2.7	109	122	15	AEROBIC RESPIRATION CONTROL PROTEIN ARCA;
9:	2zwm-A	10.1	2.9	111	120	17	TRANSCRIPTIONAL REGULATORY PROTEIN YYCF;
10:	4lda-B	9.9	3.5	117	128	18	TADZ;
11:	3ilh-A	9.9	2.8	114	133	11	TWO COMPONENT RESPONSE REGULATOR;
12:	2qr3-A	9.8	2.9	110	121	15	TWO-COMPONENT SYSTEM RESPONSE REGULATOR;
13:	6br7-B	9.8	3.2	113	125	17	TWO-COMPONENT SYSTEM RESPONSE REGULATOR PROTEIN;
14:	2zay-A	9.8	3.2	113	123	9	RESPONSE REGULATOR RECEIVER PROTEIN;
15:	3jte-A	9.8	3.3	113	126	12	RESPONSE REGULATOR RECEIVER PROTEIN;

TABLE S3: Solution structure statistics – PDB ID 6TH8, BMRB ID 34452.

	SA	$\langle SA \rangle_r$
Restraint Violations[1]		
Distance restraints (Å)		
All (676)	0.013 ± 0.001	0.013
Intra-residue (98)	0.004 ± 0.001	0.004
Inter-residue sequential (192)	0.016 ± 0.001	0.016
Medium range (85)	0.018 ± 0.001	0.016
Long range (225)	0.014 ± 0.001	0.015
H-bond (76)	0.000 ± 0.000	0.000
Persistent viol. thres.[2]	0.075	-
Dihedral restraints (°)		
All (347)	0.047 ± 0.005	0.045
Persistent viol. thres[2]	0.25	-
H-bond restraints[3]		
Distance (Å) (76)	2.18 ± 0.11	2.13 ± 0.12
Antecedent angle (°)	13.0 ± 5.1	13.7 ± 5.7
Covalent Geometry		
Bonds (Å × 10^{-3})	2.66 ± 0.03	2.65
Angles (°)	0.65 ± 0.01	0.65
Impropers (°)	1.27 ± 0.03	1.22
Structure Quality Indicators[4]		
Ramachandran Map (%)	99.1 / 0.9 / 0.0	99.2 / 0.8 / 0.0
Atomic R.M.S.D (Å)[5]		
	Backbone Heavy Atom	All Heavy Atom
SA vs $\langle SA \rangle$	0.27 ± 0.06	0.86 ± 0.06
SA vs $\langle SA \rangle_r$	0.36 ± 0.05	1.04 ± 0.08
$\langle SA \rangle$ vs $\langle SA \rangle_r$	0.24	0.72

[1] Violations are expressed as RMSD ± SD unless otherwise stated. Numbers in brackets indicate the number of restraints of each type.

[2] Persistent violations are defined as those occurring in at least 75% of all structures. The thresholds at which no persistent violations occur are tabulated.

[3] Hydrogen bonds were treated as pseudo-covalent bonds. Deviations are expressed as the average distance/average deviation from linearity for restrained hydrogen bonds.

[4] Defined as the percentage of residues in the favored/allowed/outlier regions of the Ramachandran map as determined by MOLPROBITY (Chen et al., 2010, *Acta Crystallogr. D Biol. Crystallogr.* 66:12; Davis et al., 2007, *Nucleic Acids Res.* 35:W375).

[5] Structures are labeled as follows: SA, the final set of 20 simulated annealing structures; $\langle SA \rangle$, the mean structure calculated by averaging the coordinates of SA structures after fitting over secondary structure elements; $\langle SA \rangle_r$, the structure obtained by regularizing the mean structure under experimental restraints. RMSD values were obtained based on superimpositions over ordered residue (defined as P3-L118).