Engineering soluble proteins for structural genomics

Jean-Denis Pédelacq[1], Emily Piltch[3], Elaine C. Liong[2], Joel Berendzen[2], Chang-Yub Kim[1], Beom-Seop Rho[1], Min S. Park[1], Thomas C. Terwilliger[1], and Geoffrey S. Waldo[1]*

Published online: 19 August 2002, doi:10.1038/nbt732

Structural genomics has the ambitious goal of delivering three-dimensional structural information on a genome-wide scale. Yet only a small fraction of natural proteins are suitable for structure determination because of bottlenecks such as poor expression, aggregation, and misfolding of proteins, and difficulties in solubilization and crystallization. We propose to overcome these bottlenecks by producing soluble, highly expressed proteins that are derived from and closely related to their natural homologs. Here we demonstrate the utility of this approach by using a green fluorescent protein (GFP) folding reporter assay to evolve an enzymatically active, soluble variant of a hyperthermophilic protein that is normally insoluble when expressed in *Escherichia coli*, and determining its structure by X-ray crystallography. Analysis of the structure provides insight into the substrate specificity of the enzyme and the improved solubility of the variant.

Pilot structural genomics projects have established the feasibility of rapid structure determination for a small fraction of natural proteins, but considerable bottlenecks remain for the vast majority. For example, a structural genomics project focused on the hyperthermophile crenarchaeon *Pyrobaculum aerophilum* showed that nearly 50% of the targeted proteins partitioned into inclusion bodies or insoluble aggregates when expressed in *Escherichia coli*[1,2]. Current strategies for overcoming these bottlenecks can be divided into two categories. One is to use native protein sequences and employ extensive testing to find conditions that yield sufficient amounts of soluble protein. The complementary strategy is to use protein sequences that have been modified to optimize the protein's suitability for structure determination while maintaining its native conformation.

The advantage of using native protein sequences is that the protein is likely to maintain its function, provided it has the correct post-translational modifications. The disadvantage is that time-consuming and expensive searches for expression[3], refolding[4,5], solubilization, and crystallization conditions are carried out at relatively late stages in the process, allowing only a limited number of trials for each targeted protein.

In contrast, the advantage of using modified protein sequences is that an extensive search for sequences suitable for high-throughput structure determination can be carried out at an early step at which far more possibilities can be readily tested. Our GFP-based *in vitro* evolution technique for engineering protein solubility can use simple colony-plating techniques to identify mutations that yield highly soluble protein[6,7]. The potential disadvantage of this approach is that the resulting proteins may sometimes be non-functional; however, extensive studies have shown that the structure of proteins with one or several mutations is generally very similar to that of the wild-type proteins[8,9].

We propose an efficient two-part strategy for structural genomics. First, target sequences are chosen from protein families or on the basis of important biochemical function. Second, in cases in which these targets are misfolded or insoluble and are recalcitrant to conventional solubilization methods such as refolding, genetic variants that are nearby in sequence space and that are suitable for high-throughput structure determination are identified and isolated by *in vitro* evolution or screening. Here we applied directed evolution using the GFP folding reporter[6] to improve the folding and solubility of a nucleoside diphosphate kinase (NDP-K) and two other recalcitrant proteins from the hyperthermophile *P. aerophilum*. We readily obtained crystals for the evolved, enzymatically active NDP-K containing six point mutations and solved its structure by x-ray diffraction. The fold and active-site constellation of the engineered protein are essentially identical to those of the closest structural homolog. This work demonstrates the utility of the GFP-based directed evolution approach for facilitating high-throughput structural genomics.

Results and discussion

Refolding trials of recalcitrant target proteins. We chose methyl transferase (MT), tartrate dehydratase β-subunit (TD-β), and nucleoside diphosphate kinase (NDP-K) from the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum* (Pa) as candidates for engineering more soluble variants. We have shown previously that these proteins were directed into inclusion bodies when expressed in *E. coli* at 37°C [6] (Fig. 1A). Expression at either 27°C or 10°C failed to improve soluble protein yields (data not shown). SDS–PAGE revealed that much of wild-type TD-β was truncated and that this fraction increased with expression time, suggesting proteolytic cleavage perhaps resulting from unstable or misfolded conformations (Fig. 1A). We reasoned that if these targets could be refolded, the full-length proteins, expressed with C-terminal His_6 tags, could be purified using metal-affinity[10] and size-exclusion[11] chromatography. Refolding screens[4,5] failed to identify conditions yielding useful amounts of either MT or TD-β. After processing 80 mg of inclusion bodies, we recovered less than 400 μg of MT or TD-β after refolding and metal-

[1]*Bioscience Division, MS-M888 and* [2]*Biophysics Group, MS-P244, Los Alamos National Laboratory, Los Alamos, NM 87545.* [3]*University of Rochester, Rochester, NY 14627.* *Corresponding author (waldo@telomere.lanl.gov).*

affinity purification. Refolding 80 mg of washed NDP-K inclusion bodies yielded ~2 mg of soluble protein after metal-affinity chromatography, but the refolded protein irreversibly aggregated during dialysis and was unsuitable for structural studies. SDS–PAGE revealed that 90% of the soluble refolded NDP-K with a C-terminal His_6 motif failed to bind the metal affinity resin. This result was unexpected as the C terminus of NPD-K is not buried according to existing structures[12]. Denatured NDP-K bound quantitatively to metal affinity resin in the presence of 8 M urea, suggesting that much of the solubilized refolded protein might remain misfolded.

Directed evolution of target proteins for increased solubility. The wild-type proteins were tested for folding ability using the GFP folding reporter assay[6]. Briefly, the genes were expressed in *E. coli* strain BL21(DE3) as N-terminal fusions with GFP, and the fluorescence of colonies was assessed by visualizing through a 520 nm long-pass filter using 488 nm illumination. All three proteins strongly interfered with the folding and chromophore formation of the fused GFP domain (Fig. 1B).

We screened libraries of mutants of the three proteins for closely related variants with improved folding and solubility. Each protein was subjected to forward evolution using the GFP reporter[6]. During each round of evolution, the 40 brightest colonies were screened from ~40,000 colonies for subsequent recombination by gene shuffling[6,13]. After four rounds of forward evolution, there was no further increase in GFP fusion fluorescence. We carried out two rounds of backcrossing[13] to remove non-essential mutations. We chose 16 of the optima for each of the three proteins for further analysis. Colonies expressing GFP fusions of these optima were much brighter than colonies expressing the wild-type fusions (Fig. 1B). Each was subcloned and expressed in *E. coli* at 37°C without the GFP tag, and the solubility was determined by SDS–PAGE as previously described[6]. The evolved MT, TD-β, and NDP-K were 50%, 95%, and 90% soluble, respectively (Fig. 1A). The evolved TD-β protein was expressed predominantly as the full-length product, and was not proteolyzed as had been previously observed. In marked contrast with the soluble refolded native NDP-K, the evolved NDP-K bound quantitatively to metal-affinity resin, indicating that the C-terminal His_6 motif was exposed to the solvent, consistent with correct folding.

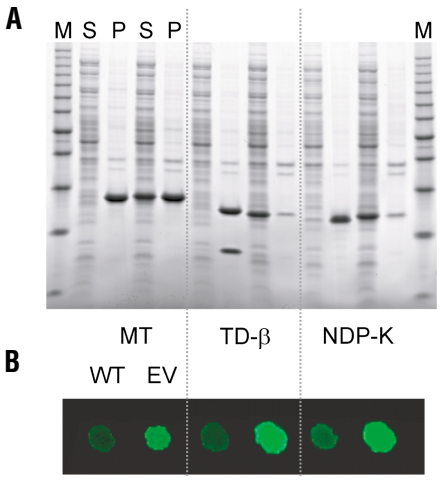

Figure 1. Protein solubility and fluorescence measurements. Solubility of wild-type (WT) and evolved (EV) recombinant *P. aerophilum* proteins MT, methyl transferase; TD-β, tartrate dehydratase β subunit; and NDP-K, nucleoside diphosphate kinase. (A) Coomassie-stained 4–20% gradient acrylamide SDS–PAGE of soluble (S) and pellet (P) fractions of proteins expressed without GFP tags at 37°C. M, 10 kDa molecular-weight ladder (Invitrogen). The lowest molecular-weight band is 10 kDa. (B) Photograph of *E. coli* colonies expressing GFP fusions of proteins at 37°C.

Table 1. Relative kinase activity measured as burst luminescence using a discontinuous luciferase–luciferin assay of generated ATP

NDP-Ks[a]	Substrate	Normalized burst luminescence[b] 25°C	50°C	75°C
Pa native	dGTP	30	20	25
	dTTP	360	2,870	6,630
	dATP	480	2,270	4,420
	dCTP	1,255	6,570	11,230
Pa evolved	dGTP	0.8	1.5	2.9
	dTTP	8.0	186.9	603.6
	dATP	36.0	251.5	646.8
	dCTP	107.8	653.9	1,408.4
Baker's yeast	dGTP	64,660	83,840	220
	dTTP	34,360	65,660	120
	dATP	62,640	100,000	0
	dCTP	20,220	28,300	0

[a]Approximate concentration (mg/ml) of proteins used in reactions with dNTP + ADP: Pa wild type refolded = 1.48×10^{-3}; Pa evolved = 1.23×10^{-2}; baker's yeast = 8.73×10^{-5}. Under these conditions ~25% of the best substrate is consumed at 50°C (Experimental Protocol).
[b]All measurements were taken in triplicate. Raw data blanks (no enzyme) corresponding to ADP plus the substrates (dGTP, dTTP, dCTP) yielded burst magnitudes of ~19 ± 2; the corresponding blank for ADP + dATP was 129 ± 4, indicative of substantial ATP contamination. Blank values were independent of incubation temperature. Blanks were subtracted and data normalized by dividing by the protein concentration (see above). A scaling factor (8.838×10^{-3}) was applied to all data to scale the highest-activity datum (baker's yeast with dATP at 50°C) to 100,000. Relative uncertainty, ~4%.

Structural analysis of evolved NDP-K. We obtained diffracting crystals of the evolved NDP-K. The structure is of interest for several reasons. First, in higher eukaryotes, the multimeric enzyme plays an important role in cell differentiation, oncogenic transformation, development, apoptosis, and signal transduction[14]. Second, amino acid residues important for catalytic activity have been identified in several X-ray structures with bound nucleotides[12], and the catalytic mechanism is well understood[15]. Third, Pa NDP-K contains unique sequence elements not found in any of the ~80 known NDP kinases (J.-D. Pédelacq, G.S. Waldo, and T.C. Terwilliger, unpublished data). Fourth, in preliminary studies, neither the refolded wild type nor the evolved Pa NDP-K optima had appreciable kinase activity using dGTP as the phosphate donor substrate. This result was puzzling given the preference of NPD-K for guanine nucleotides[12,15]. Finally, it was unclear why the wild-type Pa NDP-K was insoluble, as all eight native NDP kinases for which structures have been previously determined were expressed in soluble form[12]. The DNA sequences of the four most soluble Pa NDP-K clones were determined by dye-terminator sequencing. All four were identical, and contained the six amino-acid mutations A10D, G33D, E40K, R71Q, S107N, and I117N.

NDP-Ks share between 35% and 45% amino-acid sequence identity, and regions of high structural similarity can be superimposed with a root-mean-square deviation (r.m.s.d.) <1.5 Å. All monomeric subunits contain an α/β domain that is conserved in our evolved Pa protein structure[12]. Using DALI[16], we determined the closest structural homolog to be the human isoform Nm23-H2 (ref. 17; Protein Data Bank access code 1NSK). Of the 182 Cα atoms in the monomer of the Pa enzyme, 137 could be aligned to the α/β domain of Nm23-H2 with an r.m.s.d. of 1.1 Å. The remaining 45 residues are located mostly in two unique loops (Fig. 2, indicated in blue), increasing by 8–20% the accessible surface relative to the known NDP-K structures (J.-D. Pédelacq, G.S. Waldo, and T.C. Terwilliger, unpublished data).

Residues important for kinase activity[15] are conserved in both the wild-type and evolved Pa NDP-Ks (Fig. 3A), and their side-chain ori-

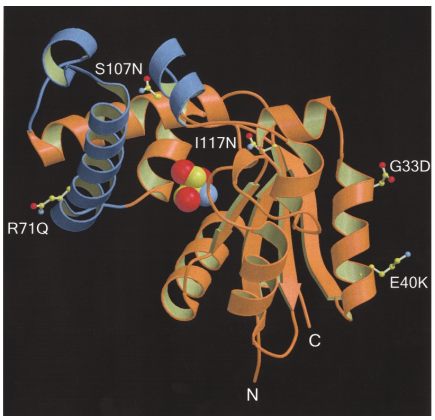

Figure 2. The NDP kinase fold. Ribbon representation illustrating the X-ray structure of the evolved Pa NDP-K monomeric subunit with a Tris molecule in the active site. Side chains of the mutated residues G33D, E40K, R71Q, S107N, and I117N are labeled. The 13 amino-terminal residues, including A10D, are highly disordered. Drawn with Molscript[31] and Raster 3D[32].

entations are very similar to those of the human isoform Nm23-H2 (Fig. 3B). The distances of each mutated residue from the closest atom of a GDP molecule modeled in the evolved Pa active site were deduced from a superimposition of the Pa Cα trace onto a Nm23-H2 monomer containing one GDP molecule in the active site. The distances range from 5.3 Å to 26.0 Å. N107, the closest residue, points away from the active site and is presumed not to perturb the enzyme's activity. Given the preference of the NDP-K enzyme family for guanine nucleotide substrates[12,15], it is surprising that under the conditions used in the *in vitro* assay[18], the Pa native and evolved enzymes have no substantial activity with dGTP. We predict that one of the novel loops could come into close contact (0.8 Å) with the NH_2 group of the adenine base of GDP (Fig. 3A). We also modeled TDP in the active site of the evolved Pa NDP-K based on the *Dictyostelium discoideum* structure with TDP[19]. The closest distance between TDP and the nearest residue of the Pa novel loop is 2.4 Å, suggesting that dTTP might be an acceptable substrate for the Pa NDP-K (Fig. 3C). We evaluated the kinase activity of Pa NDP-K using dTTP, dATP, and dCTP as phosphate-donor substrates. dCTP is the best substrate (Table 1), whereas cytosine nucleotides are the poorest substrates for other NDP-Ks[15], including the enzyme from baker's yeast. The absence of the NH_2 group may not be the only important factor for activity because dTTP and dCTP are not equally good substrates for the Pa NDP-K. We observed higher kinase activity at 75°C relative to 25°C or 50°C for the Pa enzyme, consistent with the hyperthermophilicity of *P. aerophilum*. In contrast, the NDP-K from baker's yeast was inactive at 75°C.

Structural clues to improved solubility of evolved Pa NDP-K. Given the substantially increased brightness of *E. coli* cells expressing the evolved Pa NDP-K–GFP fusion protein (Fig. 1B), the mutations must at least reduce interference with GFP folding and chromophore formation, possibly by eliminating off-pathway folding intermediates of the wild-type Pa protein. To evaluate the contribution of each mutation to improved folding and solubility, we generated the six single point mutants. Each was measurably brighter as a GFP fusion relative to the wild type (data not shown). We subsequently expressed the mutants without the fused GFP and determined the solubility of each using SDS–PAGE. All were insoluble with the exception of A10D and E40K, which were each ~10% soluble. This observation suggests that the mutations act synergistically, given that the evolved enzyme is mostly soluble. A detailed examination of the structure of the evolved Pa NDP-K reveals additional clues to its improved solubility. The mutations A10D, G33D, and I117N replace hydrophobic residues with charged or polar amino acids. Rational structure-guided site-directed substitutions of apolar hydrophobic residues with polar amino acids has been shown to improve protein solubility in some cases[20]. A10D is disordered and can be presumed to be accessible to solvent. I117N, which is deeply buried inside the monomer (Table 2), does not make any polar contact with the neighboring residues. G33D is buried in the dimer interface and makes hydrogen bonds to E40K (Fig. 4A).

To assess changes in the dimer interface caused by the mutations G33D and E40K, we compared the experimentally determined evolved structure with that of the native enzyme modeled *in silico*. Seven NDP-K structures, including the Pa enzyme, are hexameric in the crystal state. Only the *Myxococcus xanthus* enzyme was identified as a tetramer[21]. Both oligomeric forms result from the assembly of superimposable dimers[22]. G33 and E40 are strongly conserved in all NDP-Ks including the native Pa enzyme. Starting with the three-dimensional structure of the evolved Pa NDP-K dimer (Fig. 4A) and positioning each subunit as in Nm23-H2, we generated a model of the wild-type Pa dimer interface (Fig. 4B). Examination of the dimer interface suggests two reasons for the improved solubility of the evolved Pa NDP-K relative to the wild-type protein. First, in the evolved protein, K40 forms the apex of a pyramidal hydrogen-bond network involving the side-chain residues of D33, E34, and G30 from the symmetry-related subunit (Fig. 4A). In contrast, in the model of the wild-type protein, E40 makes only two hydrogen bonds

Table 2. Solvent accessibility of the mutated residues in the evolved Pa and corresponding residues in the human isoform Nm23-H2

Evolved Pa					Human Nm23-H2				
			Surface area (Å²)					Surface area (Å²)	
Residue	Chain	Total	Side chain	Main chain	Residue	Chain	Total	Side chain	Main chain
D33[a]	A	9.91	9.42	0.49	G22	A	0.00	0.00	0.00
	B	9.92	9.45	0.47		B	0.00	0.00	0.00
K40[a]	A	65.02	55.58	9.44	E29	A	30.40	14.82	15.58
	B	65.80	55.48	10.32		B	30.31	15.29	15.02
Q71[b,c]	A	65.58	65.58	0.00					
	B	65.81	65.81	0.00					
N107[c]	A	43.17	40.20	2.96	G63	A	44.98	34.32	10.66
	B	43.39	40.44	2.95		B	47.19	33.27	13.92
N117	A	2.06	2.06	0.00	V73	A	0.03	0.03	0.00
	B	1.68	1.68	0.00		B	0.01	0.01	0.00

[a]Surface area calculated for the dimer interface.
[b]Q71 is located on one of the unique loops of the Pa, and is absent from Nm23-H2.
[c]Q71 and N107 are both on the solvent-exposed surface of the monomer and the hexamer.

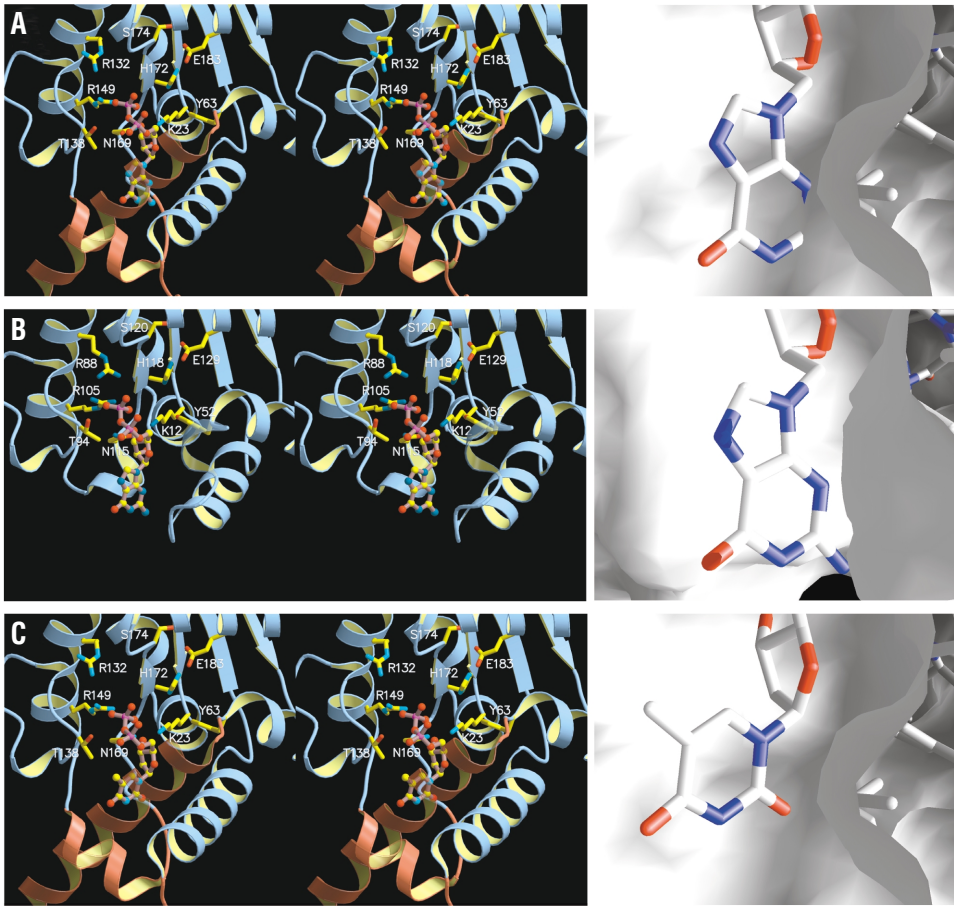

Figure 3. The nucleotide binding site. Stereo view of the NDP-K active site of (A) evolved Pa NDP-K with modeled GDP; (B) human isoform Nm23-H2 with bound GDP; and (C) evolved Pa NDP-K with modeled TDP. The corresponding molecular surface around the base is also shown. Residues important for catalytic activity are drawn as stick models and the dashed lines indicate the polar interactions (hydrogen bonds and salt bridges). The Pa novel loops are shown in orange. Drawn with Molscript[31], Raster 3D[32], and Grasp[33].

with the main-chain nitrogen residues in positions 32 and 33 (Fig. 4B). Second, the dimer interface of the evolved NDP-K shows a more favorable charge distribution relative to the wild-type protein. In the evolved NDP-K, K40 forms a patch of positive charge that balances the negative potential of D33 and E34 from the symmetry-related subunit (Fig. 4C). In contrast, in the model of the wild-type protein, E34 and E40 form a single large patch of negative charge across the dimer interface (Fig. 4D). The favorable hydrogen-bond network and charge distribution afforded by G33D–E40K should help direct the formation of correctly assembled dimers, thus suppressing the formation of nonspecific off-pathway aggregates. Compensatory mutations across subunit interfaces and the concomitant effects on protein solubility have been noted previously[7].

We have used the GFP folding reporter assay to evolve soluble variants of Pa NDP-K and two other proteins that could not be obtained in a soluble form by conventional means. Most importantly, the structure of the adapted, soluble Pa NDP-K indicates that the six mutations (A10D, G33D, E40K, R71Q, S107N, and I117N) have no appreciable effect on the overall fold of the protein. The side-chain positions of residues in the active site are conserved and the protein is active with the appropriate substrate. Remarkably, *in vitro* and natural evolution have converged to selection of the charge-balanced G33D and E40K pair, as illustrated in the evolved Pa NDP-K and the wild-type human isoform Nm23-H8C[23].

The approach described here is suitable for high-throughput engineering of protein solubility, and we plan to use it to determine the structures of proteins from *Mycobacterium tuberculosis* strain H37Rv (http://www.doe-mbi.ucla.edu/TB). The NDP-K described here is a relatively compact single-domain protein. Our work with *M. tuberculosis* will include larger, more complex multi-domain proteins. Since we performed the evolution of our NDP-K, we have, in collaboration with S.W. Suh and colleagues, evolved a triple mutant of the β-ketoacyl carrier protein (ACP) reductase homolog (Rv2002) from *M. tuberculosis* using the GFP method. All three mutations avoid the conserved regions among the reported sequences of ACP reductase and its homologs from *Mycobacterium* spp.[24].

Our method of engineering protein folding is independent of protein function. This is useful because nearly 40% of the predicted open reading frames of the *M. tuberculosis* strain H37Rv genome sequence represent proteins with no functional assignment (http://www.sanger.ac.uk/Projects/M_tuberculosis). In cases in which an assay for activity is available, our method can be combined with an activity screen to identify soluble, active variants. An ordered set of protein variants with decreasing levels of mutation can be generated by backcrossing the DNA encoding the evolved optimum with wild-type DNA by shuffling[6,13], cloning into the GFP folding reporter vector, and picking colonies with progressively less GFP-fusion fluorescence. Protein variants with desired combinations of activity, solubility, and level of mutation can then be more efficiently selected

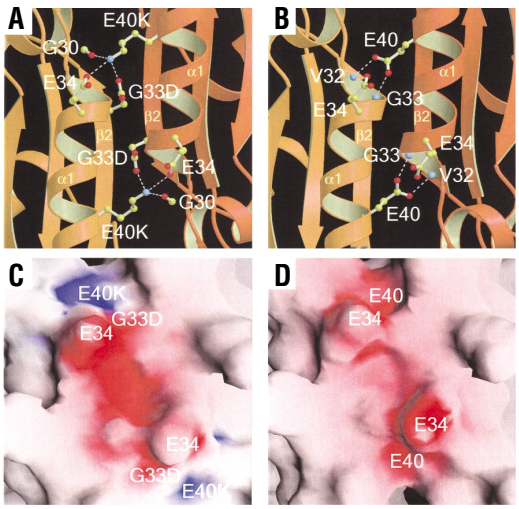

Figure 4. The dimer interface. Hydrogen-bond network of (A) evolved and (B) modeled wild-type Pa NDP-K proteins. Corresponding molecular surface of (C) evolved and (D) modeled wild-type Pa NDP-K proteins, colored according to electrostatic potential: uncharged (white), negative (red; Asp and Glu), and positive (blue; Lys). Both figures were generated with GRASP[33].

from this ordered set. Soluble protein variants provided by directed evolution methods should play an increasingly important role in structural proteomics.

Experimental protocol

Cloning, protein expression, and protein quantification. Cloning, expression of GFP-tagged fusion proteins, quantification of protein by densitometry of SDS–PAGE, and directed evolution were carried out substantially as previously described[6]. To facilitate purification, proteins were expressed with C-terminal His_6 tags. The 3′ end of the gene was cloned without a stop codon in frame with a *Bam*HI restriction site and the downstream His_6 motif, followed by the TAA stop codon. The resulting C-His_6-tagged proteins had the amino-acid extension GSHHHHHH.

Protein refolding. Washed inclusion bodies were prepared from pelleted *E. coli* liquid cultures (Luria–Bertani medium) using CelLytic B-II reagent (Sigma-Aldrich, St. Louis, MO) according to the manufacturer's instructions. Inclusion body proteins were unfolded using freshly prepared 8 M urea (~1 ml 8 M urea per 20 mg of inclusion body). After centrifugation (30,000*g*, 30 min) to remove insoluble debris, protein refolding screens were performed using the FoldIt Kit (Hampton Research, Laguna Niguel, CA) according to the manufacturer's instructions. Aggregation was monitored by measuring the concomitant increase in right-angle light scattering using a Perkin-Elmer LS-50B spectrofluorimeter with overlapping excitation and emission instrument settings (excitation 450 nm, emission 460 nm, each with 5 nm bandwidth), or by measuring the increase in absorbance at 450 nm[25]. Refolding buffers resulting in the minimum scattering were used to construct a second screen with a finer mesh. Final refolding buffers used for preparative refolding were: 0.05 M Tris(hydroxymethyl)aminomethane, pH 8.2, 0.15 M NaCl, 0.2 M guanidine for MT; 0.1 M Tris, pH 7.0, 0.05 M NaCl for TD-β; and 0.15 M Tris, pH 8.5, 0.15 M NaCl, 10% (vol/vol) glycerol for NPD-K. Denatured proteins were rapidly renatured by diluting 20-fold into refolding buffer to a final protein concentration of 0.1 mg/ml.

Production, purification, and crystallization of selenomethionine-substituted Pa NPD-K. The selenium-substituted protein was expressed in BL21(DE3) (Novagen, Madison, WI) using minimal medium supplemented with selenomethionine, six other amino acids, various salts, and sulfate[26]. Cells were grown to the mid-log phase at 37°C, induced with 1 mM isopropylthiogalactoside (IPTG) and then incubated for another six hours. The cell pellets were harvested by centrifugation at 4°C at 3,000 rpm (6,000*g*) for 20 min. The His_6-tagged selenomethionine-containing protein was purified by metal-affinity chromatography (Talon Superflow, Clontech, Palo Alto, CA) and gel filtration (Sephacryl S-100 HR, Amersham Biosciences, Piscataway, NJ). The protein was dialyzed against 100 mM Tris, pH 8.5, 15 mM β-mercaptoethanol (βME), and concentrated to ~7 mg/ml. The optimized crystallization conditions were found to be 10% (wt/vol) PEG 4000. Small plates appeared within two days, and reached their maximum size of $600 \times 600 \times 60\ \mu m^3$ in a week.

Data collection and structure determination. The crystals were cryocooled in liquid nitrogen after a rapid transfer in 30% (wt/vol) PEG 4000. Multiwavelength anomalous diffraction (MAD) data were collected to a resolution of 2.4 Å at beamline X8C at National Synchrotron Light Source (Upton, NY). Crystals belong to the monoclinic space group C2 with cell parameters $a = 125$ Å, $b = 72$ Å, $c = 105$ Å, $\beta = 133.3°$. Three monomers form the asymmetric unit of the crystal, two form a dimer through a non-crystallographic axis, and the third forms a dimer through a crystallographic two-fold axis. Details of the data processing, phasing, and refinement will be discussed elsewhere (J.-D. Pédelacq, G.S. Waldo, and T.C. Terwilliger, unpublished data). Briefly, the CCP4 suite of programs[27] was used to merge and scale these intensities and to compute the structure-factor amplitudes. Anomalously scattering-atom refinement and MAD phasing were conducted using the SOLVE package[28]. Experimental phases were improved by density modification using RESOLVE[29] and DM[27]. The model was refined to 2.5 Å using CNS[30], applying strict non-crystallographic constraints. The non-crystallographic symmetry restraints were relaxed during the later stages and the final cycle was carried out with no restraints. The 13 N-terminal amino-acid residues were highly disordered and were absent from the experimental electron density map. The crystallographic Rfactor and Rfree values are 0.22 and 0.28, respectively.

NDP kinase activity test. A luciferase assay kit (Sigma) was used to detect the ATP generated by the NDP kinase–catalyzed synthesis of ATP from dNTP and ADP. We reasoned that the hyperthermophilic Pa NDP kinase would exhibit optimal reaction kinetics at elevated temperatures incompatible with the luciferase enzyme, so we used a discontinuous version of a real-time assay[18]. Each 100 µl of ATP-generating reaction mix (0.1 M Tris, pH 8.5, 0.15 M NaCl, 10 mM $MgSO_4$ (Buffer A)) was equilibrated for 15 min at the target temperature (25°C, 50°C, or 75°C) with the test protein. A mix of 2 mM of the phosphate donor (dGTP, dTTP, dCTP, or dATP) and 2 mM ADP was added to initiate the reaction. The ATP luciferase–luciferin assay mix was prepared according to manufacturer's instructions. The blocking agent BSA was not stable at 75°C and was omitted. The concentration of NDP-K was maintained above 1.0×10^{-3} mg/ml to minimize wall-absorption losses. The activity was determined by measuring the burst luminescence produced when 100 µl of a 25-fold dilution of ATP luciferase–luciferin assay solution was added to 100 µl of the ATP-generating reaction using a Turner Designs (Sunnyvale, CA) Luminometer Model TD-20e. The specific activity of the baker's yeast as stated by the supplier (Sigma) was ~1,300 AU/ml.

Acknowledgments

We thank Leon Flaks for his assistance in performing data collection. We also thank James Jett, Andrew Bradbury, and Kathleen Sandman for review of the manuscript, and the National Institutes of Health and University of California Campus Laboratory Collaboration Program for generous support.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology website (http://www.nature.com/naturebiotechnology) for details.

Received 5 October 2001; accepted 7 May 2002

1. Christendat, D. *et al.* Structural proteomics: prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol.* **73**, 339–345 (2000).
2. Gaasterland, T. Archaeal genomics. *Curr. Opin. Microbiol.* **2**, 542–547 (1999).
3. Makrides, S.C. Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* **60**, 512–538 (1996).
4. Armstrong, N., De Lencastre, A. & Gouaux, E. A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein Sci.* **8**, 1475–1483 (1999).
5. Rudolph, R. & Lilie, H. *In vitro* folding of inclusion body proteins. *Faseb J.* **10**, 49–56 (1996).
6. Waldo, G.S., Standish, B.M., Berendzen, J. & Terwilliger, T.C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691–695 (1999).
7. Kim, C.A. *et al.* Polymerization of the SAM domain of TEL in leukemogenesis and transcriptional repression. *EMBO J.* **20**, 4173–4182 (2001).
8. Skinner, M.M. & Terwilliger, T.C. Potential use of additivity of mutational effects in simplifying protein engineering. *Proc. Natl. Acad. Sci. USA* **93**, 10753–10757 (1996).
9. Heinz, D.W., Baase, W.A. & Matthews, B.W. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc. Natl. Acad. Sci. USA* **89**, 3751–3755 (1992).
10. Porath, J. Immobilized metal ion affinity chromatography. *Prot. Exp. Purif.* **3**, 263–281 (1992).
11. Hagel, L., Lundstrom, H., Andersson, T. & Lindblom, H. Properties; in theory and practice; of novel gel-filtration media for standard liquid-chromatography. *J. Chromato.* **476**, 329–344 (1989).
12. Janin, J. *et al.* Three-dimensional structure of nucleoside diphosphate kinase. *J. Bioenerg. Biomemb.* **32**, 215–225 (2000).
13. Stemmer, W.P.C. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
14. De la Rosa, A., Williams, R.L. & Steeg, P.S. Nm23/nucleoside diphosphate kinase: toward a structural and biochemical understanding of its biological functions. *Bioessays* **17**, 53–62 (1995).
15. Lascu, I. & Gonin, P. The catalytic mechanism of nucleoside diphosphate kinases. *J. Bioenerg. Biomemb.* **32**, 237–246 (2000).
16. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
17. Morera, S., Lacombe, M.L., Xu, Y.W., Lebras, G. & Janin, J. X-ray structure of human nucleoside diphosphate kinase B complexed with GDP at 2 Å resolution. *Structure* **3**, 1307–1314 (1995).
18. Karamohamed, S., Nordstrom, T. & Nyren, P. Real-time bioluminometric method for detection of nucleoside diphosphate kinase activity. *Biotechniques* **26**, 728 (1999).
19. Cherfils, J., Morera, S., Lascu, I., Veron, M. & Janin, J. X-ray structure of nucleoside diphosphate kinase complexed with thymidine diphosphate and Mg2+ at 2 Å resolution. *Biochemistry* **33**, 9062–9069 (1994).
20. Dale, G.E., Broger, C., Langen, H., Darcy, A. & Stuber, D. Improving protein solubility through rationally designed amino acid replacements: solubilization of the

trimethoprim-resistant type S1 dihydrofolate reductase. *Protein Eng.* **7**, 933–939 (1994).
21. Williams, R.L. *et al.* Crystal structure of *Myxococcus xanthus* nucleoside diphosphate kinase and its interaction with a nucleotide substrate at 2 Å resolution. *J. Mol. Biol.* **234**, 1230–1247 (1993).
22. Lascu, I., Giartosio, A., Ransac, S. & Erent, M. Quaternary structure of nucleoside diphosphate kinases. *J. Bioenerg. Biomemb.* **32**, 227–236 (2000).
23. Lacombe, M.L., Milon, L., Munier, A., Mehus, J.G. & Lambeth, D.O. The human Nm23/nucleoside diphosphate kinases. *J. Bioenerg. Biomemb.* **32**, 247–258 (2000).
24. Yang, J.K. *et al.* Crystallization and preliminary X-ray crystallographic analysis of the Rv2002 gene product from *Mycobacterium tuberculosis*, a β-ketoacyl carrier protein reductase homologue. *Acta Crystallogr. D. Biol. Crystallogr.* **58**, 303–305 (2002).
25. Trivedi, V.D., Raman, B., Ramakrishna, T. & Rao, C.M. Detection and assay of proteases using calf lens β-crystallin aggregate as substrate. *J. Biochem. Biophys. Meth.* **40**, 49–55 (1999).
26. Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. & Clardy, J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124 (1993).
27. Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D. Biol. Crystallogr.* **50**, 760–763 (1994).
28. Terwilliger, T.C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D. Biol. Crystallogr.* **55**, 849–861 (1999).
29. Terwilliger, T.C. Reciprocal-space solvent flattening. *Acta Crystallogr. D. Biol. Crystallogr.* **55**, 1863–1871 (1999).
30. Brunger, A.T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D. Biol. Crystallogr.* **54**, 905–921 (1998).
31. Kraulis, P.J. Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
32. Merritt, E.A. & Bacon, D.J. Raster3D: Photorealistic molecular graphics. *Meth. Enzym.* **277**, 505–524 (1997).
33. Nicholls, A., Sharp, K.A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).