

# Structure and Evolutionary Analysis of a Non-biological ATP-binding Protein

Sheref S. Mansy<sup>1</sup>, Jinglei Zhang<sup>2</sup>, Rainer Kümmerle<sup>3</sup>, Mikael Nilsson<sup>2</sup>  
James J. Chou<sup>4</sup>, Jack W. Szostak<sup>1</sup> and John C. Chaput<sup>2\*</sup>

<sup>1</sup>Howard Hughes Medical Institute, and Department of Molecular Biology Massachusetts General Hospital, Boston, MA 02114 USA

<sup>2</sup>Center for BioOptical Nanotechnology, The Biodesign Institute, and Department of Chemistry and Biochemistry Arizona State University Tempe, AZ 85287-1604, USA

<sup>3</sup>Bruker BioSpin AG Fällanden, Switzerland

<sup>4</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston MA 02115, USA

We present a structural and functional analysis of the evolutionary optimization of a non-biological protein derived from a library of random amino acid sequences. A series of previously described *in vitro* selection experiments transformed a low-affinity ancestral sequence into a stably folded, high affinity ATP binding protein structure. While the evolutionarily optimized protein differs from its ancestral sequence through the accumulation of 12 amino acid mutations, the means by which those mutations enhance the stability and functionality of the protein were not well understood. We used a combination of mutagenesis, biochemistry, and NMR spectroscopy to investigate the structural and functional significance of each mutation. We solved the three-dimensional structure of the folding optimized protein by solution NMR, which revealed a fourth strand of the  $\beta$ -sheet of the  $\alpha/\beta$ -fold that was not observed in an earlier crystallographic analysis of a less stable version of the protein. The structural rigidity of the newly identified  $\beta$ -strand was confirmed by  $T_1$ ,  $T_2$ , and heteronuclear nuclear Overhauser enhancement (NOE) measurements. Biochemical experiments were used to examine point mutations that revert the optimized protein back to the ancestral residue at each of the 12 sites. A combination of structural and functional data was then used to interpret the significance of each amino acid mutation. The enhanced ATP affinity was largely due to the emergence of a patch of positive charge density on the protein surface, while the increased solubility resulted from several mutations that increased the hydrophilicity of the protein surface, thereby decreasing protein aggregation. One mutation may stabilize the hydrophobic face of the  $\beta$ -sheet.

© 2007 Elsevier Ltd. All rights reserved.

\*Corresponding author

**Keywords:** mRNA display; solution NMR; *in vitro* evolution

## Introduction

We previously used mRNA display<sup>1</sup> to isolate a non-biological ATP-binding protein from an unbiased random-sequence library.<sup>2</sup> The library from which our ancestral ATP-binding protein was isolated consisted of  $>10^{12}$  unique proteins each containing a central random region of 80 resi-

dues. After eight rounds of mRNA display based *in vitro* selection and amplification, four families of ATP binding proteins were identified from the starting library. None of these four families exhibited any significant sequence homology to any other known amino acid sequence found in nature. One of these protein families, family B, was further optimized for improved binding activity through mutagenesis and several additional rounds of selection and amplification.<sup>2</sup> Analysis of the protein sequences from this selection revealed two conserved CXXC motifs that bound a single zinc-metal ion. Complete biophysical characterization of the binding optimized proteins from the output of this selection was complicated by the presence of additional amino acid sequences outside the core-binding domain of the selected ATP-binding protein.

Abbreviations used: NOE, nuclear Overhauser enhancement; MBP, maltose binding protein; ABP, ATP binding protein; RCD, residual dipolar coupling; FOB, folding optimized family B.

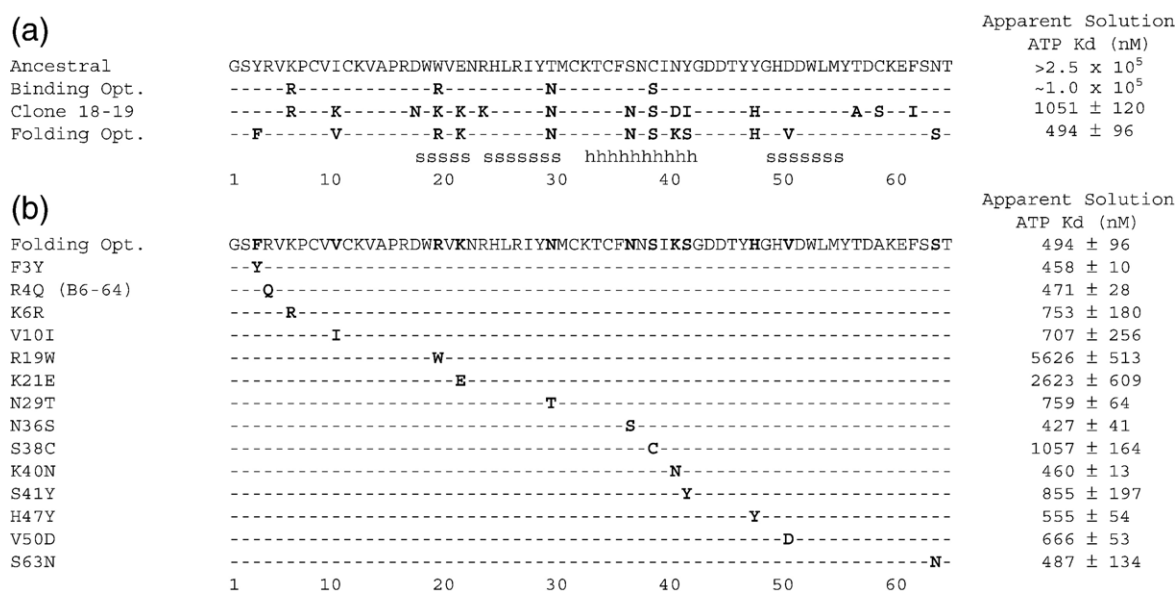
E-mail address of the corresponding author: [john.chaput@asu.edu](mailto:john.chaput@asu.edu)

Identification of the core-binding domain by the characterization of truncation constructs allowed for the *in vivo* expression and isolation of monomeric protein in pure form. The X-ray crystal structure of one highly divergent variant (protein 18–19) from this selection was reported<sup>3</sup> and reveals a novel  $\alpha/\beta$  fold not found in biology with a metal binding site remarkably similar to the treble clef zinc binding motif.<sup>4</sup>

All of the variants examined required high concentrations of free ATP in order to remain stably folded and soluble. In an attempt to overcome this limitation, mRNA display was used to select for new protein variants that remained stably folded in the absence of excess ATP ligand.<sup>5</sup> The selection strategy involved using guanidine hydrochloride to enrich for protein variants that remained folded, and therefore functional, in the presence of increasing concentrations of chemical denaturant. For each round of selection, the guanidine hydrochloride concentration was gradually increased to ensure that less than 10% of the input into each round of selection was recovered from the previous round. After six rounds of selection, many of these folding optimized ATP binding proteins remained water soluble and monomeric when expressed in *Escherichia coli*, and bound ATP with high affinity and specificity. We were then able to identify a 62 amino acid residue core domain of this protein that retained the same ATP binding function as the full-length construct. Comparison of the consensus sequences for each stage in the evolution of this protein (Figure 1(a)), from ancestral sequence to the folding optimized

variant, revealed the location and identity of 12 amino acid substitutions that arose during the *in vitro* evolution of this protein from a low affinity ancestral state to a well folded high affinity structure.

We used structural and functional studies to investigate this *in vitro* evolutionary processes in greater detail. We solved the tertiary structure of the core ATP binding domain of the folding optimized protein by nuclear magnetic resonance spectroscopy. Analysis of the folded structure revealed a fourth strand in the  $\beta$ -sheet that was not present in the crystal structure of a less stable and highly divergent variant. The rigidity of the fourth  $\beta$ -strand was confirmed by  $T_1$ ,  $T_2$ , and heteronuclear nuclear Overhauser enhancement (NOE) measurements. To investigate the functional properties of the individual mutations, we constructed single-point revertants for each mutation present in the core-binding domain of the protein and evaluated each revertant for its affect on ATP binding affinity, protein solubility, and thermal stability. We then used the structure of the folding optimized protein to interpret the effect of each mutation on ATP affinity and protein solubility. Taken together, the data suggest that a poorly functional, artificial protein can be significantly optimized through the accumulation of single-point mutations that do not dramatically affect the fold of the protein or the mode of ligand binding. This suggests that it might be possible to evolve many biologically relevant proteins that are structurally unstable or poorly soluble to a state of increased folding stability.



**Figure 1.** Sequence alignment of ATP binding protein variants of family B. (a) Alignment of primordial, binding-optimized, and folding-optimized consensus sequences from rounds 8 and 18 of our previous selection<sup>2</sup> and round 6 of our denaturing selection,<sup>5</sup> respectively, and protein 18–19. (b) Amino acid sequences of individual single-point mutant proteins used to identify genetic mutations with structural and/or functional significance. In each case, the selected mutation was reverted back to its ancestral residue. Invariant residues are indicated by dashes and selected mutations are shown in bold. Apparent solution binding affinity for ATP ( $K_d$ ) was determined by displacement equilibrium filtration.<sup>14</sup>

## Results

### Effects of mutations on ATP-binding

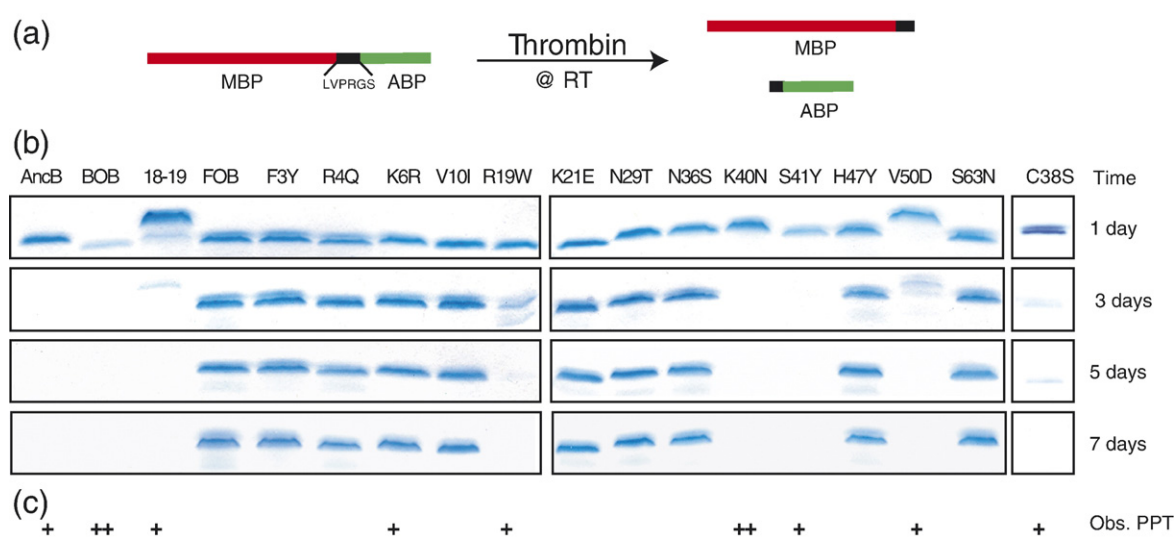
Our initial goal was to follow changes in ATP-affinity during the evolutionary optimization of the ancestral ATP-binding protein and to determine the effects of the individual mutations on ATP affinity. We therefore prepared the ancestral, binding optimized, folding optimized, and single-mutant revertant proteins as maltose binding protein (MBP) fusion constructs, since most of the proteins investigated were not soluble as autonomous folding units. Each ATP binding protein was constructed in the 62 amino acid sequence context defined by the core-binding domain of the folding optimized variant. Comparison of the dissociation constants for the ancestral, binding optimized, and folding optimized proteins (Figure 1(a)) revealed a dramatic shift in ATP binding. The  $K_d$  value of the ancestral protein for ATP was estimated to be between 0.5 mM and 1 mM, while the binding optimized protein had a  $K_d$  value of 100  $\mu$ M. Remarkably, the  $K_d$  of the folding optimized protein decreased dramatically to  $\sim$ 500 nM, suggesting that selection for greater stability of the folded state had indirectly led to an increase in ATP-affinity.

To evaluate the significance of each of the mutations that arose during the *in vitro* evolutionary optimization of the ATP-binding protein, we characterized 12 single point mutants of the folding optimized protein (Figure 1(b)). These mutations were designed to individually revert each of these 12 positions to their original residue observed in the ancestral protein. We also tested two additional

single-point mutations, R4Q and K6R. R4Q was chosen because the protein with the highest ATP affinity generated from the folding optimized selections differed from the consensus sequence only by this change.<sup>5</sup> The K6R mutation was chosen for analysis because of the unusual change from K6 in the ancestral protein to R6 in the binding optimized protein, then back to K6 in the folding optimized protein. Comparison of the dissociation constants for each mutant protein with the folding optimized consensus sequence (Figure 1(b)) enabled us to determine which amino acid substitutions had the greatest effect on ATP binding affinity. Of the 12 mutations, the revertants R19W and K21E resulted in a ten- and fivefold decrease in ATP affinity, respectively, relative to the folding optimized consensus sequence. Five other reversion mutations (V10I, N29T, S38C, S41Y and V50D) and K6R showed much smaller drops (twofold or less) in ATP binding affinity. Although the significance of these latter mutations is difficult to assess individually, if all of the effects are real and independent, their combined influence could be as much as ninefold.

### Effects of mutations of protein solubility

We wished to understand the contribution of each of the mutations that arose during the evolution of the optimized ATP-binding protein to the solubility of that protein. The solubilities of the ATP binding proteins were tested by proteolytically separating the fusion constructs into MBP and the free ATP binding protein (Figure 2(a)), incubating with or without ATP for up to seven days, pelleting aggregated protein and examining the remaining



**Figure 2.** Thrombin cleavage assay for protein solubility. (a) MBP fusion proteins were cleaved with thrombin to separate the individual ATP binding proteins from the maltose binding protein. (b) MBP fusion proteins were assayed for solubility at 2 mg/ml concentration over the course of seven days. The presence or absence of free ATP binding protein was determined by SDS-PAGE analysis. (c) The presence or absence of insoluble precipitate in the reaction mixture was noted with the following abbreviations: – (no precipitate), + (some precipitate), and ++ (significant precipitate). In most cases the presence of observable precipitate correlated with the absence of soluble protein in the gel.

soluble ATP-binding protein by analyzing band intensities on SDS–gels (Figure 2(a)). Comparison of the ancestral, binding optimized, and folding optimized proteins (Figure 2(b)) revealed that only the folding optimized protein remained soluble after seven days at room temperature, while the ancestral and binding optimized proteins formed visible precipitates during the first 24 h and become completely insoluble after three days. Surprisingly, the core domain of the divergent protein 18–19 also formed a visible precipitate after 24 h and was almost completely insoluble by three days. However, in the presence of 10 mM ATP (Supplementary Data), protein 18–19 (but not the ancestral and binding optimized proteins) remained soluble for greater than seven days. Only the folding optimized protein remained stably folded and soluble in the absence of ATP. Thus, differences in the amino acid sequence between the folding optimized protein and its ancestral forms, as well as the 18–19 variant, must contribute to the observed stability of the protein.

Analysis of the single amino acid revertant proteins (Figure 2(b)) indicated that five (R19W, S38C, K40N, S41Y, and V50D) formed visible precipitates within three days in the absence of ATP. The least soluble single reversion mutant protein was S41Y. This is the only amino acid substitution to result from two genetic mutations in the DNA sequence (UAC to AGC). Thus, optimization of the ancestral protein required a rare double mutation to change this amino acid, supporting the importance of a Ser residue at this position. The change from the ancestral W19 to R19 was noted above as the mutation with the strongest effect on ATP binding. Two of the remaining changes in the ancestral sequence (Y41, D50) to the optimized sequence (S41, V50) also affect ATP binding, but to a lesser extent. It is also interesting to note that although the K6R version of the folding optimized protein was largely soluble after seven days, a slight precipitate was observed (Figure 2(c)). This modest change in solubility may be why the folding optimized protein reverted to K6 rather than retain the R6 seen in the binding optimized protein and in the 18–19 variant.

### Effects of mutations on thermal stability

In addition to ligand binding affinity and protein solubility, we also investigated the effects of evolutionary optimization on the thermal stability of the folding optimized protein by monitoring changes in circular dichroism (210 nm) as a function of temperature. Although each of the proteins examined gave sigmoidal curves consistent with a single transition between the native and denatured states, none of the melts were reversible. Nevertheless, we were able to determine the relative thermal stability (Table 1) of each protein by comparing the melting transition ( $T_{50}$ ) for protein 18–19, FOB, and four single-point reversions found to remain soluble after removal of the MBP protein affinity tag. We were not able to obtain thermal stability data for the F3Y,

**Table 1.** Thermal stability of the folding optimized protein

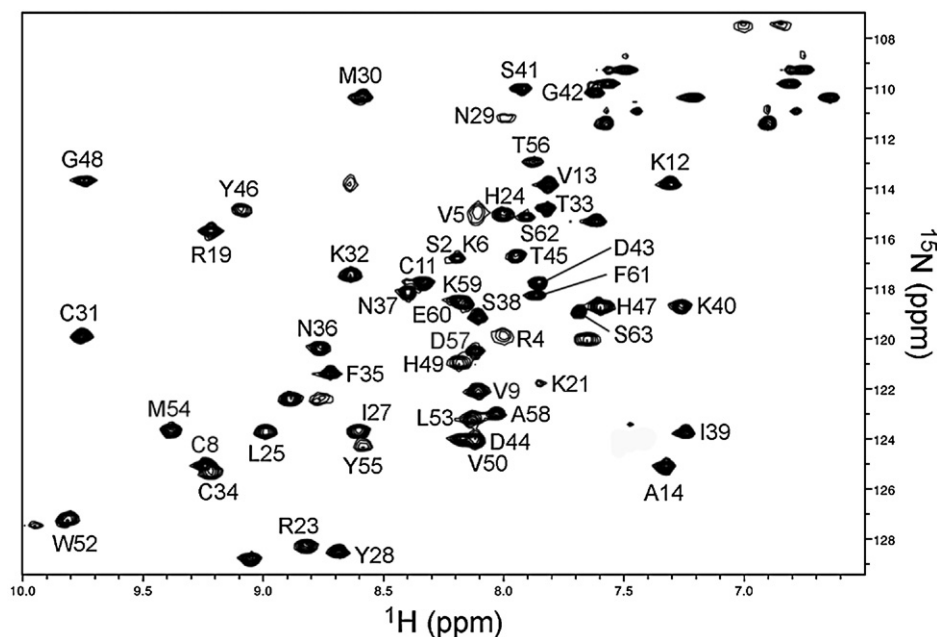
Entry	Sequence	Length	$T_{50}$ (°C)
1	18–19	81	56
2	FOB	81	76
3	FOB	64	74
4	FOB (K21E)	64	73
5	FOB (N29T)	64	78
6	FOB (N36S)	64	58
7	FOB (H47Y)	64	63

V10I, and S63N reversions, which remained soluble in the time dependent precipitation assay, due to limited expression and decreased solubility at elevated temperatures. Since protein 18–19 was not soluble as a 64 amino acid construct, we measured the thermal stability of a longer 81 amino acid construct that was similar to the previously crystallized sequence. Protein 18–19 was found to be much less thermally stable, with a  $T_{50}$  20 °C lower than the folding optimized protein. The additional amino-terminal residues that keep protein 18–19 soluble apparently do not contribute greatly to the thermal stability of the folding optimized protein, since both 64 and 81 amino acid constructs yielded nearly identical  $T_{50}$  values (76 °C *versus* 74 °C, respectively). Of the four individual revertant mutations examined, N36S has the greatest effect on thermal stability, as this reversion lowered the  $T_{50}$  from 74 °C to 58 °C. The H47Y reversion also significantly reduced the  $T_{50}$  to 63 °C. No appreciable change was observed for the K21E and N29T reversions ( $T_{50}$  = 73 °C and 78 °C, respectively).

### Structure and dynamics of the folding optimized ATP-binding protein

The crystal structure of protein 18–19 provided the first look at the structure of the ATP binding protein.<sup>3</sup> However, the folding optimized protein differs from protein 18–19 at 12 out of 62 positions in the core binding domain. In order to ensure that the observed biochemical data were not the result of gross structural changes, we used NMR spectroscopic data to calculate the tertiary structure of the folding optimized protein. Resonances were assigned for 58 of the residues of the folding optimized protein (Figure 3). The unassigned amino acid resonances were Gly1, Phe3, Pro7, Pro15, Val20, and Asn22. In addition, the assignments for residues Ser2, Arg4, Val5, and Lys6 are tentative, since few connectivities were observed. Excluding the proline residues, the unassigned residues correspond to loops and turns connecting regions of secondary structure. The inability to identify resonances within these regions and the lack of connectivities *via* through bond magnetization transfer for the remaining resonances of the amino terminus are consistent with conformational flexibility, which is typical of the termini and loop regions of proteins. The chemical shifts and calculated dihedral angles of the assigned backbone resonances were consistent





**Figure 3.**  $^1\text{H}$ - $^{15}\text{N}$  HSQC of the folding optimized protein at 600 MHz. Resonances are labeled with their corresponding residue number.

with the secondary structure seen in the crystal structure of protein 18–19 with the exception of the loss of an amino-terminal  $\alpha$ -helix and the presence of an additional  $\beta$ -strand at the carboxy terminus. The N-terminal half of the amino-terminal  $\alpha$ -helix present in the crystal structure of protein 18–19 derived from a FLAG-tag used for protein purification, whereas our investigation of the folding optimized protein used only the core domain of the protein. We were able to avoid extra sequences from protein tags because of the increased solubility of the core domain of the folding optimized protein, which allowed the structure and function of the protein to be evaluated without the influences of additional amino acid residues. The secondary structural elements were further confirmed by NOE assignments with the observance of characteristic intra-helical  $d_{\text{NN}}(i, i+2)$ ,  $d_{\alpha\text{N}}(i, i+2)$ , and  $d_{\alpha\text{N}}(i, i+3)$  NOEs and long-range inter-strand NOEs for the  $\beta$ -sheet. Taken together, the secondary structure of the folding optimized protein consisted of one  $\alpha$ -helix (K32–G42) and a four-stranded antiparallel  $\beta$ -sheet composed of  $\beta$ 1 (W18–K21),  $\beta$ 2 (H24–Y28),  $\beta$ 3 (H49–L53) and  $\beta$ 4 (D57–F61).

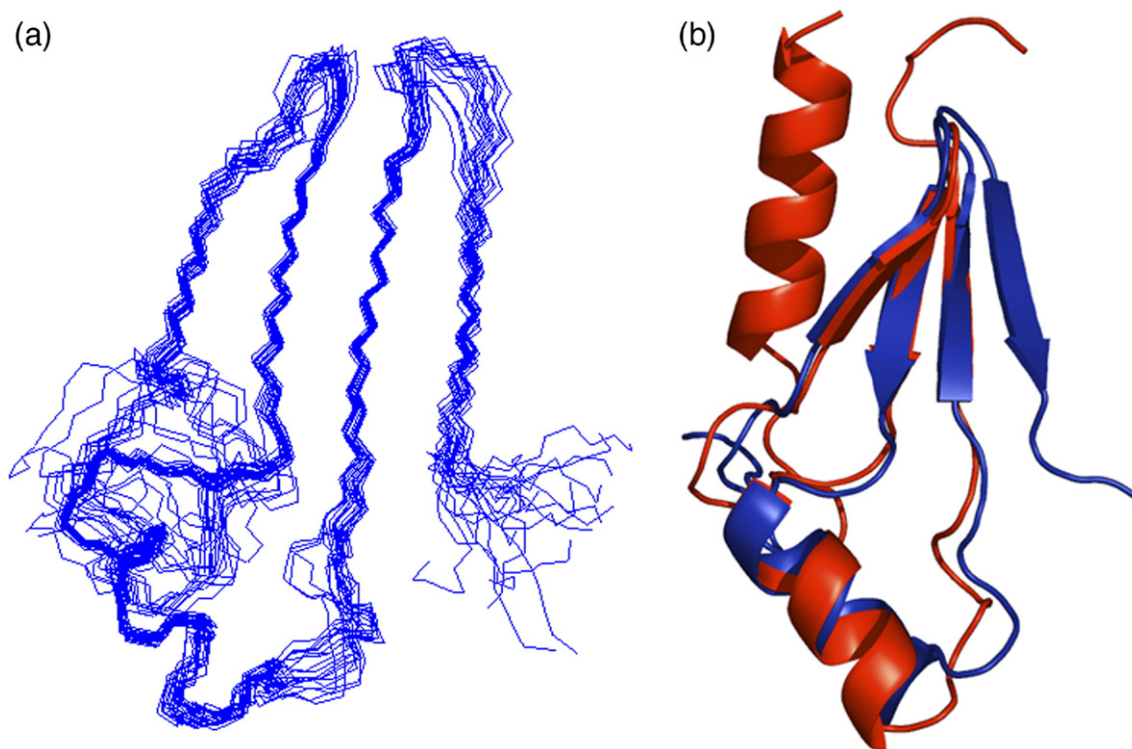
The RDC data also confirmed the similarity in tertiary structure between the folding optimized protein and protein 18–19 between residues 8 and 53 with a correlation coefficient of  $\sim 0.97$ . Long-range constraints were derived from 3D  $^{13}\text{C}$ - and  $^{15}\text{N}$ -NOE spectroscopy (NOESY) heteronuclear single quantum coherence (HSQC) spectra and used to calculate the structure of the folding optimized protein. The final structure had no NOE violations greater than  $0.5 \text{ \AA}$  and no dihedral angle violations greater than  $5^\circ$ . The structure was evaluated with MolProbity<sup>6</sup> and found to have 98% of its residues within the Ramachandran favored region for the

portions of the folding optimized protein that were structured. The value was 85% when considering the entire molecule. The backbone pairwise rmsd value for the family of structures over regions of secondary structure was  $0.6 \text{ \AA}$  and from residues 8 to 61 it was  $1.3 \text{ \AA}$ . The backbone rmsd between the folding optimized protein and protein 18–19 between residues 8 and 53 was  $1.7 \text{ \AA}$  (Figure 4).

Despite being an abiotic protein, the folding optimized protein shows many characteristics of a natural protein fold, including regular secondary structure and a compact globular tertiary fold (hydrodynamic radius =  $1.8 \text{ nm}$ ). The average  $T_1$ ,  $T_2$ , and heteronuclear NOE values of  $0.51 \text{ s}$ ,  $0.11 \text{ s}$ , and  $0.86$ , respectively, are consistent with a rigid, monomeric, and globular fold for a protein of  $7.5 \text{ kDa}$  and further confirmed the rigidity of the carboxy-terminal region of the folding optimized protein (Figure 5).

## Discussion

The *in vitro* evolutionary optimization of a low affinity ancestral ATP binding protein into a well folded, high affinity protein occurred through the accumulation of 12 amino acid substitutions. We have investigated the affect of each of these substitutions on ATP affinity and folding stability by examining single-point revertants at each site in the core-binding domain of the native protein. Two of the mutations (F3Y and S63N) have no measurable effect on either ATP affinity or protein stability, and as far as we can tell are completely neutral. Of the remaining ten mutations, two affect ATP binding strongly and five have possible weaker affects. In addition, seven mutations affect folding,



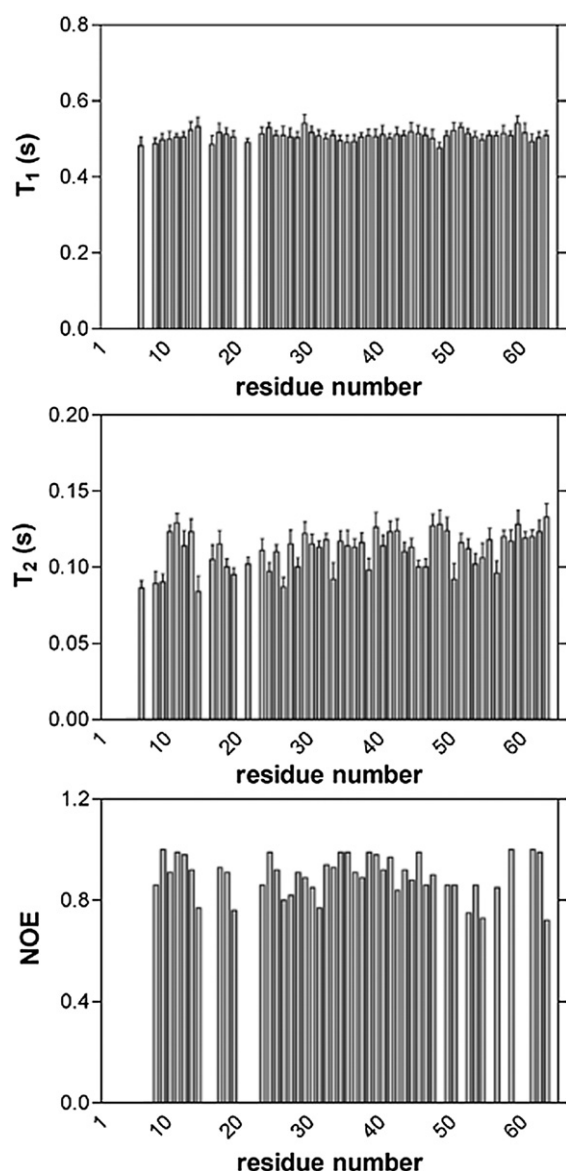
**Figure 4.** NMR solution structure of the folding optimized protein. (a) Family of structures of the folding optimized protein. (b) Ribbon diagram of the NMR structure of the folding optimized protein (blue) superimposed against the previously determined X-ray crystal structure of protein 18–19 (red).

as evidenced by solubility and thermal stability assays, and four of these affect both binding and folding. We will discuss these mutations in the context of the structure of the folding optimized protein.

Despite being a synthetic, abiotic protein, this *de novo* evolved protein contains all of the hallmarks of a naturally occurring globular structure. The folding optimized protein adopts an  $\alpha/\beta$ -fold characterized by a four-stranded antiparallel  $\beta$ -sheet and one  $\alpha$ -helix. The burying of a hydrophobic surface area upon binding of the adenine nucleobase and the tetrahedral coordination of four invariant Cys residues presumably drives the tertiary structure to a folded state. The first three strands of the  $\beta$ -sheet are composed of residues with high  $\beta$ -sheet propensity (Val, Ile, Tyr, and Trp)<sup>7</sup> and have a standard arrangement of alternating hydrophobic (W18, V20, L25, I27, V50, and W52) and hydrophilic (R19, K21, H24, R26, Y28, H49, and D51) residues.<sup>8</sup> Several residues within the hydrophilic face of the  $\beta$ -sheet (R26, Y28, H49, and K21) mediate polar interactions with the ATP ligand. The CXXCX<sub>n</sub>CXXC zinc-binding site, although a common motif in many proteins,<sup>9</sup> and remarkably similar to the “treble-clef” zinc-binding structure, is unique in the sense that no other nucleotide binding protein domains are zinc nucleated structures.<sup>4</sup> Comparison of the three-dimensional structure of the folding optimized ATP-binding protein with other nucleotide binding proteins, such as those containing the well-characterized Rossmann fold, reveals that while the ATP-

binding protein adopts a completely different protein fold, the process of *in vitro* evolution has ultimately led to the rediscovery of some of the same ligand binding contacts as those previously discovered by nature.<sup>10</sup> For instance, the adenine nucleobase (Figure 6) is precisely positioned between two aromatic side-chains in a readily observable pi-stacking interaction and the  $\alpha$  and  $\beta$ -phosphate groups of ATP coordinate one or more polar side-chains. Such interactions appear in most known nucleotide binding proteins.<sup>11</sup> One reason for this might be that there are only a limited number of ways in which certain types of chemical function (such as ligand-binding affinity) can be achieved, but a large number of ways in which proteins can fold to achieve this function.

The most dramatic difference between the solution structure of the folding optimized protein and the crystal structure of protein 18–19 is the presence of a fourth  $\beta$ -strand in the folding optimized protein. However, whether this represents a true structural difference between these two proteins is presently unclear. Support for altered structures stems from the fact that there are sequence differences between these two proteins in this short five-residue  $\beta$ -strand. For example, the construct used for our solution structural analysis had Ala and Phe at positions 58 and 61, respectively, whereas the crystallographically characterized protein contained Ser58 and Ile61. Furthermore, one of the evolved positions within  $\beta$ 3, V50 for the folding optimized protein and D50 for ancestral, binding



**Figure 5.**  $T_1$ ,  $T_2$ , and heteronuclear NOEs obtained for the folding optimized protein.

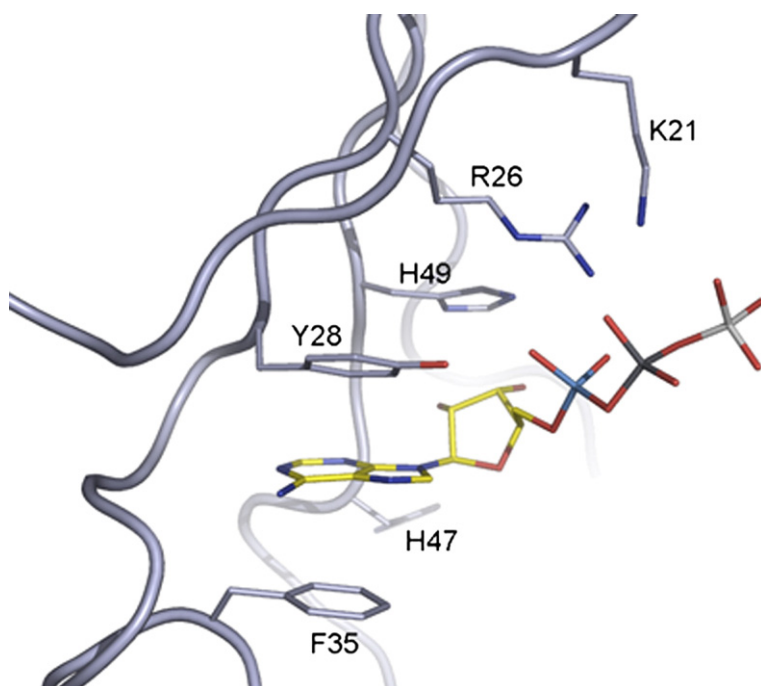
optimized, and 18–19 proteins, may contribute to the stabilization of the  $\beta$ -sheet (see below), and thus the stabilization of  $\beta_4$ . Nevertheless, the ancestral and binding optimized consensus sequences are identical to that of the folding optimized protein within  $\beta_4$ , and so loss of rigidity of the carboxy-terminus may be a peculiarity of protein 18–19. Another possibility is that the crystal state of the protein does not reflect its folding properties in solution, as has been occasionally observed.<sup>12</sup> Further mutagenic and spectroscopic studies are necessary to definitively characterize the structural differences between the different ATP-binding proteins.

One of the most remarkable features of the evolutionary optimization of the ATP-binding protein is that the basic protein structure, and the core protein–ATP interactions, were all present in the

ancestral sequence, i.e. the original sequence present in the random library. All of the mutations that affect ATP affinity affect peripheral aspects of ATP binding. The two revertant mutations (i.e. changes from the optimized to the ancestral state) with the strongest deleterious effects on ATP binding, R19W and K21E, have clear structural interpretations in terms of interactions with the triphosphate of the ATP ligand. The change from the ancestral W19 to R19 was first observed in the binding optimized protein sequence, and replaces a hydrophobic residue with a positively charged side-chain that interacts with the phosphates of the nucleotide. This position is K19 in protein 18–19, and the  $\epsilon$ -amino group of this lysine H-bonds with the  $\alpha$  and  $\beta$ -phosphate groups of ATP in the crystal structure.<sup>3</sup> The original E21 of the ancestral and binding optimized proteins evolved to K21 in the folding optimized protein and protein 18–19. K21 appears to hydrogen bond with the  $\gamma$  phosphate of ATP, since structural calculations placed the  $\epsilon$ N of Lys21 within 3.4 Å of the  $\gamma$  phosphate of ATP (Figure 6). More generally, the region of the protein surface defined by positions 19 and 21 has evolved an electropositive patch that is complementary to the high negative charge of the triphosphate of bound ATP (Figure 7(a)). Mutations at five additional sites have weaker effects on ATP binding (V10I, N29T, S38C, S41Y and V50D). The residues at these sites do not interact directly with the ATP ligand, either in the crystal structure or in our solution structure of the folding optimized protein. These mutations may lead to indirect effects on the positioning of neighboring residues, or may be due to altered folding stability in the case of S38C, S41Y and V50D. It is interesting to note that if all of these mutations act independently on ATP binding, the combined effect on the  $K_d$  would be roughly 500-fold, consistent with the observed change in  $K_d$  from the ancestral protein ( $>250$   $\mu$ M) to the folding optimized protein (500 nM).

The five single-site revertants (R19W, S38C, K40N, S41Y, V50D, from the optimized to the ancestral state) that decreased protein solubility are readily explainable from the structure of the folding optimized protein. Three of these changes (R19W, K40N, and S41Y) significantly decrease the hydrophilicity (Figure 7(b)) of the protein surface. Position 19, however, is unique in that it contributes both to improved protein solubility and ATP affinity. In the absence of ATP this residue is likely to be solvent exposed, and so R19, as found in the folding and binding optimized proteins, is more favorable than the W19 in the ancestral sequence. Following ATP binding, R19 hydrogen bonds to the  $\alpha$  and  $\beta$ -phosphate groups of ATP, as discussed above. The K40N results in loss of a surface charge, and S41Y increases the exposed hydrophobic surface area. S38C replaces a cysteine thiol on the protein surface, where it could potentially lead to dimerization and subsequent unfolding and further aggregation. The influence of position 50 (V50D) is more complex. In the crystal structure of protein 18–19, there is a salt





**Figure 6.** The adenine-binding pocket present in the NMR structure of the folding optimized protein. Residues that define the ATP-binding site are labeled. The side-chain amino-group of K21 is within hydrogen-bonding distance of the  $\gamma$ -phosphate of ATP, an interaction not present in the ancestral and binding optimized proteins.

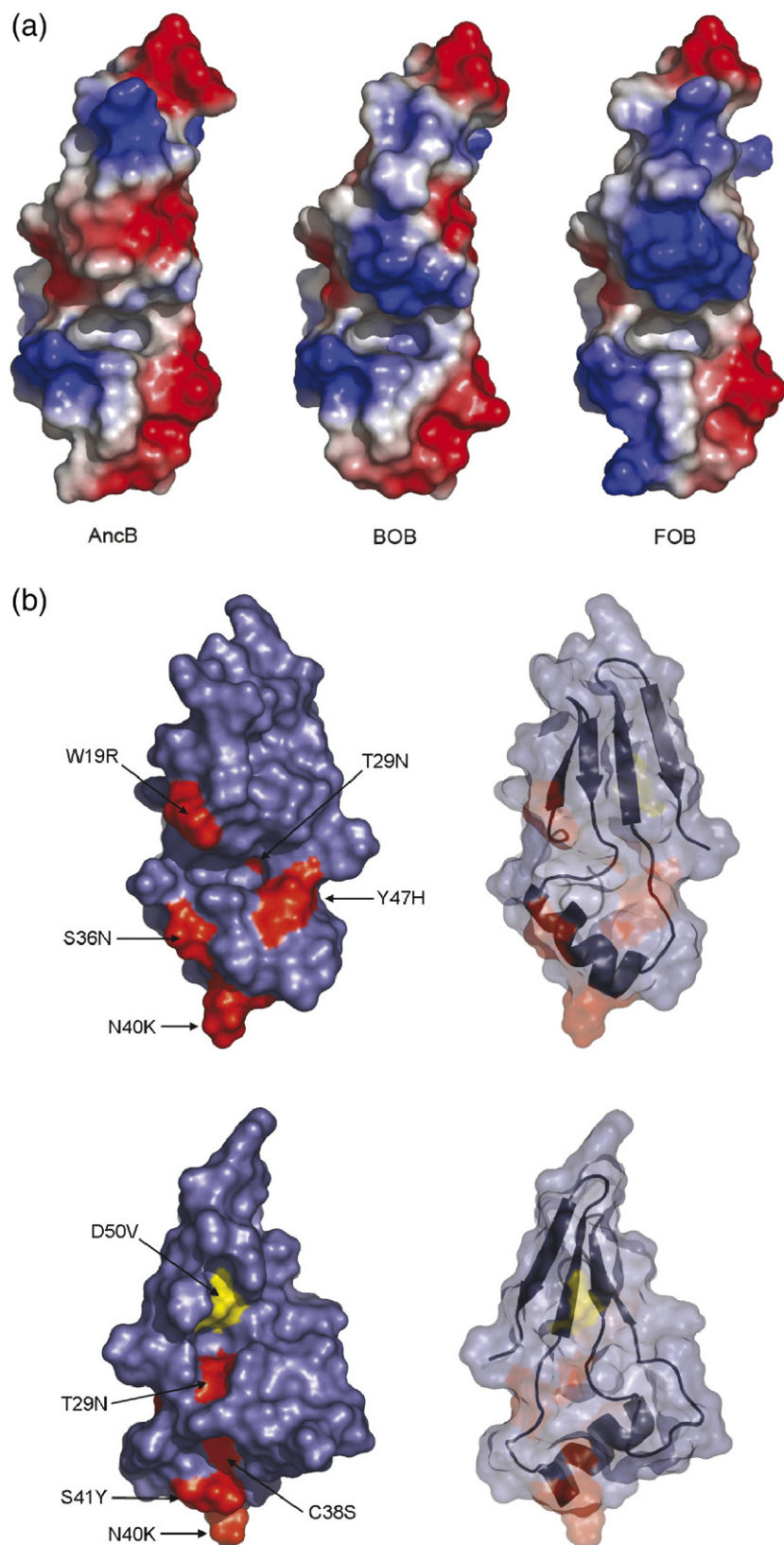
bridge between residues D50 and R6. However, because R6 is located in the amino-terminal  $\alpha$ -helix, this interaction is lost in the folding optimized protein. Since position 50 is on the hydrophobic face of the  $\beta$ -sheet it may be more energetically favorable to have a hydrophobic residue at this position, rather than a charged polar residue. Indeed, for the folding optimized protein, the proximity of Val50 in  $\beta$ 3 and Ile27 of  $\beta$ 2 is seen *via* NOEs between the  $\gamma$  and  $\delta$  protons of the Ile side-chain and the backbone amide of Val50. This would account for the presence of V50 in all of the clones examined in the output of the folding optimized selection. However, whether Val50 stabilization is sufficient to allow for the formation of  $\beta$ 4 requires further experimental evaluation. The possible influence of Val50 on  $\beta$ 4 can be observed by NOEs between this residue and Lys59 of  $\beta$ 4. For example, NOEs between the  $\alpha$  and  $\gamma$  protons of Val50 and the  $\alpha$  and  $\delta$  protons of Lys59 were detected. Interestingly, different mutations are found at three of these five positions in protein 18–19 (K19, D40, I41), suggesting that there are different ways of optimizing the functionality of the ancestral protein.

The thermal stability data show that several residues have important effects not detected by the binding and solubility assays. The N36S reversion, for example, results in a significant drop in thermal stability (58 °C *versus* 76 °C), probably due to the loss of the Asn36 hydrogen bond to the backbone carbonyl of K32. The interaction of N36 and K32 is supported by a number of NOE interactions observed between these two residues, including NOEs involving the side-chain amide of Asn36. The H47Y reversion, which also causes a significant drop in thermal stability (63 °C *versus* 76 °C), appears to de-stabilize the evolutionarily optimized protein by loss of a hydrogen bond between the N- $\epsilon$ 2 position

of His47 and the Ser38 side-chain. One of the more interesting results to come from these experiments was the observation that the K21E reversion had no measurable effect on protein folding stability. This was somewhat surprising given the dramatic effect that this mutation has on ATP binding affinity, and demonstrates that ligand binding and protein folding are not thermodynamically coupled. Taken together, the thermal stability data highlights the importance of hydrogen bonding and side-chain packing in protein folding stability.

All previous examples of the use of directed evolution to optimize protein function have started with biological proteins that have a long evolutionary history; typically the optimization is with respect to a new environment or a new function. In contrast, our work started with a sequence isolated from a large collection of synthetic, random protein sequences. When we began our efforts to optimize the initial, poorly functional sequence, it was not clear *a priori* that the sequence could be greatly optimized. For example, most random sequences with modest degrees of function might quickly reach a local optimum that cannot be surpassed without major structural changes. The fact that our primordial ATP-binding protein could be so extensively optimized, both with respect to affinity and with respect to protein stability, demands explanation in terms of the distribution of function in sequence space. One hypothesis that is consistent with our observations and with the general history of protein mutagenesis studies is that peaks in the protein fitness landscape are steep, in the sense that they consist of relatively few highly fit sequences, and vastly more sequences that are less fit. A randomly chosen sequence with a low level of functionality is therefore most likely to come from the base of a peak in the fitness landscape, and should therefore be optimizable by directed





**Figure 7.** Surface view of the ancestral, binding optimized, and folding optimized ATP binding proteins. (a) A positive electrostatic patch evolved during the course of the selections that improved ATP affinity. Positive, negative, and neutral regions are shown in blue, red, and white, respectively. Structures represent residues 8–54. (b) Surface representation of the folding optimized protein showing mutations resulting in increased hydrophilicity (red) and increased hydrophobicity (yellow) based on Kyte–Doolittle hydrophobicity data.<sup>41</sup> Top and bottom images refer to front and back views, respectively. Structures represent residues 8–62.

evolution. If most peaks in the fitness landscape have the same or similar steepness, then sequences from near the base of a very high peak (i.e. very good optimal function) may greatly outnumber sequences of similar activity that are near the top of lower peaks.

If this hypothesis is correct, then it should be possible to significantly optimize most randomly selected functional sequences. Testing this hypothesis will clearly require the selection and optimization of many independent examples of novel, functional proteins.

## Significance

We have begun to elucidate a pathway by which a primordial protein sequence attains increasing degrees of functionality through the systematic accumulation of point mutations resulting from *in vitro* evolution. Through biochemical and structural investigation of the ATP binding protein we have been able to suggest explanations for how changes at these positions contribute to improved solubility and ATP affinity. Enhanced ATP affinity results from the emergence of a patch of positive charge density on the protein surface, while improved protein solubility appears to result from mutations that increase the hydrophilicity of the protein surface and thereby decrease protein aggregation. Protein stability may be further improved by an additional mutation that stabilizes the hydrophobic face of the  $\beta$ -sheet of the protein.

The solution structure of the first synthetic protein evolved from a library of random amino acid sequences more faithfully reflects the true structure of the ATP-binding protein, at least with respect to regions that independently evolved without the influence of flanking sequences. The observation of a previously undetected  $\beta$ -strand presents an opportunity to further probe the association between the evolution of ligand binding, protein stability, and protein solubility with the appearance of an additional region of secondary structure.

*In vitro* directed protein evolution provides a novel approach to understanding the changes necessary to evolve an ancestral protein from a weakly folded state to a well-folded three-dimensional structure. The structure of the *de novo* evolved ATP binding protein reveals the exploitation of the same strategies used by natural proteins to stabilize tertiary folds and to recognize ligands. The application of directed evolution provides an exciting opportunity to build new abiotic proteins with tailor-made functionality.

## Materials and Methods

### Protein expression and purification

Individual ATP binding protein (ABP) sequences were constructed by PCR (see Supplementary Data, Figure S1, Tables S1 and S2) and inserted into the pMal expression vector (pIADL14) between BamHI and HindIII restriction sites. The resulting C-terminal fusion proteins have a thrombin cleavage site (LVPRGS) separating the MBP from the ATP binding protein (ABP).<sup>13</sup> Plasmid DNA was transformed into One-shot Chemically Competent *E. coli* TOP10 cells (Invitrogen) and sequenced to verify each mutation (ASU Core Facility). Plasmid DNA containing the correct mutation was transformed into *E. coli* BL21 (DE3) strain (Invitrogen). Bacteria were grown in LB medium supplemented with 0.1 mM ZnSO<sub>4</sub> at 37 °C to an A<sub>600</sub> of 0.8 and induced with 1 mM IPTG. Following an additional 4 h of growth, bacteria were harvested by centrifugation and re-suspended in 10 ml of phosphate buffer (25 mM NaH<sub>2</sub>PO<sub>4</sub> (pH 8.0), 250 mM NaCl). Cells

were lysed by sonication in the presence of 1 mg/ml lysozyme. The crude lysate was clarified by filtration through a 0.45  $\mu$  PES filter (Millipore). MBP-fusion proteins were bound to an amylose column (New England Biolabs), washed with phosphate buffer and eluted with phosphate buffer supplemented with 10 mM maltose. MBP-fusion proteins were concentrated using an Amicon ultrafiltration device (Millipore). Protein concentration was determined by UV-spectroscopy at 280 nm using molar extinction coefficients calculated for each sequence. The yield of pure MBP-ABP fusion protein obtained from a 250 ml culture of induced *E. coli* was ~10–15 mg.

### Dissociation constants

Equilibrium dissociation constants ( $K_d$ ) for purified MBP-ABP fusion proteins were measured by equilibrium ultrafiltration.<sup>14</sup> Apparent  $K_d$  values were measured using trace [ $\gamma$ -<sup>32</sup>P]ATP (Amersham Biosciences) and a series of concentrations of MBP-ABP fusion protein spanning the  $K_d$  app. The data were iteratively fit (through non-linear regression) using the computer program Deltagraph 4.0 (Red Rock Software) to the  $y = b + c((x/(x + K_d \text{ app}))$ ), where  $y$  is the experimentally measured value (top counts-bottom counts/top counts),  $x$  is the protein concentration,  $b$  represents non-specific binding to the protein and/or filter, and  $c$  is the maximum fraction of counts that can be bound.

### Solubility assay

MBP-ABP fusion proteins (2 mg/ml) were incubated overnight at room temperature with thrombin (1 mg/ml; Sigma) to cleave the MBP-ABP fusion into separate proteins. The relative amount of free ABP that remained in solution was monitored over time (one, three, five and seven days) by centrifuging the reaction mixture and analyzing a small aliquot (5  $\mu$ l) of the supernatant on a 10%–20% gradient SDS-PAGE (Tris-Glycine buffer system, BioRad). Comparing the band intensity of each mutant protein to the folding optimized protein provided a direct assessment of relative solubility. The presence or absence of insoluble precipitate was noted with the following abbreviations: - (no precipitate), +equation (some precipitate), and ++ (significant precipitate).

### Circular dichroism

Circular dichroism (CD) spectra were acquired using an Aviv CD Spectrometer model 202. Spectra were recorded in phosphate buffer without exogenous ATP by monitoring the wavelength dependence of  $[\theta]$  in 1 nm increments with a sampling time of 10 s. Melting curves were generated by monitoring the change in wavelength dependence of  $[\theta]$  at 210 nm over a temperature gradient of 24 °C –90 °C.  $T_{50}$  values were determined by curve fitting using the program Deltagraph 4.0 (Red Rock Software).

### <sup>15</sup>N and <sup>13</sup>C-isotope labeled protein

Samples used for NMR were purified from cells grown in M9 minimal media supplemented with <sup>15</sup>NH<sub>4</sub>Cl and [<sup>13</sup>C]glucose (Cambridge Isotopes). The protein construct used for NMR measurements was a truncated version of the folding optimized protein analogous to clone B6 (ABP

**Table 2.** Structural restraints and statistics

<i>NMR restraints</i>	
Number of distance restraints	
Total inter-proton	677
Sequential ( <i>i</i> , <i>i</i> +1)	349
Medium-range from ( <i>i</i> , <i>i</i> +2) to ( <i>i</i> , <i>i</i> +5)	202
Long-range (> <i>i</i> , <i>i</i> +5)	126
Number of dihedral angles	
Total	60
Phi	30
Psi	30
Total number of constraints	737
Total number of constraints per residue	11.5
<i>Structural statistics</i>	
Favored region (%)	77
Additional allowed region (%)	21
Average pairwise rmsd	
Backbone (Å)	1.3
Heavy (Å)	1.9

B6-62), except that a carboxy-terminal non-Zn ligating cysteine (C58) was replaced with an alanine. NMR samples were extensively dialyzed against 10 mM sodium phosphate, 50 mM KCl, 0.5 mM ATP (pH 6.5) and concentrated with a centricon YM-3 (Amicon) to ca 0.3 mM. Samples used for residual dipolar coupling measurements were partially aligned with 10 mg/ml Pf1 phage (ASLA biotech).

### NMR spectroscopy and structure calculations

NMR spectra were either acquired on a Bruker 600 MHz equipped with a cryo-probe at the MIT Francis-Bitter Magnetic Laboratory or on a Bruker 700 MHz equipped with a cryo-probe at Bruker BioSpin AG, Fällanden, Switzerland. Spectra were either processed with NMRPipe<sup>15</sup> or Bruker TopSpin 1.3 and analyzed with CARA<sup>16</sup> and NMRDraw. Backbone resonances were assigned *via* HNCA<sup>17</sup> and HNCACB<sup>18</sup> at 600 MHz and HNCO<sup>19</sup> at 700 MHz. Side-chain assignments were made by a (H)CCH-total correlated spectroscopy (TOCSY)<sup>20</sup> (with a mixing time of 22 ms) and <sup>1</sup>H-<sup>15</sup>N-TOCSY-HSQC (with a mixing time of 60 ms) at 700 MHz. HNHA<sup>21</sup> and <sup>1</sup>H-<sup>15</sup>N-NOESY-HSQC<sup>22</sup> spectra at 700 MHz were used for torsion angle calculations. The HNHA spectrum was used to calculate backbone dihedral  $\phi$  angles from <sup>3</sup>J<sub>HNH $\alpha$</sub>  coupling constants determined from the appropriate Karplus relationship.<sup>23,24</sup> The  $\psi$  dihedral angles were calculated from NOE peak intensity ratios of  $d_{\alpha\text{N}}(i-1, i)$  and  $d_{\text{N}\alpha}(i, i)$  in a <sup>1</sup>H-<sup>15</sup>N-NOESY-HSQC spectrum.<sup>25</sup> Dihedral angles were additionally analyzed by correlations with chemical shifts *via* TALOS.<sup>26</sup> Residual dipolar couplings were calculated from 2D IPAP <sup>15</sup>N-<sup>1</sup>H HSQC<sup>27</sup> spectra by comparing  $J_{\text{NH}}$  couplings between isotropic and partially aligned ABP FOB samples. The RDC values were then compared with those expected from the crystal structure of protein 18–19 (PDB ID, 1UW1) using PALES.<sup>28</sup> NOEs were measured from 3D <sup>1</sup>H-<sup>13</sup>C-HSQC-NOESY<sup>22</sup> (with a mixing time of 120 ms) and 3D <sup>1</sup>H-<sup>15</sup>N-HSQC-NOESY<sup>22</sup> (with a mixing time of 120 ms) spectra collected at 700 MHz.

Spectra for dynamics analysis were carried-out at 700 MHz. A series of <sup>1</sup>H-<sup>15</sup>N HSQC spectra were collected with flip-back<sup>29</sup> and phase-sensitive<sup>30–32</sup> pulse sequences. <sup>15</sup>N longitudinal relaxation rates,  $T_1$ , were measured with delays of 5.0, 100, 200, 400, 600, 1000, and 1500 ms. <sup>15</sup>N

transverse relaxation rates,  $T_2$ , were measured with delays of 16.5, 49.5, 99 (2 $\times$ ), 165, and 398 ms using a  $\tau_m$  of 900  $\mu$ s. Heteronuclear <sup>1</sup>H-<sup>15</sup>N NOEs were measured by taking the ratio of peak volumes acquired with and without <sup>1</sup>H saturation.<sup>29</sup> Relaxation rates  $T_1$  and  $T_2$  were calculated by fitting cross-peak intensities (*I*) as a function of the delay time (*t*) to a single-exponential decay ( $I = I_0 e^{(t/T)}$ , where *T* is either  $T_1$  or  $T_2$ ).

Structures were calculated with a hybrid distance geometry - simulated annealing protocol<sup>33,34</sup> with XPLOR-NIH.<sup>35</sup> Parameter files for ATP and zinc were from HIC-Up.<sup>36</sup> Distance restraints for the ATP and Zn<sup>2+</sup> were taken from the crystal structure of protein 18–19 (PDB ID, 1UW1). A total of 677 distance restraints were used for the folding optimized protein, including 349 sequential (*i*, *i*+1), 202 medium-range (from (*i*, *i*+2) to (*i*, *i*+5)), and 126 long-range (greater than (*i*, *i*+5)) constraints (Table 2). Two-hundred structures were calculated to generate a family of 20 structures. The hydrodynamic radius of the folding optimized protein was calculated with HYDROPRO<sup>37</sup> and theoretical  $T_1$ ,  $T_2$ , and heteronuclear NOE values were calculated with HYDRONMR.<sup>38</sup> Models of the ancestral and binding optimized proteins were calculated with MODELLER 8v0.<sup>39</sup> Structures were visualized with PyMol† and MOLMOL.<sup>40</sup>

### Protein Data Bank accession code

The coordinates of the folding optimized protein have been deposited in the RCSB Protein Data Bank (PDB) with accession code 2P0X.

## Acknowledgements

This work was supported by grants from the NASA Astrobiology Institute (NNA04CC12A) (to J.W.S.) and from the BioDesign Institute at ASU (to J.C.C.). J.W.S. is an Investigator of the Howard Hughes Medical Institute. S.S.M. is a NIH post-doctoral fellow (F32 GM07450601). We thank Dr Cristina Del Bianco and Dr Karel Kubicek for providing helpful scripts.

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.jmb.2007.05.062](https://doi.org/10.1016/j.jmb.2007.05.062)

## References

1. Wilson, D. S., Keefe, A. D. & Szostak, J. W. (2001). The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl Acad. Sci. USA*, **98**, 3750–3755.
2. Keefe, A. D. & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Naturem*, **410**, 715–718.

† <http://www.pymol.org>



3. Lo Surdo, P., Walsh, M. A. & Sollazzo, M. (2004). A novel ADP- and zinc-binding fold from function-directed *in vitro* evolution. *Nature Struct. Mol. Biol.* **11**, 382–383.
4. Grishin, N. V. & Krishna, S. S. (2004). Structurally analogous proteins do exist! *Structure*, **12**, 1125–1127.
5. Chaput, J. C. & Szostak, J. W. (2004). Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability. *Chem. Biol.* **11**, 865–874.
6. Lovell, S. C., Davis, I. W., Arendall, W. B., III, de Bakker, P. I. W., Word, M., Prisant, M. G. *et al.* (2003). Structure validation by C-alpha geometry: phi, psi, and C-beta deviation. *Proteins: Struct. Funct. Genet.* **50**, 437–450.
7. Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry*, **33**, 5510–5517.
8. Hecht, M. H., Das, A., Go, A., Bradley, L. H. & Wei, Y. (2004). *De novo* proteins from designed combinatorial libraries. *Protein Sci.* **13**, 1711–1723.
9. Cowan, J. A. (1997). *Inorganic Biochemistry: An Introduction*, 2nd edit, Wiley-VCH, New York.
10. Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**, 241–256.
11. Cappello, V., Tramontano, A. & Koch, U. (2002). Classification of proteins based on the properties of the ligand-binding site: The case of adenine-binding proteins. *Proteins: Struct. Funct. Genet.* **47**, 106–115.
12. Garbuzynskiy, S. O., Melnik, B. S., Lobanov, M. Y., Finkelstein, A. V. & Galzitskaya, O. V. (2005). Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray and NMR-resolved protein structures. *Proteins: Struct. Funct. Genet.* **60**, 139–147.
13. McCafferty, D. G., Lessard, I. A. D. & Walsh, C. T. (1997). Mutational analysis of potential zinc-binding residues in the active site of the enterococcal D-Ala-D-Ala dipeptidase VanX. *Biochemistry*, **36**, 10498–10505.
14. Jenison, R. D., Gill, S. C., Pardi, A. & Polisky, B. (1994). High-resolution molecular discrimination by RNA. *Science*, **263**, 1425–1429.
15. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
16. Keller, R. L. J. (2004). *The Computer Aided Resonance Assignment Tutorial*. 1st edit. vol. 1, CANTINA Verlag, Goldau.
17. Yamazaki, T., Lee, W., Revington, M., Mattiello, D. L., Dahlquist, F. W., Arrowsmith, C. H. & Kay, L. E. (1994). An HNCA pulse scheme for the backbone assignment of  $^{15}\text{N}$ ,  $^{13}\text{C}$ ,  $^2\text{H}$ -labeled proteins: application to a 37-kDa Trp repressor-DNA complex. *J. Am. Chem. Soc.* **116**, 6464–6465.
18. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. (1990). Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89**, 496–514.
19. Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wuthrich, K. (1998). TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc. Natl Acad. Sci. USA*, **95**, 13585–13590.
20. Kay, L. E., Xu, G. Y., Singer, A. U., Muhandiram, D. R. & Forman-Kay, J. D. (1993). A gradient-enhanced HCCH-TOCSY experiment for recording side-chain  $^1\text{H}$  and  $^{13}\text{C}$  correlations in H<sub>2</sub>O samples of proteins. *J. Mag. Reson. ser. B*, **101**, 333–337.
21. Vuister, G. W. & Bax, A. (1993). Quantitative J correlation: a new approach for measuring homonuclear three-bond  $J(\text{HNHA})$  coupling constants in  $^{15}\text{N}$ -enriched proteins. *J. Am. Chem. Soc.* **115**, 7772–7777.
22. Wider, G., Neri, D., Otting, G. & Wuthrich, K. (1989). A heteronuclear three-dimensional NMR experiment for measurements of small heteronuclear coupling constants in biological macromolecules. *J. Magn. Reson.* **85**, 426–431.
23. Karplus, M. (1959). Contact electron-spin coupling of nuclear magnetic moments. *J. Phys. Chem.* **30**, 11–15.
24. Karplus, M. (1963). Vicinal proton coupling in nuclear magnetic resonance. *J. Am. Chem. Soc.* **85**, 2870–2871.
25. Gagne, R. R., Tsuda, S., Li, M. X., Chandra, M., Smillie, L. B. & Sykes, B. D. (1994). Quantification of the calcium-induced secondary structural changes in the regulatory domain of troponin-C. *Protein Sci.* **3**, 1961–1974.
26. Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
27. Ottiger, M., Delaglio, F. & Bax, A. (1998). Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J. Magn. Reson.* **131**, 373–378.
28. Zweckstetter, M. & Bax, A. (2000). Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J. Am. Chem. Soc.* **122**, 3791–3792.
29. Grzesiek, S. & Bax, A. (1993). The importance of not saturating H<sub>2</sub>O in protein NMR. Application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc.* **115**, 12593–12594.
30. Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. A. (1992). Pulse sequences for removal of the effects of cross correlation between dipolar and chemical shift anisotropy relaxation mechanisms on the measurement of heteronuclear T1 and T2 values in proteins. *J. Magn. Reson.* **97**, 359–375.
31. Kay, L. E., Torchia, D. A. & Bax, A. (1989). Backbone dynamics of proteins as studied by  $^{15}\text{N}$  inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry*, **28**, 8972–8979.
32. Marion, D. & Wuthrich, K. (1983). Application of phase sensitive two-dimensional correlated spectroscopy (COSY) for measurements of  $^1\text{H}$ - $^1\text{H}$  spin-spin coupling constants in proteins. *Biochem. Biophys. Commun.* **113**, 967–974.
33. Kuszewski, J., Nilges, M. & Brunger, A. T. (1992). Sampling and efficiency of metric matrix distance geometry: a novel “partial” metrization algorithm. *J. Biomol. NMR*, **2**, 33–56.
34. Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). Determination of the three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters*, **229**, 317–324.
35. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73.
36. Kleywegt, G. J. & Jones, T. A. (1998). Databases in protein crystallography. *Acta Crystallog. sect. D*, **54**, 1119–1131.

37. de la Torre, J. G., Huertas, M. L. & Carrasco, B. (2000). Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.* **78**, 719–730.
38. de la Torre, J. G., Huertas, M. L. & Carrasco, B. (2000). HYDRONMR: prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. *J. Magn. Reson. B*, **147**, 138–146.
39. Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
40. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
41. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

*Edited by F. Schmid*

(Received 15 December 2006; received in revised form 1 May 2007; accepted 22 May 2007)  
Available online 26 May 2007