JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Exploring Symmetry as an Avenue to the Computational Design of Large Protein Domains

Carie Fortenberry, Elizabeth Anne Bowman,[†] Will Proffitt,[‡] Brent Dorr,[§] Steven Combs, Joel Harp, Laura Mizoue, and Jens Meiler*

Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology, and Institute for Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States

S *Supporting Information*

ABSTRACT: It has been demonstrated previously that symmetric, homodimeric proteins are energetically favored, which explains their abundance in nature. It has been proposed that such symmetric homodimers underwent gene duplication and fusion to evolve into protein topologies that have a symmetric arrangement of secondary structure elements—"symmetric superfolds". Here, the ROSETTA protein design software was used to computationally engineer a perfectly symmetric variant of imidazole glycerol phosphate synthase and its corresponding symmetric homodimer. The new protein, termed FLR, adopts the symmetric $(\beta\alpha)_8$ TIM-barrel superfold. The protein is soluble and monomeric and exhibits two-fold symmetry not only in the arrangement of secondary structure elements but also in sequence and at atomic detail, as verified by crystallography. When cut in half, FLR dimerizes readily to form the symmetric homodimer. The successful computational design of FLR demonstrates progress in our understanding of the underlying principles of protein stability and presents an attractive strategy for the *in silico* construction of larger protein domains from smaller pieces.

Structural studies of globular proteins have demonstrated that, despite there being thousands of unique proteins within living organisms, almost all tertiary structures can be categorized into one of 10 fundamental protein folds.[1] Six of these fundamental "superfolds" exhibit symmetry at the level of the tertiary fold: a set secondary structure element is repeated at least twice in a defined sequential order and internally symmetric spatial arrangement.[2] It has been postulated that these symmetric superfolds have evolved via gene duplication and fusion events from homo-oligomeric proteins (Figure 1). Fusion of monomer units into a single domain removes the entropic cost of assembling the oligomer, increasing thermodynamic stability and kinetic foldability.[3] Diversification on the sequence level achieves more complex biological functions and removes evidence of symmetry at the level of the primary sequence.[4] However, the overall fold remains symmetric.

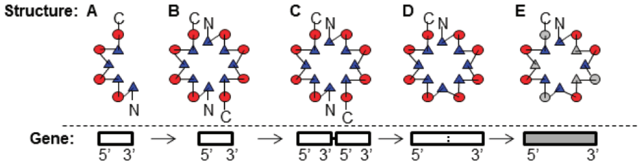

Figure 1. Self-attraction of a monomeric protein (A) yields a homodimeric complex with N- and C-termini close in space (B), and thereby a symmetric interface. If N- and C-termini are spatially proximal, gene duplication (C) and fusion (D) preserve the energetically favorable interaction across the interface. Diversification on the sequence level (E) allows for more complex function to be achieved (introduction of mutations is represented by gray shading). Circles represent α-helices; triangles represent β-strands.

Interestingly, the vast majority of homodimeric complexes in the Protein Data Bank (PDB) exhibit a symmetric arrangement of the two monomer units.[5] Andre et al. used explicit energy docking calculations with the ROSETTA protein design software to investigate the bias toward very-low-energy complexes in symmetric homodimeric complexes.[6] The study found a 2-fold greater variance in the interaction energy of random symmetric protein–protein docking arrangements, leading to an increased chance of observing highly attractive interactions. It is concluded that symmetric homodimers are selected for in evolution, explaining their abundance in nature.

This bias toward symmetry in homodimers would be preserved on the level of folds that arose through gene duplication and fusion. A boundary condition for this evolutionary strategy is that the C-terminus of one domain is spatially close to the N-terminus of the other domain in the tertiary structure of the homodimer so that, after gene duplication and fusion, the new protein domain can fold without disrupting the structure of the symmetric subunits. Figure S-1 in the Supporting Information (SI) demonstrates that 6.5% of 461 representative symmetric homodimers in the PDB have N- and C-termini closer than 20 Å. The 12 homodimers with the shortest distance between their N- and C-termini are also displayed in Figure S-1.

The $(\beta\alpha)_8$-barrel superfold is one of the most frequently observed folds in nature, comprising 10% of proteins with known structures.[7] The domain is composed of eight $\beta\alpha$ units, linked together by loops which wrap around to form a cylinder of parallel β-strands (β-barrel, Figure S-2) surrounded by a layer of parallel α-helices. The wide variety of amino acid sequences that adopt the fold makes it difficult to determine an evolutionary history. It is possible that the $(\beta\alpha)_8$-barrel fold was created several times independently and via different evolutionary routes.[8] However, one of the most popular hypotheses is that it arose through gene duplication and fusion of $(\beta\alpha)_{2n}$ units (Figure 1). Wierenga

Received: June 3, 2011

suggested that the $(\beta\alpha)_4$-half-barrel might be the smallest evolutionary unit because of its prominence in two-fold-symmetric $(\beta\alpha)_8$-barrel proteins.[9] However, structure-based multiple sequence alignments reveal a common GXD motif in the loops that precede even-numbered β-strands, suggesting evolution from $(\beta\alpha)_2$-quarter-barrel units. Soding et al.[10] detected a distinct two- and four-fold internal symmetry in members from several different SCOP superfamilies of the $(\beta\alpha)_8$-fold.

More evidence indicating the evolution of $(\beta\alpha)_8$-barrels from gene duplication and fusion comes from imidazole glycerol phosphate synthase (HisF, Figure S4-A). The HisF $(\beta\alpha)_4$-half-barrel structures have a sequence identity of only 16% but superimpose with root-mean-square distance deviations (rmsd) of 2.1 Å. In addition, the N- and C-terminal halves of HisF can be expressed separately and self-associate to form inactive homodimers.[11] When co-expressed *in vivo* or refolded *in vitro*, the two half-barrels combine to form an active heterodimer. In an attempt to reconstruct the evolutionary events that gave rise to HisF, Sterner et al. fused two copies of the gene encoding the C-terminal HisF $(\beta\alpha)_4$-half-barrel. Although the resulting protein, "CC", was poorly soluble and unfolded with low cooperativity,[12] an iterative process combining rational redesign followed by random mutagenesis and selection generated a stable protein, "C***C", with native-like properties.[13] However, although it gave impressive results, this strategy has disadvantages: (a) The resulting protein C***C is no longer perfectly symmetric on the sequence level, as rational redesign and random mutagenesis introduce different mutations in both subunits. (b) The approach assumes that the C-terminal half of the barrel was duplicated and all mutations accumulated in the N-terminal half during evolution, which is highly unlikely. (c) The process involves an iterative improvement of the designed protein through trial-and-error, offering limited insight into the fundamental forces that determine protein stability and limiting its application to other proteins.

Assembly of larger proteins from symmetric subunits not only presents an attractive strategy in evolution; it could also facilitate the computational design of large proteins, as the symmetry constraint reduces the sequence and conformational search space. Further, it enables a stepwise protocol that designs and characterizes stable subunits before optimizing interfaces between them for self-assembly. Both strategies will reduce the computational resources needed, enabling the design of larger proteins.

The present study reverse-engineers a perfectly two-fold-symmetric $(\beta\alpha)_8$-barrel based on a well-defined energy potential and with a reproducible *in silico* protocol (Figures 2, S-4, and S-5). It overcomes above-mentioned limitations of previous studies and explores the potential to exploit protein symmetry for the design of larger protein domains. The promising results obtained by Seitz et al. when fusing the C-terminal half of HisF to form CC[12] inspired this research to systematically test 62 symmetrized HisF variants *in silico*. We expected to identify energetic hotspots in the CC protein and determine a low-energy symmetric version of HisF. Note that this study was completed independently and before the experimental structure of the asymmetric C***C became available.[13] While the resulting protocol is based on the HisF structure as a template, the general strategy can be applied for *de novo* design of larger proteins. Specifically, the symmetry constraint reduces the sequence and conformational search space by a factor of 2, making the respective computer simulations feasible.

HisF was first superimposed on itself with a 180° rotation around the main β-barrel axis using a structure−structure alignment algorithm[14] (Figure 2A,C,E). As a result of the two-fold symmetry on the topology level, 62 sequence position pairs superimpose at 2.1 Å in the protein backbone. These 62 positions reside in parts of the structure that follow the two-fold symmetry most closely, i.e., α-helices and β-strands. At each of these 62 positions it is possible to cross over from one HisF copy to the other. For example, looking at the position pair (94::215)—starting at amino acid 94 of copy 1, follow the HisF backbone trace to amino acid 215 of copy 1. It superimposes with amino acid 94 of the 180° rotated copy 2 of HisF. Then, continue tracing on copy 2 until residue 215 is reached and it is possible to jump back to amino acid 94 of copy 1 (Figure 2D,F). As a result, 62 cyclic symmetric HisF variants were created, each duplicating a different half of the original protein. Note that, depending on the cut point, a different set of HisF loops is kept and duplicated, resulting in symmetric variants of different lengths. In essence, this protocol is a protein design experiment with a constraint on the sequence and conformational space.

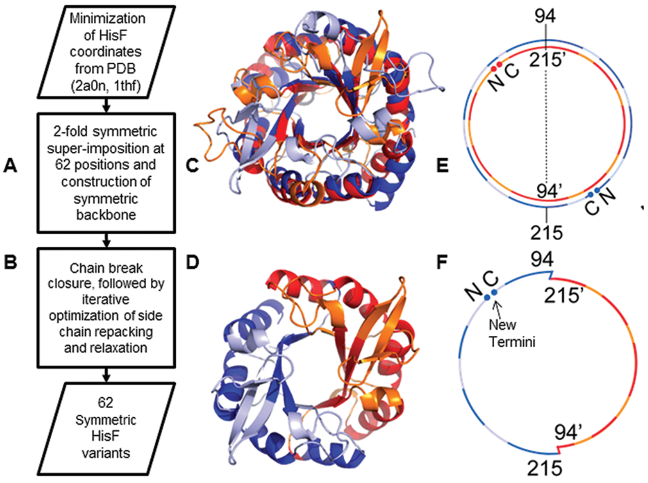
Minimization of HisF coordinates from PDB (2a0n, 1thf)
2-fold symmetric super-imposition at 62 positions and construction of symmetric backbone
Chain break closure, followed by iterative optimization of side chain repacking and relaxation
62 Symmetric HisF variants
A
B
C
D
E
F
94
215'
N C
94'
C N
215
94
N C
215'
New Termini
94'
215

Figure 2. Steps taken to computationally create the symmetric variants. Panel C is the superimposition of HisF, where the 62 cut sites are shown in dark blue on one copy and red on another. The noncut sites are shown in light blue and orange, respectively. Panel D is the symmetric variant created from duplicating cut sites 94−215 on each of the superimposed halves, which is termed FLR. Panels E and F are schematic representations of the same process, showing the cut sites of FLR at 94 and 215. Panel F shows the location of the new termini. A larger version of this figure can be found in the Supporting Information, Figure S-3.

Cyclic coordinate descent (CCD)[15] was used to rectify the slight geometry imperfections at the jump points. N- and C-termini are reintroduced into the cyclic proteins at positions equivalent to the termini positions in HisF between $\beta1$ and $\alpha8$. Iterative energy optimization, including backbone perturbation, side chain repacking, and gradient-based energy minimization,[16,17] was applied to the structure. For each of the 62 symmetric HisF variants, this energy minimization protocol was repeated 40 times in independent runs that started from either one of two experimental structures (1thf, 2a0n)[18] and one of 10 backbone conformations created in the CCD loop closure protocol. Repeating the protocol from slightly different backbone conformations ensures dense sampling of the local conformational space, providing an accurate determination of the minimum energy.

To prioritize symmetric HisF variants for experimental validation, the 62 variants were ranked by energy. Depending on the length of the loops, the variants had between 238 and 248 residues. To remove a bias toward larger proteins, the energy was normalized by the number of amino acids prior to ranking and is hence reported as ROSETTA energy units per amino acid (REU/AA). The top panel of Figure S-4 graphically depicts the lowest energy for each of the 62 backbones, ranging between −2.80 (68::189) and −3.16 REU/AA (94::215). Most stable variant 94::215 was termed FLR based on the amino acid sequence at the cut point. To obtain a baseline for comparison, the experimental structures of HisF (1thf, 2a0n) were minimized using an identical protocol and yielding −3.06 REU/AA.

All designs with energies better than −3.10 REU/AA were located in regions between sequence position pairs 93::214 (last turn of $\alpha3$ through $\beta4$) and 102::223 (last turn of $\alpha7$ through $\beta8$). Consistently low energies throughout these regions of secondary structure suggest that the half-barrel that contains $\beta4$-$\alpha4$-$\beta5$-$\alpha5$-$\beta6$-$\alpha6$-$\beta7$-$\alpha7$ of HisF yields the most stable two-fold-symmetric variants, largely independent of the precise position of the cut points. This region contains the elongated $\beta5$-$\alpha5$ loop, which consists of a three-stranded β-sheet. As this region is duplicated, the β-strand content of these symmetric variants increased from 24% in HisF to 30% in the symmetric HisF variants. The α-helical content remained constant at 35%.

Interestingly, the variant that is most similar to the fusion of the C-terminal half CC described by Sterner et al., (120::244), scored among the best (−3.06 REU/AA), giving an indication of why the experiments by the Sterner group were successful. The Sterner group further noted a salt-bridge cluster in HisF which contained R5 ($\beta1$), E46 ($\beta2$), K99 ($\beta4$), and E167 ($\beta6$). The cluster is irregular in the sense that not all four amino acids originate from β-strands with even numbers; i.e., they fail to form a single layer. The uncharged amino acid A220 in $\beta8$ cannot contribute to the salt-bridge cluster and is replaced with R5 ($\beta1$). This irregularity is responsible for the absence of the salt-bridge cluster in CC. Reintroduction of this salt-bridge cluster into the fusion of the C-terminal half of HisF greatly improved the proteins' stability experimentally[19,20] and also in our simulations from −3.06 to −3.10 REU/AA. The lowest energy symmetric HisF variants of the present study, including FLR, duplicate $\beta4$ instead of $\beta8$ when compared to CC. These proteins thereby contain the salt-bridge cluster at the base of the β-barrel consisting of E46 ($\beta2$), K99 ($\beta4$), E167 ($\beta6$), K220 ($\beta8$).

The active site of HisF is located at the C-terminal face of the barrel. The conserved and catalytically essential residues in HisF are located in positions D11 ($\beta1$) and D130 ($\beta5$). In FLR, D130 is duplicated and D130′ is placed in a position equivalent to D11. Further, HisF binds two phosphate groups of the substrate through residues G82, N103, T104 in site 1 and D176, G177, G203, A224, S225 in site 2. FLR duplicates N103, T104, D176, G177, G203, forming two intact phosphate binding sites.

A truncated variant consisting of amino acids 1::121 of FLR was constructed and termed halfFLR (see SI for sequence details). The ROSETTA energy of the monomer is substantially reduced when compared to FLR (−2.82REU/AA). A symmetric homodimer of halfFLR mimicking the structure of FLR is predicted to regain full stability (−3.16 REU/AA). The dimer interface is ∼1700 Å^2. Dimerization therefore stabilizes the protein by ∼11% in REU/AA and is predicted to occur spontaneously. This property of halfFLR further validates the hypothesis of the creation of symmetric superfolds from symmetric homodimers through the generation of a hypothetical, ancestral homodimer for HisF (Figure 1C).

In an additional step, the sequence of all 62 variants was optimized, enforcing a symmetry constraint to test if additional mutations can further stabilize the protein. While mutations were introduced in many of the 62 variants, FLR remained unaltered, indicating that its sequence is optimal. Even after optimization of the sequence of all 62 variants, FLR maintained the best overall energy and was therefore selected for experimental verification.

Details on the construction of the genes and expression for FLR and halfFLR are given in the SI. We observed a monodispersed particle size distribution with an average hydrodynamic radius of 50 ± 20 Å for FLR and 60 ± 20 Å for halfFLR, both within error of the expected values. Analytical size-exclusion chromatography (SEC) indicated a single symmetric peak at a volume corresponding to a 30 kDa species for both proteins (Figure S-6). Secondary structure element percentages were calculated based on far-UV circular dichroism spectra and confirmed the predicted constant α-helical and increased β-strand content relative to HisF (Figure S-7). The stability of halfFLR and FLR was assessed by guanidine-induced denaturation and indicated slightly decreased stability (2.6 and 2.8 M guanidine, respectively) compared HisF (3.5 M guanidine) but showed cooperative unfolding (Figure S-8). Differential scanning calorimetry indicates that the protein aggregates at high temperatures. Two-dimensional NMR (^{1}H−^{15}N HSQC) indicated compactly folded proteins with approximately half the number of peaks as HisF (140 vs 252 peaks, Figure S-9). The number can be slightly larger than precisely half of the 252 signals for HisF as the perfect two-fold symmetry is broken at the N-terminus (see SI).

Analytical ultracentrifugation (AUC) of the halfFLR species was performed to assess the percent dimerization of the protein. Sedimentation velocity AUC experiments indicate a single dimeric species. Similarly, SEC and dynamic light scattering experiments display a single dimeric species. No monomer or other oligomeric state can be observed under any conditions, preventing determination of the dissociation constant. Using a protein concentration of 160 μM in the AUC experiment and assuming the fraction of the monomeric version is <1%, we determine a conservative upper limit for K_d of 20 nmol, confirming the tight interaction predicted computationally.

The experimental structure of FLR was determined by X-ray crystallography to a resolution of 1.4 Å using the computational model for molecular replacement (PDB code 3DTN). The experimental structure shows that the protein is folded into the predicted $(\beta\alpha)_8$-barrel structure with 0.87 Å rmsd between the computational and experimental model backbones. Amino acid side chain conformations agree to 87% between model and experiment. Interestingly, 1.5 copies of FLR reside in each unit cell, which is diagnostic of the structural symmetry. The distances between equivalent positions agree to a rmsd of 0.29 Å The two halves superimpose to a rmsd of 0.34 Å for Cα positions, making FLR perfectly symmetric within the resolution of the experiment. The two halves superimpose for Cα positions to an rmsd of 0.339 Å. The FLR structure also indicates that the predicted salt-bridge cluster at the base of the β-barrel, consisting of residues E46 ($\beta2$), K99 ($\beta4$), E167 ($\beta6$), and K220 ($\beta8$), is intact (Figure 3A). The catalytic aspartate residues D9/130 and the phosphate binding sites N103/224, T104/225, D55/176, G56/177, G82/203 are largely unperturbed (Figure 3C).

The experimental structure of halfFLR was determined with a resolution 2.3 Å (PDB code 3DTM). The computational model of halfFLR was used for phasing and all comparisons. The experi-

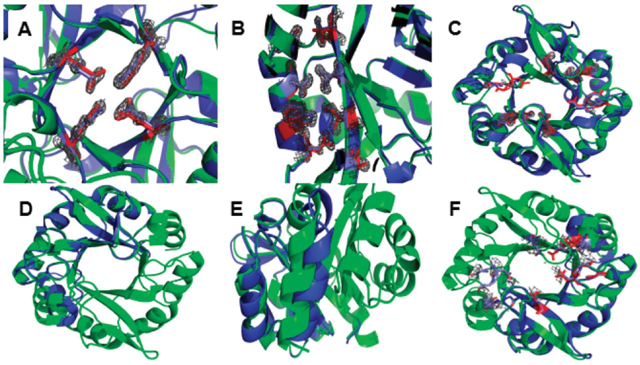

Figure 3. Computationally predicted models (blue) and experimental structures (green). (A) Density of the salt-bridge cluster (gray) of FLR superimposed with the computational side chains in red. These are residues E46, K99, E167, K220. (B) Density of the contacts between helix 1 and strand 1 superimposed with the computational side chains in red, revealing an excellent side-chain recovery. (C) Catalytically important residues, shown superimposed with the predicted model, appearing unperturbed. However, the missing density in the loops could explain the loss of activity in FLR. (D) Overall agreement between the model and the experimental structure of halfFLR. (E) Interface between the two halves of dimeric halfFLR, showing slight deviations from the computational model, likely due to the model being a monomeric half. (F) Catalytically important residues of halfFLR, showing the same flexibility as the FLR protein, again possibly explaining the lack of activity.

mental structure shows the protein is folded into the predicted $(\beta\alpha)_8$-barrel structure with 0.49 Å difference between the computational and experimental backbone coordinates. HalfFLR clearly shows that the two monomeric halves of the protein are assembled as a symmetric dimer (Figure 3D). Important structural features such as interface contacts and catalytically important residues are represented in Figure 3E,F, showing agreement with the predicted model. HalfFLR's phosphate binding sites are occupied with phosphate ions that were present in the crystallization buffer at a concentration of 12.5 mmol/L. Crystallography data collection and refinement statistics are listed in the SI.

Wild-type HisF converts N1-[(5′-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide to 5-aminoimidazole-4-carboxamide ribonucleotide and imidazole glycerol phosphate. Although the catalytic residues were largely unperturbed, FLR and halfFLR lack catalytic activity. Missing density indicates flexibility in the loop containing D176 and G177 which could explain the loss of activity (Figure 3C).

This study presents the first structure and sequence of a structurally symmetric $(\beta\alpha)_8$-barrel protein that is soluble and monomeric and folds cooperatively. Primary structure can be constrained to conform to the symmetry of the tertiary structure, and the protein still folds properly. The results of the present study are consistent with the gene duplication and fusion hypothesis of symmetric superfolds. Moreover, it creates two hypothetical ancestral variants of HisF: a sequence-symmetric variant of HisF and a related half-barrel protein that spontaneously dimerizes to a symmetric homodimeric $(\beta\alpha)$-barrel. Conserved structural traits such as salt-bridges and core packing are noted in these symmetric designs. The computational design protocol was highly accurate, as the X-ray structures agreed within 0.87 and 0.49 Å with the predicted models. To date, the largest *de novo*-designed protein consists of 106 amino acids. By taking advantage of the inherent symmetry of the $(\beta\alpha)_8$-barrel fold in the protein HisF, a protein of 242 amino acids was computationally designed, although arguably not *de novo*. However, the strategy to connect identical small proteins to larger architectures can be extended to the *de novo* design of larger domains.

ASSOCIATED CONTENT

S Supporting Information. Experimental procedures and crystallographic data. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

jens.meiler@vanderbilt.edu

Present Addresses

[†]Emory University, Atlanta, GA
[‡]Johns Hopkins University, Baltimore, MD
[§]Harvard University, Boston, MA

ACKNOWLEDGMENT

This work was supported by Defense Advanced Research Projects Agency, Protein Design Project and by the National Science Foundation grant 0619789.

REFERENCES

(1) Orengo, C. A.; Jones, D. T.; Thornton, J. M. *Nature* **1994**, *372*, 631–634.

(2) Brych, S. R.; Kim, J. W.; Logan, T. M.; Blaber, M. *Protein Sci.* **2003**, *12*, 2704–2718.

(3) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14249–14255.

(4) Pereira-Leal, J. B.; Levy, E. D.; Kamp, C.; Teichmann, S. A. *Genome Biol* **2007**, *8*, R51.

(5) Levy, E. D.; Boeri Erba, E.; Robinson, C. V.; Teichmann, S. A. *Nature* **2008**, *453*, 1262–1265.

(6) Andre, I.; Strauss, C. E.; Kaplan, D. B.; Bradley, P.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 16148–16152.

(7) Gerlt, J. A. *Nat. Struct. Biol.* **2000**, *7*, 171–173.

(8) Nagano, N.; Orengo, C. A.; Thornton, J. M. *J. Mol. Biol.* **2002**, *321*, 741–765.

(9) Wierenga, R. K. *FEBS Lett.* **2001**, *492*, 193–198.

(10) Soding, J.; Remmert, M.; Biegert, A. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W137–142.

(11) Hocker, B.; Beismann-Driemeyer, S.; Hettwer, S.; Lustig, A.; Sterner, R. *Nat. Struct. Biol.* **2001**, *8*, 32–36.

(12) Seitz, T.; Bocola, M.; Claren, J.; Sterner, R. *J. Mol. Biol.* **2007**, *372*, 114–129.

(13) Hocker, B.; Lochner, A.; Seitz, T.; Claren, J.; Sterner, R. *Biochemistry* **2009**, *48*, 1145–1147.

(14) Ortiz, A. R.; Strauss, C. E. M.; Olmea, O. *Protein Sci.* **2002**, *11*, 2606–2611.

(15) Canutescu, A. A.; Dunbrack, R. L. *Protein Sci.* **2003**, *12*, 963–972.

(16) Bradley, P.; Misura, K. M.; Baker, D. *Science* **2005**, *309*, 1868–1871.

(17) Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D. *Nature* **2007**, *450*, 259–264.

(18) Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M. *Science* **2000**, *289*, 1546–1550.

(19) Hocker, B.; Claren, J.; Sterner, R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16448–16453.

(20) Sterner, R.; Hocker, B. *Chem. Rev.* **2005**, *105*, 4038–4055.