THE PROTEIN SOCIETY

A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding

Liam M. Longo,[1] Connie A. Tenorio,[1] Ozan S. Kumru,[2] C. Russell Middaugh,[2] and Michael Blaber[1]*

[1]Department of Biomedical Sciences, Florida State University, Tallahassee, Florida 32306-4300
[2]Department of Pharmaceutical Chemistry, University of Kansas, Lawrence, Kansas 66047

Received 13 September 2014; Accepted 6 October 2014
DOI: 10.1002/pro.2580
Published online 9 October 2014 proteinscience.org

Abstract: The halophile environment has a number of compelling aspects with regard to the origin of structured polypeptides (i.e., proteogenesis) and, instead of a curious niche that living systems adapted into, the halophile environment is emerging as a candidate "cradle" for proteogenesis. In this viewpoint, a subsequent halophile-to-mesophile transition was a key step in early evolution. Several lines of evidence indicate that aromatic amino acids were a late addition to the codon table and not part of the original "prebiotic" set comprising the earliest polypeptides. We test the hypothesis that the availability of aromatic amino acids could facilitate a halophile-to-mesophile transition by hydrophobic core-packing enhancement. The effects of aromatic amino acid substitutions were evaluated in the core of a "primitive" designed protein enriched for the 10 prebiotic amino acids (A,D,E,G,I,L,P,S,T,V)—having an exclusively prebiotic core and requiring halophilic conditions for folding. The results indicate that a single aromatic amino acid substitution is capable of eliminating the requirement of halophile conditions for folding of a "primitive" polypeptide. Thus, the availability of aromatic amino acids could have facilitated a critical halophile-to-mesophile protein folding adaptation—identifying a selective advantage for the incorporation of aromatic amino acids into the codon table.

Keywords: protein evolution; proteogenesis; prebiotic; protein design; protein folding

Introduction

Abiogenesis (the origin of living systems) is hypothesized to have used the simple chemical building blocks that were freely available in the prebiotic environment (the Oparin–Haldane "heterotroph hypothesis"). A number of abiotic processes have been proposed to generate the critical organic compounds required for life to develop, including spark discharge chemistry,[1] hydrothermal vent chemistry,[2] high energy particle synthesis,[3] and deep space

Abbreviations: ADA, N-(2-acetamido) iminodiacetic acid; Aro, aromatic; BCA, bicinchoninic acid; CD, circular dichroism; DSC, differential scanning calorimetry; EPD, empirical phase diagram; ΔGunf, unfolding Gibbs energy; IPTG, Isopropyl β-D-1-thiogalactopyranoside; LUCA, last universal common ancestor; OD, optical density; NaPi, sodium phosphate; Ni-NTA, Nickel-nitrilotriacetic acid; OH, hydroxyl; PDB, protein databank; PEG, polyethylene glycol; PV1, designed primitive β-trefoil protein version 1; PV2, designed primitive β-trefoil protein version 2; RMSD, root mean square deviation; Sol, water solvent; T_m, melting temperature; Tween, polysorbate; U.V., ultraviolet.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: FSU College of Medicine (L.M.L); Grant sponsor: National Science Foundation Research [Experience for Undergraduates (NSF-REU); project (Award Number: 1156900); C.A.T.].

*Correspondence to: Michael Blaber; Department of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL 32306-4300. E-mail: michael.blaber@med.fsu.edu

chemistry with subsequent delivery to the Earth's surface by comets and meteorites.[4,5] Strikingly, these processes produce a consistent set of 10 of the 20 common α-amino acids (termed the "prebiotic set") comprised of A, D, E, G, I, L, P, S, T, and V.[6,7] Notably, this set has also been confirmed in recently reanalyzed original spark discharge samples of Miller.[8]

The prebiotic set of α-amino acids has several remarkable and compelling features as regards potential fitness for proteogenesis. For example, while the set resides at the theoretical minimum of complexity required for foldability[9,10] it contains amino acids having among the highest propensity values for formation of all three types of common protein 2° structure (i.e., α-helix, β-strand, reverse turn). With regard to hydrophobic/hydrophilic patterning essential for folding of soluble globular proteins,[11] the prebiotic set contains five hydrophobic and five hydrophilic amino acids. The prebiotic set is also U.V. transparent—indicating the potential for persistence and accumulation in a high-U.V. flux environment, as would be present prior to development of oxygen/ozone in the atmosphere,[12] (for a detailed discussion of such properties see Ref. 6). Given the above, and the ubiquity of proteins as the molecular workhorses in all extant life, it is likely that polypeptides were incorporated early in the abiogenic process; that is, proteogenesis (the origin of polypeptides) was a key event in the overall process of abiogenesis.

Consistent with the proteogenic hypothesis is recent experimental evidence that the prebiotic set of amino acids likely defines a "foldable set" (i.e., is supportive of protein folding) within a halophilic (high salt) environment.[13] The compatibility of prebiotic protein folding and the halophilic environment is due to the unique composition of the prebiotic amino acid alphabet, which is devoid of both basic and aromatic amino acids and is also—a distinctive hallmark of halophile proteomes.[7,14–16] High salt serves to stabilize protein structures having reduced hydrophobic packing volume, shields surface acidic charges, and promotes solubility through carboxylate binding of hydrated Na^+ cations. Salt-induced peptide formation (SIPF) also promotes favorable condensation reactions of peptide bonds in aqueous solution.[17] Thus, rather than being a curious niche that life adapted into, the halophile environment has been proposed as the likely site of origin of both proteogenesis and abiogenesis.[13,17,18]

The general consensus that aromatic amino acids (both canonical and noncanonical) were essentially absent when life first emerged is supported by several observations. The aromatic amino acids are the largest and most complex of the common α-amino acids[19] and prebiotic aromatic amino acid synthesis appears highly inefficient (with most abiotic chemical syntheses failing to yield aromatics altogether).[6,7] Furthermore, due to an essential lack of ozone in the atmosphere, abiotically generated aromatic compounds (i.e., aromatic amino acids and nucleic acid bases having absorption wavelengths falling within the U.V. range) would have been highly susceptible to photodegradation. As such, the concentrations of aromatic amino acids in unprotected surface environments on the early Earth are expected to have been marginal and accumulation unlikely.[12] Attempts at reconstructing the order of amino acid incorporation into the genetic code are in agreement: coevolution theory identifies the aromatics as being part of the "Phase 2" amino acids (that is, those amino acids incorporated well after establishment of the genetic code).[20] A detailed multifactorial analysis by Trifonov identifies the three aromatic amino acids (F, Y, and W) as being the last amino acids to be incorporated into the genetic code (along with aliphatic M).[21] Evolutionary analysis of W biosynthesis strongly suggests that biosynthetic pathways for this aromatic amino acid evolved only once and spread between species by horizontal gene transfer, sometime after the last universal common ancestor (LUCA).[22] An evolutionary analysis of F and Y biosynthesis was unable to establish whether the LUCA could synthesize these amino acids, although synthesis of chorismate (a key metabolic precursor to the aromatic amino acids) was possible.[23] Taken together, the above data identify aromatic amino acid biosynthesis as a key adaptation acquired sometime after the emergence of life, separate from the initial proteogenesis/abiogenesis event, and concurrent with, or preceding, the LUCA.

A reasonable assumption is that aromatic amino acids provided a selective advantage, initially as metabolites, and subsequently upon incorporation into polypeptides. As metabolites, aromatic amino acids could have provided protection from damaging U.V. radiation (as in the aqueous humor of avian eyes)[24] enabling nascent living systems to move from protective environments (i.e., physically shielded from U.V. radiation) to more open expanses. Subsequent incorporation of aromatic amino acids into polypeptides may have provided a selective advantage by their ability to stabilize proteins via improved core packing (through increased hydrophobic volume and combinatorial packing efficiency). Increased stability would enable broader functional mutations (which typically occur at the expense of stability)[25,26] or to adapt to novel (destabilizing) environments that would otherwise be inaccessible. To date, however, there has been no formalism or testable hypothesis of how aromatic amino acids might have affected protein evolution.

In this study, we test the hypothesis that the availability of aromatic amino acids could have played a major role in protein evolution by enabling a halophile-to-mesophile adaptive transition in

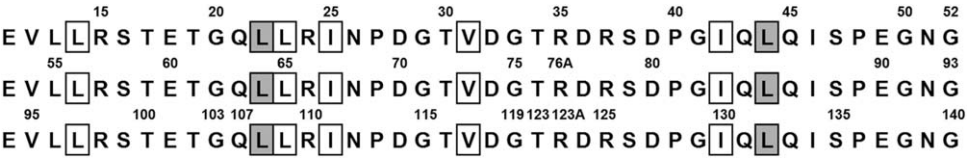

Figure 1. The 1° structure of the PV2 protein (using single letter amino acid code) and aligned to reflect the threefold symmetric architecture of the overall β-trefoil fold. The boxed positions comprise the hydrophobic core packing group. The shaded positions are the locations where aromatic amino acids were substituted as point mutations or in various combinations (see Table I).

protein folding. This hypothesis is evaluated by testing the effects of incorporating aromatic amino acids into a "primitive" designed protein that is highly enriched for the prebiotic amino acids, devoid of aromatics, and is an obligate halophile with regard to foldability.[13] Previous studies of the primitive protein model system showed that a mesophile/halophile folding transition occurred concomitant with six simultaneous substitutions of buried aromatic amino acids. The requirement of multiple simultaneous substitutions is a steep barrier to evolutionary change in comparison to a single substitution. In this study, we test whether incorporation of only a single aromatic amino acid can obviate the need for halophile conditions for the efficient folding of a "primitive" obligate halophile protein. The results show this to be the case, supporting the hypothesis that incorporation of a Shikimate-like pathway into the genome of early HaloArchea could relax the requirement of salt for protein foldability, thereby facilitating expansion into a low-salt mesophile environment, and demonstrating a plausible biophysical basis for the evolutionary selection of the "Phase 2" aromatic α-amino acids.

Results

Mutant sequence characteristics

The design of the "primitive" protein ("PV2") utilized in this study has previously been described.[13] PV2 is a small β-trefoil protein made up of three identical repeats of 42-amino acids, comprising an amino acid alphabet of only 12 letters. PV2 is highly enriched (~80%) for the prebiotic set of amino acids,[6] is devoid of aromatic amino acids, and has an acidic pI = 4.36. Importantly, the hydrophobic core of PV2 (21 of 126 residues total, or 17% of amino acid positions) is entirely prebiotic (i.e., comprised of only L, I, and V). PV2 was designed by Top-Down Symmetric Deconstruction[27] and, as a consequence, the identical sequences of the three 42-amino acid structural subdomains that form the β-trefoil architecture define a threefold rotational symmetry that substantially simplifies mutational design (Fig. 1).[28]

The aromatic amino acids F, Y, or W were incorporated at two buried locations within the PV2 scaffold known to exhibit a high statistical preference for aromatic amino acids in the β-trefoil fold[29]: symmetry-related positions 22, 64, and 108 (comprising three independent hydrophobic "mini-core" regions) and symmetry-related positions 44, 85, and 132 (participating in a cooperatively packing central hydrophobic core). L → V mutation at the mini-core and central core positions resulted in markedly reduced expression and solubility and were not pursued further. L→I mutation was tolerated in both the mini-core and central core, but was less stable than L in both cases (data not shown). Constructs denoted "6xAro" incorporate the indicated aromatic amino acid at all six positions (e.g., 6xF indicates a combined F mutation at positions 22, 64, 108, 44, 85, 132 in the PV2 protein). Other constructs are named according to the number of aromatics, the type of aromatic amino acid incorporated, and the positions of incorporation (e.g., 2xF(22,108) has an F residue incorporated at positions 22 and 108 in PV2).

Differential scanning calorimetry

Incorporation of aromatic residues F, Y, or W in the mini-core region of PV2 (as 3xAro(22,64,108)) is

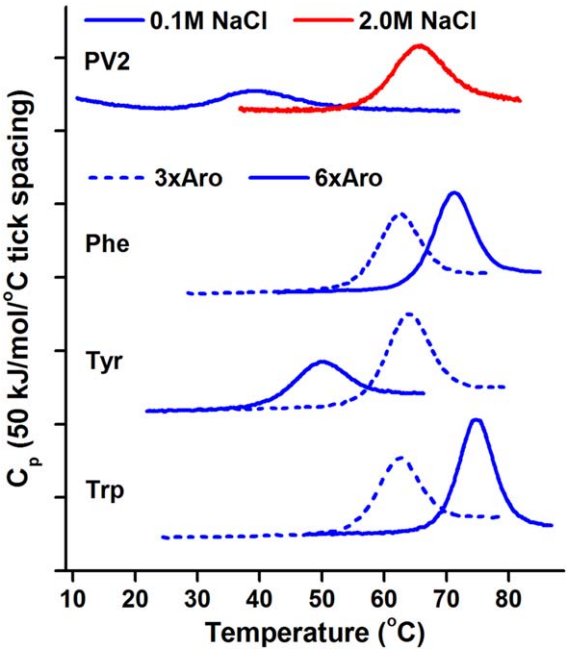

Figure 2. DSC of PV2 derivatives with aromatic amino acids at mini-core and central core positions. 3xAro constructs have the stated aromatic amino acid at positions 22, 64, 108. All data was collected under low salt conditions (0.1*M* NaCl) except where noted (red curve).

Table I. *DSC data for the thermal denaturation of PV2 and mutant proteins*

	$\Delta H(T_m)$ (kJ mol^{-1})	T_m (°C)	$\Delta H_{van't\ Hoff}/\Delta H_{cal}$	ΔT_m (°C)
Protein[a]				
PV2[b]	157 ± 5	34.2 ± 0.2	0.87 ± 0.19	–
PV2 (2.0*M* NaCl)[b]	357 ± 2	64.5 ± 0.1	0.84 ± 0.04	30.3
Mini-core mutants				
1xF(22)	306 ± 3	48.5 ± 0.1	0.71 ± 0.01	14.3
1xF(64)	302 ± 3	48.3 ± 0.2	0.74 ± 0.02	14.1
1xF(108)	300 ± 2	48.2 ± 0.1	0.75 ± 0.01	14.0
2xF(22,108)	401 ± 2	56.9 ± 0.1	0.71 ± 0.01	22.7
3xF(22,64,108)	446 ± 3	63.2 ± 0.1	0.81 ± 0.02	29.0
3xY(22,64,108)	437 ± 1	67.2 ± 0.1	1.12 ± 0.01	33.0
3xW(22,64,108)	414 ± 3	62.0 ± 0.1	1.03 ± 0.02	27.8
Central core and mini-core mutants				
6xF[b]	490 ± 3	70.7 ± 0.1	0.96 ± 0.08	36.5
6xY	301 ± 1	48.9 ± 0.1	1.04 ± 0.02	14.7
6xW	544 ± 1	74.6 ± 0.1	0.99 ± 0.01	40.4

[a] Buffer contains 0.1*M* NaCl unless otherwise noted.
[b] From Ref. 13

stabilizing in each case. Compared to PV2, the Aro3x(22,64,108) mutants (2.4% aromatic amino acid incorporation) display an increase in T_m (i.e., ΔT_m) ranging from +27.8 (W) to +33.0°C (Y). This increase is essentially equivalent to the increase in T_m exhibited by PV2 in response to a high salt environment (ΔT_m = +30.3°C in 2.0*M* vs. 0.1*M* NaCl) (Fig. 2, Table I). Likewise, all 6xAro constructs are more thermostable than PV2, with increases in T_m ranging from +14.7 (Y) to +40.4 (W) °C (Fig. 2, Table I). Comparisons between the 3xAro(22,64,108) and the 6xAro series indicate that while F or W incorporation into the central core is stabilizing, the Y mutation is destabilizing, and the melting temperature of 6xY is lowered by 18.3°C relative to 3xY(22,64,108).

To determine how many aromatics are necessary to achieve essentially complete fractional folding (i.e., ≥0.99) of PV2, 1xF, and 2xF constructs were evaluated. F was selected for further study because it is less complex and more resistant to photodegradation than either Y or W; additionally, F is considered the earliest aromatic amino acid acquisition in Trifonov's analysis (discussed above). Each of the three mini-core positions 22, 64, and 108 was mutated independently to probe for differential effects on stability. The melting temperatures and enthalpies of unfolding of 1xF(22), 1xF(64), and 1xF(108) are essentially indistinguishable, indicating that all three of the mini-core positions are structurally equivalent in the native and unfolded states. Likewise, a plot of the number of F residues in the mini-core versus ΔG_{unf} is linear (Supporting Information Fig. S1), as expected if the mini-core sites are noninteracting. Melting temperature, however, is nonlinear with respect to the number of incorporated F residues (Supporting Information Fig. S2) and it is the first F mutation that results in the greatest increase in T_m, with subsequent F mutations having diminished effects. At its temperature of maximum stability and in a low (i.e., mesophile) salt condition, PV2 is only 0.81 fractionally folded; in contrast, the 1xF mini-core variants achieve fractional folding of ≥0.99 at their respective temperatures of maximum stability in low salt (Fig. 3). A comparison of 3xY and 3xW mini-core mutant stability with 3xF shows that the Y mutation is more stable, while W is essentially isoenergetic with F. Thus, with incorporation of just a single aromatic amino acid (that is, at 0.8% of positions–an ~11-fold reduction in the typical percentage of aromatic residues found in extant, mesophile proteins[30,31]), involving either F, Y, or W, the requirement of high salt concentrations for essentially complete folding is eliminated.

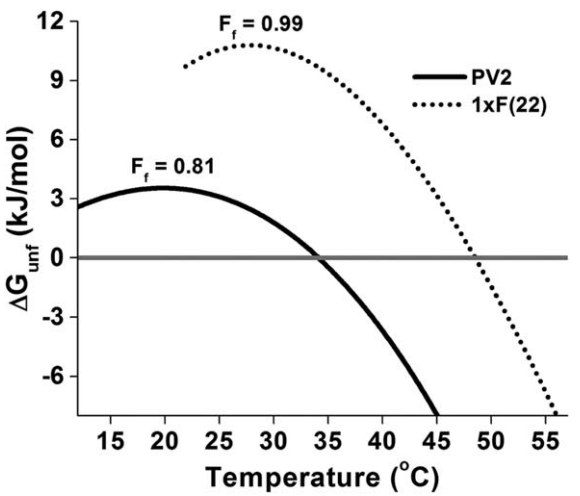

Figure 3. Stability of PV2 and 1xF(22) as a function of temperature. Stability curves were generated from fitted DSC parameters. F_f is the fraction folded at the temperature of maximum stability (i.e., ΔS = 0).

X-ray crystallography

Crystal structures for 6xY and 6xW were solved to a resolution of 1.70–1.75 Å (Table II); crystal

Table II. *Crystallographic data collection and refinement statistics*

	6xW[a]	6xY[b]
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell constants (Å)	$a = 47.1$	$a = 34.8$
	$b = 48.5$	$b = 46.8$
	$c = 69.7$	$c = 67.6$
	$\alpha = 90°$	$\alpha = 90°$
	$\beta = 90°$	$\beta = 90°$
	$\gamma = 90°$	$\gamma = 90°$
Max resolution (Å)	1.70	1.75
Highest shell (Å)	1.74–1.70	1.81–1.75
Mosaicity (°)	0.76	0.63
Redundancy	7.5	7.3
Mol/ASU	1	1
Matthews coef. ($Å^3$/Da)	2.68	1.87
Total reflections	135,200	85,201
Unique reflections	17,987	11,690
I/σ (overall)	58.0	54.3
I/σ (highest shell)	4.1	4.3
Completion overall (%)	98.8	98.3
Completion highest shell (%)	99.9	99.6
R_{merge} overall (%)	7.6	5.2
R_{merge} highest shell (%)	40.2	31.1
Nonhydrogen protein atoms	1005	1023
Solvent molecules/ion	136/2	116/1
R_{cryst} (%)	19.5	18.7
R_{free} (%)	23.2	21.6
RMSD bond length (Å)	0.007	0.007
RMSD bond angle (°)	1.04	1.10
Ramachandran plot:		
favored (%)	97.5	100.0
allowed (%)	2.5	0.0
outlier (%)	0.0	0.0
PDB accession	4QKS	4QKR

[a] 1.4*M* $(NH_4)_2SO_4$, 0.1*M* Tris pH 7.0, 0.07*M* Li_2SO_4.
[b] 30% PEG 8000, 0.1*M* Imidazole HCl pH 8.0, 0.2*M* NaCl.

structures of PV2 and 6xF have been previously reported.[13] Each mutant demonstrates the predicted β-trefoil architecture and, despite a difference of 30 buried carbons between PV2 and 6xW, there is no evidence of any significant global structural expansion or collapse. Indeed, the main chain RMSD values for the 6xAro constructs range from 0.48 (6xF) to 0.56 Å (6xY) in comparison to PV2.

Position 22 mutations ("mini-cores"). Residue positions L13 and I42, along with the aliphatic chains of R15 and R37, form a hydrophobic environment around residue position 22 (Fig. 4). This hydrophobic "mini-core" is a distinct packing environment from the central hydrophobic core-packing group, and is replicated by the threefold symmetry of the β-trefoil structure at equivalent positions 22, 64, and 108. The introduced F, W, and Y aromatic residues at position 22 are accommodated with remarkably minimal structural perturbation. Each aromatic residue adopts an identical χ1 angle as the parental L22 residue in PV2. In response to the presence of the bulkier aromatic rings at position 22, the adjacent Arg15 side chain adopts an alternative rotamer in each case to avoid a close contact (Fig. 4); all other neighbor residues are unchanged. The mutant Y hydroxyl extends into partial solvent accessibility, and its hydrogen bonding requirement is satisfied by two novel water molecules (Sol77 and Sol60, Fig. 4). Similarly, the mutant W Nε1 nitrogen of the pyrrole ring achieves partial solvent accessibility, and its hydrogen bonding requirement is satisfied by a novel water molecule (Sol33, Fig. 4).

Position 44 mutations (central core). Residue positions V12, L14, L23, and I25 form a hydrophobic environment around residue position 44 (Fig. 4). This region comprises part of the main central hydrophobic packing group, and is replicated by the threefold symmetry of the β-trefoil structure at equivalent positions 44, 85, and 132. The F, W, and Y aromatic residues introduced at position 44 are accommodated with minimal positional changes, or alternate rotamer conformations, of the adjacent residues. The introduced aromatic side chains, in each case, adopt the same χ1 angle as the parental L44 residue in PV2. The substitution of L44 by aromatic amino acids eliminates the L44 Cδ1 atom (the mutant aromatic rings are coplanar with the Cδ2 atom in each case). In response to the loss of the Leu44 Cδ1 atom the adjacent Leu14 adopts an alternative rotamer to effectively fill this space (Fig. 4). The bulkier F Cζ carbon and W Cε2 carbon introduce a close contact with L23, which is relieved by adoption of an alternate χ2 angle rotamer of L23. In the case of mutant Y44, the much longer OH group results in an alternative χ1 angle rotamer of L23 to avoid a close contact. The aromatic rings also result in a positional shift (∼1.0 Å) of adjacent I25 Cε1. In response to the bulkier indole ring of the introduced W44, the adjacent I25 residue adopts an alternative χ1 angle rotamer to avoid a steric clash. The hydrogen-bonding requirement of the mutant Y OH hydroxyl is provided by the main chain carbonyl of L23 with minimal (0.5 Å) positional shift (Fig. 4). Similarly, the hydrogen-bonding requirement of the mutant W Nε1 is also provided by the main chain carbonyl of Leu23 with minimal (0.3 Å) positional shift.

Refined coordinates and structure factors for the 6xW and 6xY mutants have been deposited in the Protein Databank (accession numbers 4QKS and 4QKR, respectively).

Empirical phase diagrams

Circular dichroism (CD), differential scanning calorimetry (DSC), and optical density at 360 nm were used to construct a temperature versus [NaCl] empirical phase diagram for PV2 and 1xF(64) (Fig. 5; raw data given in Supporting Information Fig. S3). Taken together, these probes provide a comprehensive view of the conformational state occupied by the protein, in which CD monitors secondary structure, DSC is sensitive to the heat capacity

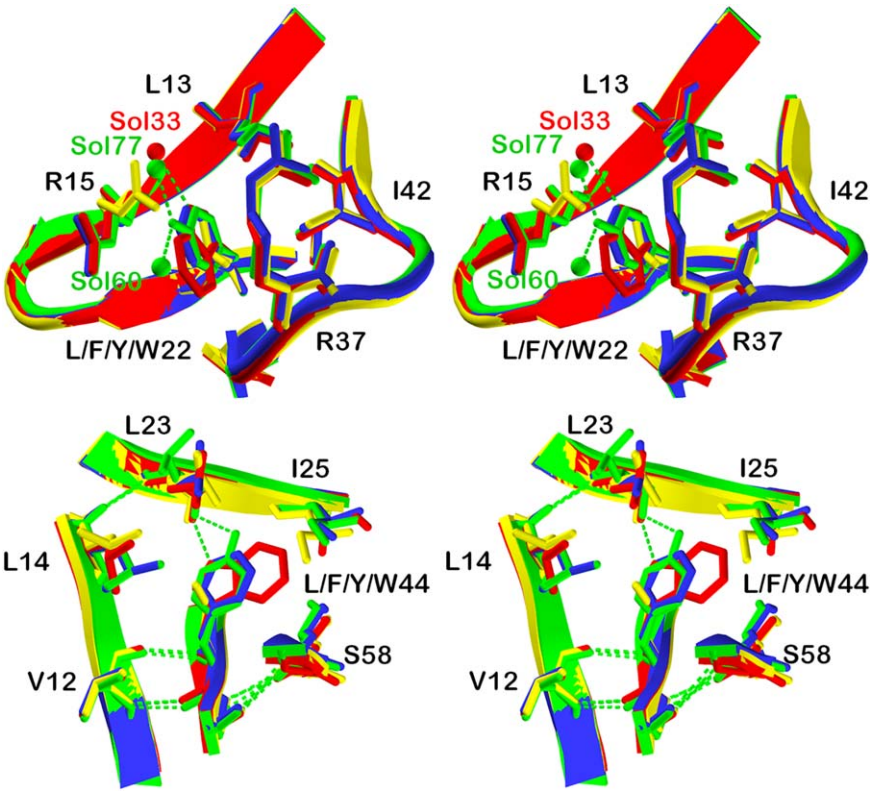

Figure 4. X-ray crystal structure overlays of aromatic substitutions in the PV2 protein. Upper panel: relaxed stereo diagram overlay of PV2 (yellow) with the 6xF (blue), 6xW (red), and 6xY (green) X-ray structures in the region of the position 22 mutations. Residue positions I42, L13 and the aliphatic side chains of R15 and R37 form a hydrophobic "mini-core" region around position 22. Lower panel: a similar overlay in the region of position 44 mutations. Residue positions L23, I25, V12, and L14 form a hydrophobic region around position 44—comprising part of the central hydrophobic core.

change associated with a conformational phase transition, and OD_{360} monitors protein aggregation. The OD_{360} data show that neither PV2 nor 1xF(64) aggregate, even at high temperatures and in the presence of 2.0*M* NaCl. Based on both DSC and CD, 1xF(64) is more thermostable than PV2, and differences in T_m as a function of salt concentration are greatest at low concentrations of NaCl: ΔT_m (0.1*M*) = +14.3°C and ΔT_m (2.0*M* NaCl) = +7.0°C (Fig. 5, panel c).

Discussion

Although abiogenesis is one of the great unsolved problems in biochemistry, practical hypotheses are notoriously difficult to formulate and test. Among the challenges is the evaluation of key physical or chemical processes that took place over geological time scales, as well as assumptions regarding uncertain conditions. Furthermore, it is highly improbable that a single experiment will arrive at a solution; as with other major scientific problems, elucidating abiogenesis will be achieved through a series of individual advances—identifying what is possible, plausible, or implausible for key aspects of the overall abiogenic process. The Miller–Urey gas discharge experiments, along with recent related studies, have identified a consensus set of 10 of the common α-amino acids (the "prebiotic set") that were plausibly available in the prebiotic soup as raw material for the very first peptides.[6,32,33] A testable hypothesis is whether this restricted "abiotic" set comprises a foldable set that is able to support complex, stably folded polypeptide architecture.[6] Basic and aromatic amino acids are notably absent from the prebiotic set, thus, salt bridges and aromatic core packing interactions are not feasible structural features to promote foldability in the earliest polypeptides. Successful studies of simplified protein design have been reported whereby foldable proteins have been constructed from a reduced α-amino acid alphabet,[34,35] and relevance for proteogenesis have been described. However, such studies have focused exclusively on achieving minimization of the alphabet size, without regard to the prebiotic relevance of the included amino acid alphabet. Thus, without exception, such minimal foldable proteins have depended on critical aromatic amino acids within the core, as well as stabilizing salt bridges (dependent on basic amino acids), to achieve a stable structure. Thus, more work is needed to elucidate the critical question of whether the prebiotic amino acids form a foldable set; however, there

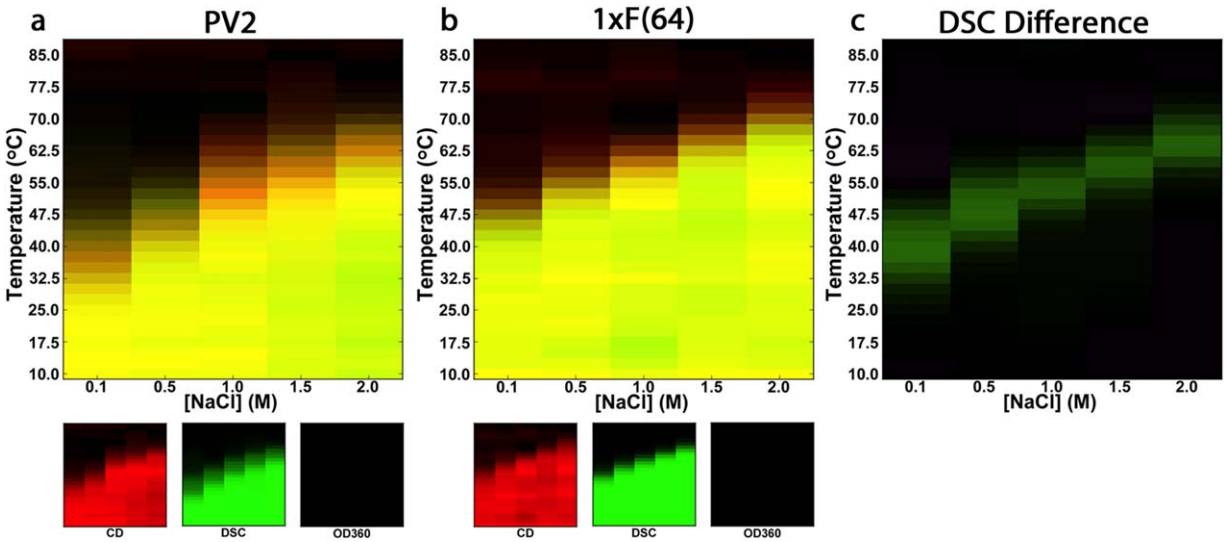

Figure 5. Temperature vs. salt concentration empirical phase diagrams of PV2 and 1xF(64). Empirical phase diagrams were generated using CD (a measure of secondary structure formation; red color indicates native-like structure), DSC (monitors heat associated with unfolding; intensity of green color corresponds integrated DSC signal, in which green color is assigned to the pretransition signal) and OD_{360} (turbidity; blue color). Phase diagrams are generated with an additive color scheme; thus, yellow color indicates that both CD and DSC report native-like interactions whereas black color indicates an unfolded protein that does not aggregate. A difference DSC EPD (c) was generated by subtracting the interpolated, integrated DSC signals of PV2 from that of 1xF(64); green indicates regions where 1xF(64) has a greater population of folded molecules than PV2.

is compelling evidence to support the prebiotic foldable set hypothesis, with studies indicating a dependency of such foldability on a halophile environment.[6,7,13]

Use of the β-trefoil architecture as a model of early folded proteins is motivated by several factors: first, the β-trefoil fold is comprised of β-strands and β-turns, which organize into β-hairpins and a β-barrel. Both the architecture itself and the structural motifs contained within it are common to every domain of life. Second, the structural evolution of the β-trefoil is well characterized, including an experimental demonstration of structural emergence by homo-oligomeric self-assembly (from a much simpler 42-mer peptide subdomain). Such data include a detailed experimentally validated path through stable, foldable sequence space linking an evolved β-trefoil protein (human fibroblast growth factor-1) to a simple 42 residue peptide "building block" (Monofoil-4P).[27,36] Furthermore, sequence simplification (a recognized feature of ancient proteins) was accomplished with the development of the PV2 protein, comprised of an alphabet of only 12 different amino acids types. Although a number of protein simplification studies have reported stable folded structure using reduced amino acid alphabets, such simplified proteins fail to achieve prebiotic relevance because they depend upon non-prebiotic amino acids for structure and stability—notably involving aromatic or basic amino acids. As such, the observation that such simplified proteins can fold within a mesophile environment does not contradict the present results. Given that the β-trefoil is a common architecture with a unique robustness to sequence simplification, we conclude that PV2—which is entirely devoid of aromatic amino acids, with a purely prebiotic protein core—represents one of the best model systems currently available for studies of the folding potential of the prebiotic set of amino acids.

The positions selected for evaluating the effects of introducing aromatic residues in PV2 (22, 44, 64, 85, 108, and 132) have the property of residing within buried hydrophobic environments and being positions statistically preferred by aromatics in consensus sequence analyses of the β-trefoil fold.[37–39]

The large hydrophobic aromatic amino acids have long been known as major contributors to efficiently packed protein cores, providing substantial stabilizing Gibbs energy.[40–42] There are two structural challenges to aromatic amino acid accommodation within a protein core: first, the adjacent residues must be able to adjust in response to the larger bulk of the aromatic amino acids, otherwise unfavorable strain ("overstuffing") among core residues will result.[43] The choice of the core and mini-core positions used in this study minimizes the potential for overstuffing as it is already known that these sites can accommodate an aromatic amino acid. Second, the protein must provide an appropriate hydrogen-bonding partner to the polar groups of Y (OH) and W (Nϵ1). In this regard, Y has both donor and acceptor requirements, while W requires only an acceptor. The X-ray structure analysis of the aromatic mutants shows that these positions in the β-trefoil have a

plasticity that facilitates ready accommodation of essentially any of the aromatic amino acids.

As expected, the added bulk of the aromatic amino acids are accommodated with minor adjacent side chain rotamer adjustments and no substantial main chain perturbations. At positions 22, 64, 108 the hydrogen-bonding requirements of Y and W are achieved by solvent—two solvent molecules (one apparent donor, one apparent acceptor) in the case of Y and one (an acceptor) in the case of W. At positions 44, 85, 132 the protein architecture itself provides an appropriate acceptor in the main chain carbonyl 23O (as a second donor interaction, 23O has an H-bond donor partner in an adjacent buried solvent). No donor is observed interacting with the Y OH; thus, while the hydrogen-bonding requirements of the introduced W may be fully satisfied, those of the introduced Y appear to be incomplete. Water/hydrophobic solvent transfer free energy values, as well as experimental values for A → S and V → T polar substitutions at hydrophobic (i.e., buried) positions in proteins, indicate an upper value of $\Delta G \sim +12$ kJ/mol for effective desolvation of such polar groups with no corresponding novel H-bond partner.[44,45] The derived $\Delta\Delta G$ value for an F → Y point mutation at symmetry-related positions 44, 85, and 132 is $\sim +10$ kJ/mol per mutation, in agreement with the expected destabilization of an unsatisfied H-bonding requirement. The stability data is consistent with the structural data: the added hydrophobic bulk of the buried aromatics provide substantial increased stability regardless of type of aromatic amino acid, with the exception of Y at positions 22, 44, 85. Thus, at the two buried environments evaluated, the protein achieves significant stability gains with a general introduction of aromatic amino acids (i.e., with 15 of 18 possible aromatic substitutions). The stability increase in response to aromatic substitution is not due to π-stacking or π-cation interactions as the prebiotic design is devoid of basic and aromatic amino acids within the core. The stability increase is due to a combination of hydrophobic effect (solvent entropy gain on aromatic burial) and more extensive van der Waals interactions within the core, combined with a structural ability of the basic β-trefoil architecture to satisfy H-bond requirements of the aromatics Y and W. Subsequently, the aromatic substitutions—potentially as point mutations—have the ability to move the folding properties of the PV2 protein from halophilic to mesophilic conditions. This ability of the β-trefoil architecture to accommodate and thermodynamically benefit from aromatic substitutions (primarily involving a main chain architecture able to provide necessary hydrogen-bonding interactions without perturbation) suggests a plausible selective advantage for this fold upon the evolutionary availability of aromatic amino acids.

Protein design studies suggest that a halophilic environment may have been involved in supporting protein folding early in abiogenesis (i.e., before the incorporation of amino acid biosynthesis).[6,13] Consistent with this view is the observation that peptide bond formation is promoted by high NaCl concentrations.[17,46] This demonstrates that polymerization (an otherwise thermodynamically unfavorable condensation reaction in water) of a key class of biopolymer is achievable under plausible prebiotic conditions. These studies suggest that the cradle of life may have resided within evaporative salt ponds, within which nonvolatile metabolites—in this case, amino acids—would have been concentrated and undergone chemical condensation to form polypeptides. Thus far, NaCl has been assumed to be the most appropriate salt of the halophile environment. Other salts are of interest to study, both as potential cosalts in halophile environments and as probes to understand the biophysical basis of enhanced stability in more detail (e.g., effects of the Hoffmeister series); such studies are currently in progress. Although it is known that copolymers (e.g., PEG, Dextran, and Ficoll) as well as various sugars can stabilize proteins, these additives (unlike simple salts) lack prebiotic relevance and their accumulation in the environment to concentrations that would significantly affect folding appears improbable.

If high salt conditions are a requirement for stable folding of the earliest polypeptides, then a key question is how life could have adapted out of such halophilic environments. Previously, it was shown that a construct with a combined total of six F residues can shift folding requirements from the halophile to mesophile environment. However, if six F substitutions are simultaneously required for a halophile-mesophile shift in folding, it would be evolutionarily implausible. We show here that incorporation of a single aromatic amino acid can effectively convert a foldable "prebiotic" polypeptide from an obligate halophile to a stable mesophile (i.e., with fractional folding of ≥ 0.99 in the absence of high concentrations of salt). Notably, the stability data demonstrate that potentially any of the aromatic amino acids (F, Y, or W) substituted into PV2 (involving the mini-core position) could enable this folding transition, and that the first aromatic amino acid yields the greatest increase in melting temperature. These results are consistent with the observation that aromatic amino acids are significantly more common in the proteomes of mesophiles than in halophiles,[7,14–16] perhaps due to the alleviated need for optimized core packing in a halophile context. Therefore, incorporation of aromatic amino acids into early proteins may have facilitated a critical halophile-to-mesophile transition. Subsequent incorporation of multiple aromatic groups within protein core regions can provided additional stability gains, enabling further adaptation into more demanding (i.e., extremophile) environments for protein folding, such as physical extremes of temperature or pH.

Materials and Methods

Protein expression and purification

Synthetic genes and mutagenesis primers were ordered from integrated DNA technologies. The 1xF and 2xF mutant proteins were constructed via site-directed mutagenesis following the Quikchange (Agilent Technologies, Santa Clara, CA) protocol. DNA sequences were verified before expression in *E. coli* BL21(DE3) competent cells. Transformed cells were grown in M9 media cultures, induced with 1 m*M* IPTG, and expressed for 8–10 h at 27°C. Cells were harvested via centrifugation at 5400*g* for 15 min at 4°C using and stored at −20°C. Cell pellets were resuspended in 5 m*M* imidazole, 50 m*M* NaPi, 500 m*M* NaCl, 0.01% Tween-80, pH 7.5. The cell suspension was lysed by passage through a French pressure cell at 1000 psi and the cell lysate was clarified by centrifugation at 29,600*g* for 60 min at 4°C. Expressed proteins contained an N-terminal $(His)_{6x}$ tag which has shown no influence upon stability or folding properties.[38] The supernatant was loaded onto a packed nickel affinity (Ni-NTA) chromatography column and the protein was eluted with 100 m*M* Imidazole, 500 m*M* NaCl, 50 m*M* NaPi, pH 7.5. Samples were further purified by gel filtration on a Superdex 75 column (GE Healthcare, Buckinghamshire, United Kingdom). The extinction coefficients for mutants containing W or Y residues were obtained by the Gill and von Hippel method.[47] Concentrations for all other mutant proteins were obtained using a bicinchoninic acid (BCA) assay using a standard curve generated against known concentrations of Symfoil-1. The purified protein was dialyzed against either phosphate buffer (100 m*M* NaCl, 10 m*M* $(NH_4)_2SO_4$, 50 m*M* NaPi pH 7.5) for crystallization studies or ADA buffer (20 m*M* *N*-(2-acetamido) iminodiacetic acid solution, 100 m*M* NaCl, pH 6.6.) for biophysical characterization and empirical phase diagram preparation.

X-ray crystallography

Purified protein was concentrated to 10–15 mg/mL in phosphate buffer. Crystal conditions were screened by the hanging drop vapor diffusion method at 25°C. 6xW crystals grew in 1.4*M* $(NH_4)_2SO_4$, 0.1*M* Tris, 0.07*M* Li_2SO4, pH 7.0; 6xY crystals grew in 30% (w/v) PEG 8,000, 0.1*M* Imidazole HCl, 0.2*M* NaCl, pH 8.0. Both crystals exhibited the same space group ($P2_12_12_1$), however, the crystal cell dimensions differ. 6xLeu (PV2; PDB ID code 4D8H) and 6xF (PDB ID code 3QYX) crystal structures have been previously reported.[13] Crystals were mounted using Hampton Research nylon cryo-loops and were cryo-cooled to 100 K by gaseous N_2 using an Oxford cryo-system (Oxford, UK). Crystals were diffracted in-house with a Rigaku RU-H3R rotating anode X-ray source (Rigaku, Tokyo, Japan) equipped with Osmic confocal mirrors (Osmic, Troy, MI) and a MarCCD165 detector (Rayonix, Evanston, IL). Data sets were analyzed with the DENZO software package to integrate, index, and scale all reflections. Molecular replacement for 6xW and 6xY was conducted using PV1 (PDB ID code 3QYX) as a search model with the PHENIX software program.[48]

Differential scanning calorimetry

DSC data was collected using a VP-DSC calorimeter (GE Healthcare, Buckinghamshire, United Kingdom). Three buffer–buffer scans were collected prior to protein loads to establish proper thermal history. 40 μ*M* protein samples in ADA Buffer were scanned from 10 to 95°C at a rate of 0.25°C/min under 34 psi. Analysis of the resulting endotherms was performed using the DSCfit software package.[49]

Empirical phase diagrams

CD was performed using a Chirascan-plus CD spectrometer (Applied Photophysics, Leatherhead, UK) equipped with a four-cuvette position Pelletier temperature controller (Quantum Northwest, Liberty Lake, WA) and a solid-state detector. The lamp, monochromator, and sample chamber were continually purged with N_2. Far-U.V. CD spectra of triplicate samples at 0.75 mg/mL were collected in the range of 260–200 nm in 1 nm steps and a 0.5 s sampling time using a quartz cuvette (0.1 cm path length) sealed with a Teflon stopper (Starna Cells Inc., Atascadero, CA). The CD signal at 230 nm was monitored as a function of temperature from 10 to 87.5°C at 2.5°C intervals. The heating rate was 1°C/min, and the equilibration time at each temperature was 1 min. The ellipticity of the buffer was subtracted from all measurements. All data were subjected to a three-point Savitzky–Golay smoothing filter using the Chirascan software (Applied Photophysics).

To quantify turbidity, the optical density at 360 nm was measured as a function of temperature (10.0–87.5°C) using a Cary-100 U.V.-Vis spectrophotometer equipped with a 12 cell-temperature controlled Pelletier (Agilent Technologies, Santa Clara CA). A 1°C/min heating rate, a 2 s integration time, and a 2 min equilibration time at each temperature were used. Samples were diluted in ADA Buffer to 0.2 mg/mL using a 1 cm path length quartz cuvette. The optical density of buffer alone was subtracted from all measurements.

DSC was performed using an Auto-VP capillary differential scanning calorimeter (MicroCal/GE Health Sciences) equipped with Tantalum sample and reference cells. Two water-water scans were taken prior to the reference and sample scans. Scans were completed from 10–90°C using a scanning rate of 15°C/h and a concentration of 1 mg/mL. Reference

subtraction and concentration normalization were performed using the instrument software.

Three-index EPDs were constructed as described[50] using the MiddaughSuite software. The DSC data was interpolated from 10.0–87.5°C at 2.5°C increments and integrated prior to EPD construction.

Acknowledgments

The X-ray Facility at Florida State University is acknowledged for assistance with data collection and processing. The authors declare no conflicts of interests.

References

1. Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117:528–529.
2. Hennet RJ-C, Holm NG, Engel MH (1992) Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: a perpetual phenomenon? Naturwissenschaften 79:361–365.
3. Kobayashi K, Kaneko T, Saito T, Oshima T (1998) Amino acid formation in gas mixtures by high energy particle irradiation. Orig Life Evol Biosph 28:155–165.
4. Wolman Y, Haverland WJ, Miller SL (1972) Nonprotein amino acids from spark discharges and their comparison with the Murchison meteorite amino acids. Proc Natl Acad Sci USA 69:809–811.
5. Chyba CF, Thomas PJ, Brookshaw L, Sagan C (1990) Cometary delivery of organic molecules to the early Earth. Science 249:366–373.
6. Longo LM, Blaber M (2012) Protein design at the interface of the pre-biotic and biotic worlds. Arch Biochem Biophys 526:16–21.
7. Longo LM, Blaber M (2014) Prebiotic protein design supports a halophile origin of foldable proteins. Front Microbiol 4:418.
8. Parker ET, Zhou M, Burton AS, Glavin DP, Dworkin JP, Krishnamurthy R, Fernandez FM, Bada JL (2014) A plausible simultaneous synthesis of amino acids and simple peptides on the primordial Earth. Angew Chem 126:8270–8274.
9. Romero P, Obradovic Z, Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 462:363–367.
10. Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152.
11. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680–1685.
12. Cockell CS, Airo A (2002) On the plausibility of a UV transparent biochemistry. Origins Life Evol Biosphere 32:255–274.
13. Longo L, Lee J, Blaber M (2013) Simplified protein design biased for pre-biotic amino acids yields a foldable, halophilic protein. Proc Natl Acad Sci USA 110: 2135–2139.
14. Oren A, Larimer F, Richardson P, Lapidus A, Csonka LN (2005) How to be moderately halophilic with broad salt tolerance: clues from the genome of *Chromohalobacter salexigens*. Extremophiles 9:275–279.
15. Paul S, Bag SK, Das S, Harvill ET, Dutta C (2008) Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol 9:R70.71–19.
16. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res 11:1641–1650.
17. Rode BM (1999) Peptides and the origin of life. Peptides 20:773–786.
18. Dundas I (1998) Was the environment for primordial life hypersaline? Extremophiles 2:375–377.
19. Cleaves HJI (2010) The origin of the biologically coded amino acids. J Theor Biol 263:490–498.
20. Wong JT-F (2005) Coevolution theory of the genetic code at age thirty. Bioessays 27:406–425.
21. Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261:139–151.
22. Merino E, Jensen RA, Yanofsky C (2008) Evolution of bacterial trp operons and their regulation. Curr Opin Microbiol 11:78–86.
23. Hernandez-Montes G, Diaz-Mejia JJ, Perez-Rueda E, Segovia L (2008) The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biol 9:R95.
24. Ringvold A, Anderssen E, Kjonniksen I (2000) UV absorption by uric acid in diurnal bird aqueous humor. Invest Ophthalmol Vis Sci 41:2067–2069.
25. Beadle BM, Shoichet BK (2002) Structural basis of stability—function tradeoffs in enzymes. J Mol Biol 321: 285–296.
26. Longo L, Lee J, Blaber M (2012) Experimental support for the foldability-function tradeoff hypothesis: segregation of the folding nucleus and functional regions in FGF-1. Protein Sci 21:1911–1920.
27. Lee J, Blaber SI, Dubey VK, Blaber M (2011) A polypeptide "building block" for the ß-trefoil fold identified by "top-down symmetric deconstruction". J Mol Biol 407:744–763.
28. Longo LM, Kumru OS, Middaugh CR, Blaber M (in press) Evolution and design of protein structure by folding nucleus symmetric expansion. Structure. doi: 10.1016/j.str.2014.08.008. [Epub ahead of print].
29. Broom A, Doxey AC, Lobsanov YD, Berthin LG, Rose DR, Howell PL, McConkey BJ, Meiering EM (2012) Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. Structure 20:1–11.
30. Dyer KF (1971) The quiet revolution: A new synthesis of biological knowledge. J Biol Educ 5:15–24.
31. King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798.
32. Doi N, Kakukawa K, Oishi Y, Yanagawa H (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. Prot Eng Des Sel 18:279–284.
33. McDonald GD, Storrie-Lombardi MC (2010) Biochemical constraints in a protobiotic Earth devoid of basic amino acids: the "BAA(-) World". Astrobiology 10:989–1000.
34. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 4:805–809.
35. Walter KU, Vamvaca K, Hilvert D (2005) An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 280:37742–37746.
36. Lee J, Blaber M (2011) Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. Proc Natl Acad Sci USA 108:126–130.

37. Alsenaidy MA, Wang T, Kim JH, Joshi SB, Lee J, Blaber M, Volkin DB, Middaugh CR (2012) An empirical phase diagram approach to investigate conformational stability of "second-generation" functional mutants of acidic fibroblast growth factor (FGF-1). Protein Sci 21:418–432.
38. Brych SR, Blaber SI, Logan TM, Blaber M (2001) Structure and stability effects of mutations designed to increase the primary sequence symmetry within the core region of a β-trefoil. Protein Sci 10:2587–2599.
39. Murzin AG, Lesk AM, Chothia C (1992) β-Trefoil fold. Patterns of structure and sequence in the kunitz inhibitors interleukins-1β and 1α and fibroblast growth factors. J Mol Biol 223:531–543.
40. Hecht MH, Sturtevant JM, Sauer RT (1984) Effect of single amino acid replacements on the thermal stability of the NH_2-terminal domain of phage λ repressor. Proc Natl Acad Sci USA 81:5685–5689.
41. Shortle D, Stites WE, Meeker AK (1990) Contributions of the large hydrophobic amino acids to the stability of Staphylococcal nuclease. Biochemistry 29:8033–8041.
42. Eriksson AE, Baase WA, Zhang X-J, Heinz DW, Blaber M, Baldwin EP, Matthews BW (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 255:178–183.
43. Lim WA, Farruggio DC, Sauer RT (1992) Structural and energetic consequences of disruptive mutations in a protein core. Biochemistry 31:4324–4333.
44. Wolfenden R, Andersson L, Cullis PM, Southgate CCB (1981) Affinities of amino acid side chains for solvent water. Biochemistry 20:849–855.
45. Blaber M, Lindstrom JD, Gassner N, Xu J, Heinz DW, Matthews BW (1993) Energetic cost and structural consequences of burying a hydroxyl group within the core of a protein determined from ala->ser and val->thr substitutions in T4 lysozyme. Biochemistry 32:11363–11373.
46. Schwendinger MG, Rode BM (1989) Possible role of copper and sodium chloride in prebiotic evolution of peptides. Analyt Sci 5:411–414.
47. Gill SC, von Hippel PH (1989) Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 182:319–326.
48. Zwart PH, Afonine PV, Grosse-Kunstleve RW, Hung LW, Loerger TR, McCoy AJ, McKee E, Moriarity NW, Read RJ, Sacchettini JC, Sauter NK, Storoni LC, Terwilliger TC, Adams PD (2008) Automated structure solution with the PHENIX suite. Methods Mol Biol 426:419–435.
49. Grek SB, Davis JK, Blaber M (2001) An efficient, flexible-model program for the analyis of differential scanning calorimetry protein denaturation data. Prot Pept Lett 8:429–436.
50. Kim JJ, Iyer V, Joshi SB, Volkin DB, Middaugh CR (2012) Improved data visualization techniques for analyzing macromolecule structural changes. Protein Sci 21:1540–1553.