doi:10.1016/j.jmb.2010.03.057 J. Mol. Biol. (2010) 398, 763–773

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computational and Experimental Evidence for the Evolution of a $(\beta\alpha)_8$-Barrel Protein from an Ancestral Quarter-Barrel Stabilised by Disulfide Bonds

Markus Richter[1], Manal Bosnali[2], Linn Carstensen[1], Tobias Seitz[1], Helmut Durchschlag[1], Samuel Blanquart[3], Rainer Merkl[1]* and Reinhard Sterner[1,2]*

[1]*Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany*

[2]*Institute of Biochemistry, University of Cologne, Otto-Fischer-Strasse 12–14, D-50674 Cologne, Germany*

[3]*Institut des Sciences de l'Evolution, Montpellier (ISEM), UMR 5554, Université de Montpellier II, CC 065, 34095 Montpellier Cedex 05, France*

Received 19 February 2010; received in revised form 19 March 2010; accepted 26 March 2010
Available online 2 April 2010

Edited by F. Schmid

The evolution of the prototypical $(\beta\alpha)_8$-barrel protein imidazole glycerol phosphate synthase (HisF) was studied by complementary computational and experimental approaches. The 4-fold symmetry of HisF suggested that its constituting $(\beta\alpha)_2$ quarter-barrels have a common evolutionary origin. This conclusion was supported by the computational reconstruction of the HisF sequence of the last common ancestor, which showed that its quarter-barrels were more similar to each other than are those of extant HisF proteins. A comprehensive sequence analysis identified HisF-N1 [corresponding to $(\beta\alpha)_{1-2}$] as the slowest evolving quarter-barrel. This finding indicated that it is the closest relative of the common $(\beta\alpha)_2$ predecessor, which must have been a stable and presumably tetrameric protein. In accordance with this prediction, a recombinantly produced HisF-N1 protein was properly folded and formed a tetramer being stabilised by disulfide bonds. The introduction of a disulfide bond in HisF-C1 [corresponding to $(\beta\alpha)_{5-6}$] also resulted in the formation of a stable tetramer. The fusion of two identical HisF-N1 quarter-barrels yielded the stable dimeric half-barrel HisF-N1N1. Our findings suggest a two-step evolutionary pathway in which a HisF-N1-like predecessor was duplicated and fused twice to yield HisF. Most likely, the $(\beta\alpha)_2$ quarter-barrel and $(\beta\alpha)_4$ half-barrel intermediates on this pathway were stabilised by disulfide bonds that became dispensable upon consolidation of the $(\beta\alpha)_8$-barrel.

Keywords: protein evolution; $(\beta\alpha)_8$-barrel; sequence reconstruction; disulfide bond

Introduction

Protein folds are characterised by distinct topological orientations of secondary-structure elements. The generation of the nearly 1200 different folds identified to date [Structural Classification of Proteins (SCOP), release 1.75, February 2009] seems to have been completed hundreds of millions of years ago, probably already in the last common ancestor of living organisms.[1] As a consequence, hypotheses of how folds have evolved must necessarily be based on circumstantial evidence. The *de novo* origin of folds is highly improbable, because the large majority of randomly generated proteins would have had an extremely low probability of adopting a well-defined structure. A more plausible assumption is that folds evolved by the duplication, fusion, and recombination of a small set of short supersecondary-structure elements such as ββ hairpins, αα hairpins, and βαβ elements.[2] In support of this hypothesis, a number of different folds are composed of repeating structural modules that consist of approximately 20–40 amino acids. Although most repeat proteins adopt elongated nonglobular structures, in some cases the repeating units form long-

**Corresponding authors.* R. Sterner is to be contacted at Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany. E-mail addresses: Rainer.Merkl@biologie.uni-regensburg.de; Reinhard.Sterner@biologie.uni-regensburg.de.

Abbreviation used: MSA, multiple sequence alignment.

0022-2836/$ - see front matter

range interactions to create a central hydrophobic core.[3] A prominent example where a repeating array of secondary-structure elements builds a well-defined globular protein is the $(\beta\alpha)_8$- or TIM-barrel, which belongs to the most frequent, versatile, and ancient enzyme folds.[4–6] The canonical $(\beta\alpha)_8$-barrel is composed of eight (βα) modules, each of which contains a minimum of about 25 residues. The β-strand and the α-helix within a given module are linked by a βα-loop, and the α-helix of module *i* is linked to the β-strand of module *i*+1 by an αβ-loop (Fig. 1a). The eight β-strands assemble to a central β-sheet, the barrel, which is surrounded by the eight α-helices (Fig. 1b).

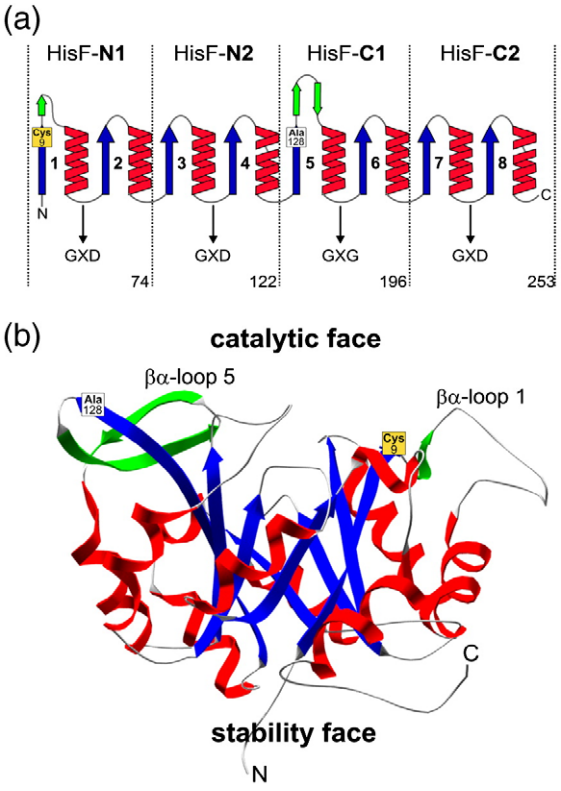

Fig. 1. Schematic depiction of the $(\beta\alpha)_8$-barrel fold of HisF. (a) Topology of the four quarter-barrels HisF-N1 $[(\beta\alpha)_{1-2}]$, HisF-N2 $[(\beta\alpha)_{3-4}]$, HisF-C1 $[(\beta\alpha)_{5-6}]$, and HisF-C2 $[(\beta\alpha)_{7-8}]$. The α-helices are marked in red, and the β-strands are marked in blue. The βα-loops and αβ-loops connect the secondary-structure elements within and between individual (βα)-units; the long βα-loops 1 and 5 contain additional β-strands, which are marked in green. The conserved GXD/GXG motif within the αβ loops 1, 3, 5, and 7 is indicative of a 4-fold symmetry of the $(\beta\alpha)_8$-barrel fold. Residue Cys9 at the end of β-strand 1 and the exchange of Ala128 for Cys at the end of β-strand 5 result in the formation of disulfide bonds in HisF-N1 and HisF-C1. The N- and C-termini are marked; the sequence position of the last residue of each quarter-barrel is indicated at the bottom. (b) Ribbon diagram of the three-dimensional structure of HisF. The active-site residues are located at the C-terminal end of the central β-barrel and within the βα loops (catalytic face). The remainder of the fold, including the αβ loops, are important for conformational stability (stability face).

In order to understand the evolution of the $(\beta\alpha)_8$-barrel fold, we earlier paradigmatically characterised two related enzymes involved in the biosynthesis of histidine, namely, *N*′-[(5′-phosphoribosyl) formimino]-5-aminoimidazole-4-carboxamide ribonucleotide isomerase (HisA) and imidazole glycerol phosphate synthase (HisF) from *Thermotoga maritima*. Both HisA and HisF display a twofold sequence and structural symmetry,[7,8] and the recombinant N- and C-terminal half-barrels HisF-N and HisF-C form stable and predominantly dimeric proteins with defined secondary and tertiary structures.[9] Furthermore, the duplication and tandem fusion of two copies of HisF-C, followed by the optimisation of the intramolecular interface, allowed us to generate a $(\beta\alpha)_8$-barrel protein from identical $(\beta\alpha)_4$ half-barrels whose X-ray structure is similar to that of wild-type HisF.[10–12] Moreover, the fusion of HisA-N with HisF-N yielded a stable $(\beta\alpha)_8$-barrel protein on which catalytic activity could be established by a combination of random mutagenesis and selection *in vivo*.[10,13] These results suggest that the $(\beta\alpha)_4$ half-barrel can be fused, mixed, and matched to yield new $(\beta\alpha)_8$-barrels.[14]

As observed for other $(\beta\alpha)_8$-barrel proteins,[15] the 2-fold symmetry of HisF can be further broken down into a 4-fold symmetry (Fig. 1a), suggesting that the enzyme has evolved by two gene duplication and fusion events from an ancestral $(\beta\alpha)_2$ quarter-barrel *via* a $(\beta\alpha)_4$ half-barrel into the $(\beta\alpha)_8$-barrel. We have tested this hypothesis by a combination of complementary computational and experimental approaches. The reconstruction of the HisF sequence of the last common ancestor (HisF-LUCA) showed that its quarter-barrels HisF-N1 $[(\beta\alpha)_{1-2}]$, HisF-N2 $[(\beta\alpha)_{3-4}]$, HisF-C1 $[(\beta\alpha)_{5-6}]$, and HisF-C2 $[(\beta\alpha)_{7-8}]$ were more similar than those of extant HisF proteins, supporting the existence of a common ancient $(\beta\alpha)_2$ predecessor. Calculated evolutionary rates and comprehensive sequence analysis identified HisF-N1 as the closest relative of this predecessor. In accordance with this finding, the purified recombinant HisF-N1 protein from *T. maritima* was more stable than the other quarter-barrels and formed a tetramer. The fusion of two HisF-N1 elements yielded the stable dimeric half-barrel HisF-N1N1. Both HisF-N1 and HisF-N1N1 contain intermolecular disulfide bonds, which was probably an efficient way to stabilise these intermediates during the evolution of the $(\beta\alpha)_8$-barrel fold of HisF.

Results

Reconstruction of ancient HisF sequences

The architecture of the central β-barrel of HisF shows a 4-fold symmetry (Fig. 1), suggesting that the $(\beta\alpha)_8$-barrel has evolved from a common ancestral $(\beta\alpha)_2$ element. However, sequence similarities of only 15–26% between the HisF quarter-barrels from *T. maritima*[16] are in or below the

twilight zone,[17] and the analysis of extant HisF sequences by highly sensitive methods[18] did not convincingly support a common origin from a $(\beta\alpha)_2$ predecessor. We reasoned that the reconstruction of the HisF-LUCA sequence should provide new insights into this problem, because in case of a common origin, the quarter-barrels of ancient HisF proteins should be more similar to each other than those of extant ones. It is important to note, however, that this result can only be expected if mutations in the pre-LUCA and LUCA era were not to hide the common origin.

For the reconstruction of HisF-LUCA we utilized programs of the PhyloBayes[19] and the nh_PhyloBayes[19] software suite, which implement homogeneous and nonhomogeneous Bayesian models of evolution. A prerequisite for the reconstruction of ancient sequences is the calculation of a highly reliable phylogenetic tree. Having in mind that horizontal transfer of *his* genes between Bacteria and Archaea is frequent,[20] we carefully selected sequences based on the HisF entry (1gpw chain A) of the DSSP database[21] according to the following criteria. Sequences that were strikingly incongruent with the "nearly universal consensus trees" of life were discarded.[22] For example, euryarcheal HisF sequences did not form a monophyletic group with crenarcheal sequences, which are closer to the root of the consensus trees, and were therefore excluded from tree reconstruction. Moreover, in order to increase the strength of the phylogenetic signal, HisF and HisH sequences originating from the same genome were concatenated. Their coevolution within the same species is highly plausible, because these two proteins form an obligate dimer within prokaryotes.[23,24] The resulting multiple sequence alignment MSA_{HisFH} contained 87 sequences from the clades Crenarchaeota, Actinobacteria, Chlorobi, Cyanobacteria, Firmicutes, Proteobacteria, and Thermotogae. Based on MSA_{HisFH}, the CAT model[19] was used to compute eight independent phylogenetic trees. Since their topologies were identical, a highly reliable consensus tree could be calculated (Fig. S1). Based on this consensus tree and the HisF section of the alignment (MSA_{HisF}; Fig. S2), ancestral sequences were computed.[19] Considering tree nodes with a posterior probability >0.95 as valid, we reconstructed the HisF sequence of an ancestor for Bacteria and Crenarchaeota (HisF-LUCA), and deepest ancestors for Cyanobacteria (*HisF-*Anc_{Cyano}), Firmicutes (*HisF-*Anc_{Firm}), and Proteobacteria (*HisF-*Anc_{Prot}) (Fig. S2).

Mean percentage sequence identities for the extant quarter-barrel pairs HisF-N1/HisF-C1, HisF-N2/HisF-C2, HisF-N1/HisF-N2, HisF-N1/HisF-C2, HisF-C1/HisF-C2, and HisF-N2/HisF-C1, as well as the sequence identities of the corresponding pairs of HisF-LUCA quarter-barrels are given in Table 1. In five out of six cases, the reconstructed quarter-barrels turned out to be more similar to each other than the extant ones with statistical significance ($p<0.001$; one sample t test); the exception is the pair HisF-N2/HisF-C2. An analogous analysis of *HisF-*Anc_{Cyano}, *HisF-*Anc_{Firm}, and *HisF-*Anc_{Prot} confirmed that quarter-barrels of ancient HisF proteins are more similar to each other than those of extant ones (Table 1). These findings support the notion that quarter-barrels have a common evolutionary origin.

Identifying the closest relative of the ancient quarter-barrel

In order to find out which of the four HisF quarter-barrels is most similar to the common $(\beta\alpha)_2$ predecessor, we computed normalized evolutionary rates for each quarter barrel, given the posterior stochastic mapping inferred from MSA_{HisFH}.[19] The results showed that HisF-N1 is the slowest-evolving quarter-barrel (mean branch length, 0.06 ± 0.01). The fastest-evolving one is HisF-C1 (mean branch length, 0.11 ± 0.03); HisF-N2 and HisF-C2 are in between (mean branch length, 0.08 ± 0.02 and 0.09 ± 0.02, respectively) (Table S1). To confirm these findings, we compared the sequences of the quarter-barrels of HisF-LUCA with those of MSA_{HisF}. The resulting data set HisF-LUCA/MSA_{HisF} yielded mean identities between ancient and extant sequences of 71% for HisF-N1, 66% for HisF-N2, 67% for HisF-C1, and 60% for HisF-C2. A Mann-Whitney rank sum test[25] proved that sequence conservation of HisF-N1 is higher than that of all other quarter-barrels with statistical significance ($p<0.001$). The comparison of the extant quarter-barrels from MSA_{HisF} yielded mean identities of 63% for HisF-N1, 57% for HisF-N2, 58% for HisF-C1, and 52% for HisF-C2 (Table S1). In all four cases, the mean identities between ancient and extant sequences are higher than those of extant ones with statistical significance ($p<0.001$; Mann–

Table 1. Cross-comparison of sequences from HisF-N1, HisF-N2, HisF-C1, and HisF-C2 quarter-barrels

Data set	HisF-N1/HisF-C1	HisF-N2/HisF-C2	HisF-N1/HisF-N2	HisF-N1/HisF-C2	HisF-C1/HisF-C2	HisF-N2/HisF-C1
HisF-LUCA	21/22	**19/15**	16/17	12/18	13/19	14/23
HisF-Anc_{Cyano}	19/22	19/21	18/21	13/14	**15/11**	**13/11**
HisF-Anc_{Firm}	19/20	15/16	**13/7**	12/16	11/23	14/16
HisF-Anc_{Prot}	23/27	21/23	**17/15**	12/15	**13/9**	16/18

For each data set, two values are given. The first value is the mean percentage of identical residues originating from cross-comparisons of extant quarter-barrels from MSA_{HisF} (line HisF-LUCA) or subsets (following lines). The second value is the percentage of identical residues when cross-comparing two quarter-barrels of the respective ancestral sequence. HisF-LUCA is the reconstructed predecessor of all MSA_{HisF} entries. *HisF-*Anc_{phylum} is the reconstructed predecessor of one of the phyla Cyanobacteria, Firmicutes, and Proteobacteria. If sequence conservation between quarter-barrels of the ancestor is lower than the mean value deduced from extant quarter-barrels, the pair of numbers is printed in bold.

Whitney rank sum test). The different degree of sequence conservation of the four quarter-barrels was further confirmed by analyzing a data set of 626 extant HisF sequences comprising seven bacterial and archeal phyla. In all cases, sequence conservation was highest for HisF-N1. For several data sets, sequence conservation did not perfectly correlate with the evolutionary rates of HisF-N2, HisF-C1, and HisF-C2. Most plausibly, this is due to the expected differences in evolutionary rates of individual genes. In summary, our sequence analysis identifies HisF-N1 as the closest relative of the ancestral quarter-barrel.

Production and purification of recombinant quarter-barrels

The finding that HisF-N1 is the most slowly evolving and most conserved quarter-barrel suggests that it has properties of the ancient quarter-barrel, which must have been a stable protein and probably formed a tetramer. To test this prediction, we attempted to produce and characterise the four *T. maritima* HisF quarter-barrels. HisF-N1 and HisF-C1 contain 74 residues and have molecular masses of 8.3 kDa each, whereas HisF-N2 and HisF-C2 contain only 48 and 58 residues, corresponding to molecular masses of 5.0 and 6.3 kDa, respectively. This difference is caused by the long βα-loops 1 and 5 in HisF-N1 and HisF-C1 (Fig. 1a), respectively, each of which contains an additional antiparallel β-sheet.[7] The genes encoding the four quarter-barrels of HisF were expressed individually in *Escherichia coli*. The analysis by SDS-PAGE showed that both His-N2 and HisF-C2 were not present either in the soluble or in the insoluble fraction of the cell extract, suggesting that both recombinant proteins were unstable and rapidly degraded *in vivo*. In contrast, large amounts of HisF-N1 and HisF-C1 were found in the insoluble cell fraction of the cell extract. Both recombinant proteins were solubilised by incubation in 6 M guanidinium chloride, refolded by dialysis against phosphate buffer, and purified to homogeneity by preparative gel-filtration chromatography.

Characterisation of HisF-N1

When purified HisF-N1 was analysed by analytical gel-filtration chromatography, the apparent

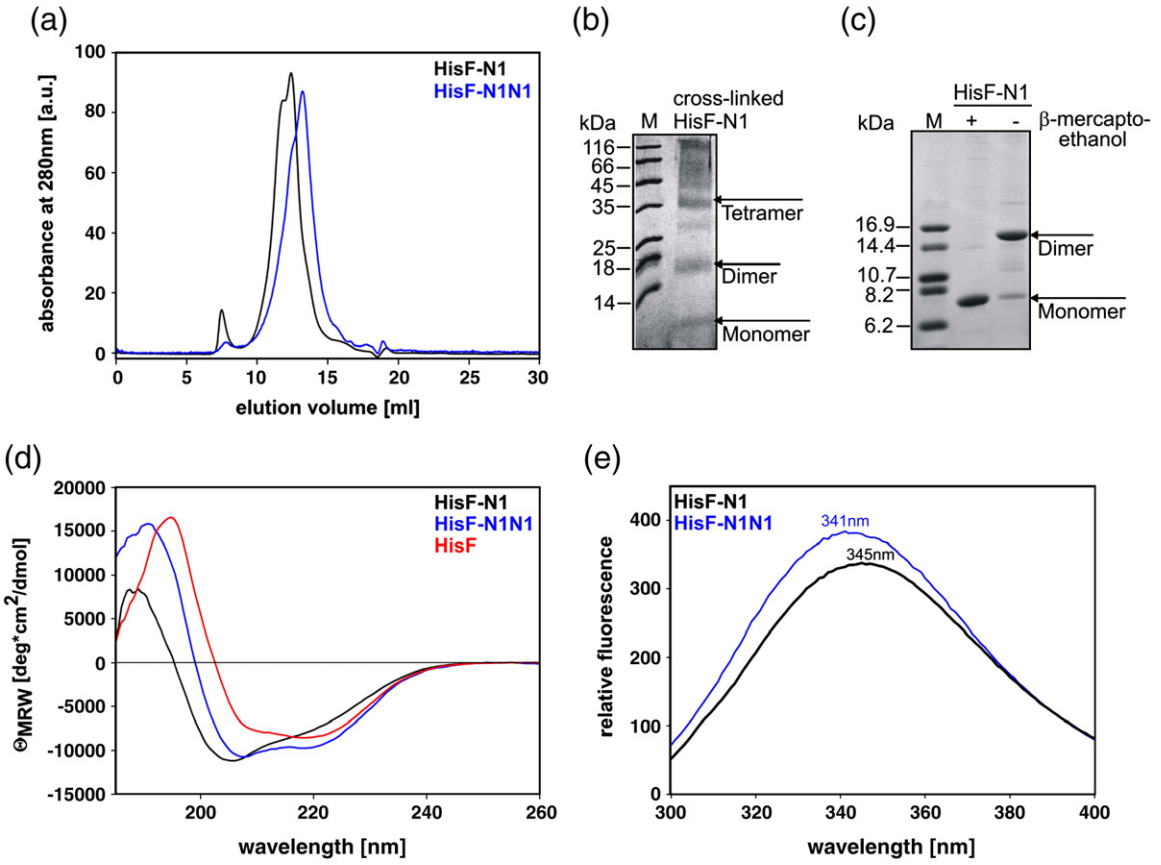

Fig. 2. Characterisation of HisF-N1 and HisF-N1N1. (a) Elution profiles of analytical gel-filtration runs (Superdex 75 column) of HisF-N1 and HisF-N1N1. (b) SDS-PAGE of HisF-N1 following cross-linking with glutardialdehyde (Tris–tricine gel; 20% polyacrylamide). The predominant association states are indicated. M, marker proteins with molecular masses on the left. (c) SDS-PAGE of purified HisF-N1 (Tris–tricine gel; 20% polyacrylamide) in the presence and absence of β-mercaptoethanol. M, marker proteins with molecular masses on the left. (d) Far-UV CD spectra of HisF, HisF-N1, and HisF-N1N1. (e) Fluorescence emission spectra following excitation at 280 nm of HisF-N1 and HisF-N1N1. The emission maxima are indicated.

molecular mass of the elution peak as deduced from a calibration curve was 43 kDa, corresponding to the pentamer (Fig. 2a). Since this stoichiometry appeared to be improbable given the 4-fold symmetry of the $(\beta\alpha)_8$-barrel fold (Fig. 1a), the association state of HisF-N1 was investigated by chemical cross-linking with glutardialdehyde. The results were analysed by SDS-PAGE, which showed that HisF-N1 predominantly associated to tetramers and a certain fraction of probably unspecific higher oligomers. However, a certain amount of dimers and monomers could also be detected, probably because the cross-linking reaction was incomplete (Fig. 2b). To obtain further information about the oligomerisation state of HisF-N1, we performed analytical ultracentrifugation. The apparent molecular mass of 37 kDa detected in a sedimentation equilibrium run is in good agreement with the calculated molecular mass of 33.6 kDa for the tetramer. Information about the stabilisation of the quaternary structure was obtained by SDS-PAGE, which showed that HisF-N1 migrated as a monomer in the presence of β-mercaptoethanol, but as a dimer in the absence of the reducing agent (Fig. 2c). This result suggests that two HisF-N1 chains are covalently linked by a disulfide bond *via* the single cysteine residue 9 from β-strand 1 (Fig. 1a). Remarkably, this cysteine residue is strictly conserved among the known HisF sequences (Fig. S2) but dispensable for the function of the enzyme.[23] The two covalently linked subunits associate with a second identical dimer to the tetramer.

The formation of native secondary structure by HisF-N1 was investigated by far-UV CD spectroscopy (Fig. 2d). The analysis of the spectrum with the ContinLL program predicted α-helical and β-strand contents of 19% and 24%, respectively, which are somewhat lower and similar to the α-helical (31%) and β-strand (22%) contents of the HisF-N1 fragment within the X-ray structure of HisF [Protein Data Bank (PDB) code 1thf]. When a CD spectrum of HisF-N1 was monitored in the presence of 10% trifluoroethanol, which promotes the formation of ordered secondary structures,[26] the predicted α-helical content in solution was increased to the one observed within the X-ray structure of HisF. The HisF-N1 protein analysed in this work contained the residue exchange Phe23Trp in βα-loop 1, which allowed us to investigate the formation of tertiary structure by fluorescence spectroscopy. The emission maximum of the native protein is 345 nm (Fig. 2e), which is blue-shifted compared to that of the denatured protein (356 nm), indicating that HisF-N1 partly shields Trp23 from the solvent. This effect can be caused by the formation of tertiary structure within the individual HisF-N1 monomers or by quaternary contacts within the tetramer. The stability of HisF-N1 was analysed by chemical unfolding in urea, which was followed by monitoring the loss of tertiary and secondary structure by fluorescence and far-UV spectroscopy. Under oxidizing conditions, the protein unfolded with moderate cooperativity and $D_{1/2}$ values of 2.5 and 2.7 M urea, respectively. In contrast, in the presence of a reducing agent, unfolding was noncooperative, and the $D_{1/2}$ value was lowered to 2.0 M urea (Fig. S3). These results demonstrate that the compactness and the conformational stability of HisF-N1 are increased by disulfide bond formation.

Characterisation and stabilisation of HisF-C1

In contrast to HisF-N1, purified HisF-C1 formed a mixture of various ill-defined oligomers and had a tendency to aggregate (Fig. 3a). We attempted to stabilise HisF-C1 by introducing the amino acid substitutions Ala128Cys and Tyr143His. The structural superposition of HisF-N1 with HisF-C1 showed that the residue corresponding to Cys9 from β-strand 1 is Ala128 from β-strand 5 (Fig. 1a). We therefore reasoned that the Ala128Cys exchange might lead to the formation of a stabilising disulfide bond. Moreover, it has been shown that the Tyr143His exchange in αβ-loop 5 results in increased solubility and stability of two fused HisF-C half-barrels,[12] and we speculated that this substitution might also be beneficial in the background of the HisF-C1 quarter-barrel. The HisF-C1-Ala128Cys + Tyr143His protein was generated by site-directed mutagenesis and heterologous expression in *E. coli*, followed by refolding from the insoluble cell extract and purification by preparative gel-filtration chromatography.

SDS-PAGE in the absence and presence of β-mercaptoethanol showed that HisF-C1-Ala128Cys + Tyr143His indeed formed a dimer *via* the newly introduced cysteine 128 (Fig. 3c). Moreover, the symmetric elution profile obtained by analytical gel-filtration chromatography indicated that HisF-C1-Ala128Cys + Tyr143His is a rather homogeneous protein (Fig. 3a), although chemical cross-linking experiments detected a mixture between the monomer, the dimer, and the tetramer (Fig. 3b). To identify the predominant association state of HisF-C1-Ala128Cys + Tyr143His in solution, we performed analytical ultracentrifugation. The sedimentation equilibrium run yielded an apparent molecular mass of 35 kDa, which is in good agreement with the calculated mass of 33.6 kDa for the tetramer. These data indicate that the HisF-C1-Ala128Cys + Tyr143His protein forms a disulfide-bond-stabilised tetramer in solution, as does HisF-N1.

The effect of the two introduced exchanges for the secondary and tertiary structure of the quarter-barrel was investigated by far-UV CD and fluorescence emission spectroscopy. The CD spectrum of HisF-C1-A128Cys + Tyr143His is very similar to that of HisF-C1 (Fig. 3d). ContinLL predicted α-helical and β-strand contents of about 27% and 17% for the quarter-barrels, which are similar and lower, respectively, than the α-helical (24%) and β-strand (37%) contents of HisF-C1 within the X-ray structure of HisF (PDB code 1thf). The addition of 10–20% trifluoroethanol resulted in calculated α-helical and β-strand contents that were identical to those of

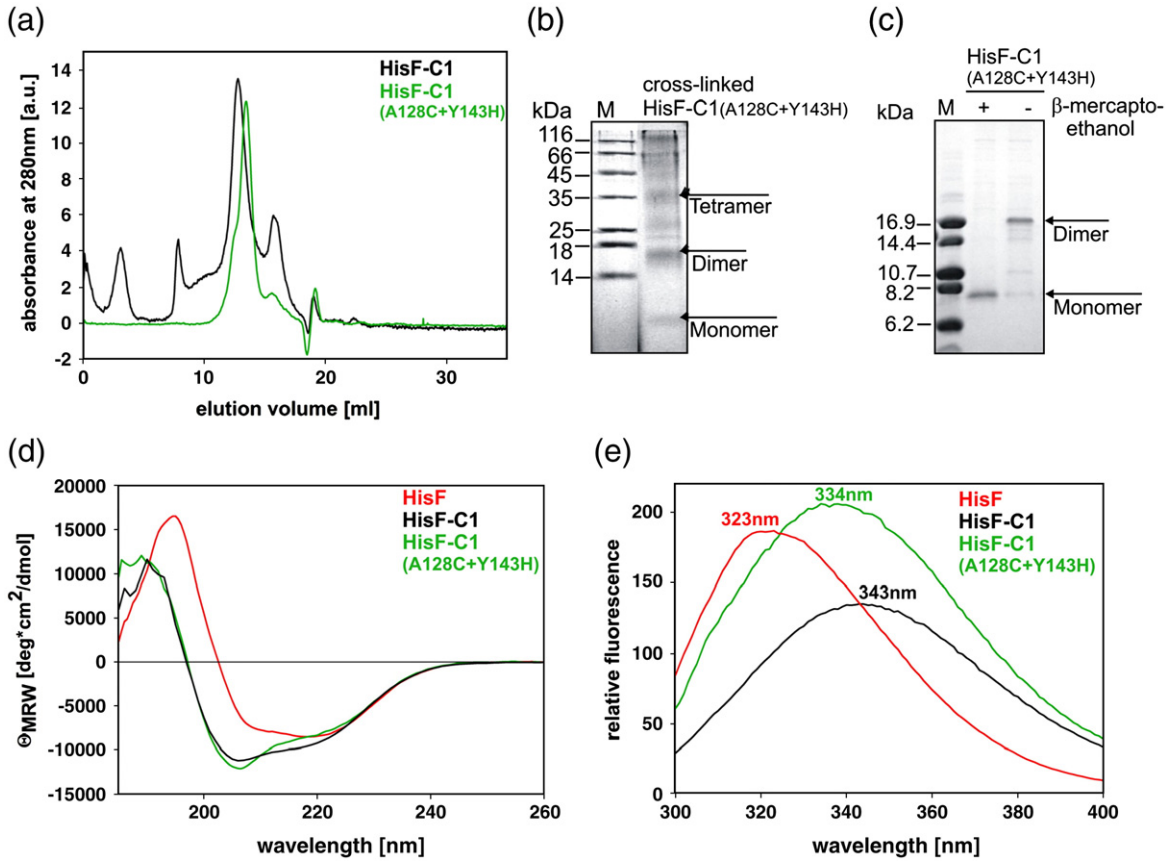

Fig. 3. Characterisation of HisF-C1 and HisF-C1-A128C+Y143H. (a) Elution profiles of analytical gel-filtration chromatography (Superdex 75 column) runs of HisF-C1 and HisF-C1-A128C+Y143H. (b) SDS-PAGE of HisF-C1-A128C+Y143H following cross-linking with glutardialdehyde (Tris–tricine gel; 20% polyacrylamide). The predominant association states are indicated. M, marker proteins with molecular masses on the left. (c) SDS-PAGE of purified HisF-C1-A128C+Y143H (Tris–tricine gel; 20% polyacrylamide) in the presence and absence of β-mercaptoethanol. M, marker proteins with molecular masses on the left. (d) Far-UV CD spectra and (e) fluorescence emission spectra following excitation at 280 nm of HisF, HisF-C1, and HisF-C1-A128C+Y143H. The emission maxima are indicated.

HisF-C1 in the HisF structure. The single tryptophan residue 156 of HisF is located in α-helix 5 and therefore also present in HisF-C1 and HisF-C1-A128Cys+Tyr143His, which allowed us to compare the fluorescence spectra of the three proteins. The emission maximum of HisF-C1-A128Cys+Tyr143His (334 nm) is located in between the maxima of HisF (323 nm) and HisF-C1 (343 nm), indicating that the Trp156 residue is shielded from solvent more efficiently in the double mutant compared to HisF-C1 (Fig. 3e). Chemical denaturation of HisF-C1 and HisF-C1-A128Cys+Tyr143His in urea, which was followed by fluorescence spectroscopy, occurred with modest cooperativity for both proteins. However, the conformational stability of the double mutant is elevated compared to that of HisF-C1, as documented by $D_{1/2}$ values of 1.7 and 3.5 M, respectively (Fig. S4).

Production and characterisation of HisF-N1N1

Following our reconstruction of an ancestral (βα)$_8$-barrel from (βα)$_4$ half-barrels,[10–12] we wished to generate (βα)$_4$ half-barrels from the ancestral-type HisF-N1 quarter-barrel. To this end, the *his*F-N1 gene was duplicated and fused in tandem to generate *his*F-N1N1, and the recombinant protein was generated by heterologous expression in *E. coli*. In contrast to HisF-N1, a considerable fraction (about 10–20%) of HisF-N1N1 was found in the soluble cell extract. Nevertheless, the recombinant protein was refolded from the insoluble cell extract and purified by preparative gel-filtration chromatography. SDS-PAGE showed that purified HisF-N1N1 in the presence of β-mercaptoethanol migrates as a monomer (calculated molecular mass, 16.6 kDa) and as a dimer in its absence, demonstrating disulfide bond formation *via* Cys9, as observed for HisF-N1. Analytical gel-filtration chromatography under oxidizing conditions revealed a symmetrical peak at an elution time corresponding to an apparent molecular mass of 28 kDa, which is in reasonable agreement with the calculated molecular mass of 33.2 kDa for the dimer (Fig. 2a). A sedimentation equilibrium run in the analytical ultracentrifuge confirmed that HisF-N1N1 mainly forms a dimer in solution.

The shape of the far-UV CD-spectrum of HisF-N1N1 lies in between the spectra of HisF-N1 and HisF (Fig. 2d). ContinLL predicted α-helical and β-

strand contents of 29% and 17%, which are in good agreement with the α-helical (31%) and β-strand (22%) contents of the HisF-N1 fragment within the X-ray structure of HisF. The emission maximum of the native protein is 341 nm, which is blue-shifted compared to that of the denatured protein (353 nm; not shown), indicating that the introduced Trp23 is shielded from solvent in HisF-N1N1 to a comparable extent as in HisF-N1 (Fig. 2e). Chemical unfolding in urea followed by fluorescence spectroscopy showed that the stability of HisF-N1N1 is comparable to that of HisF-N1, as indicated by identical $D_{1/2}$ values of 3.5 M. However, unfolding of HisF-N1N1 occurs with a higher cooperativity, indicating a better defined structure (Fig. S3). In conclusion, the duplication of the $(\beta\alpha)_2$ quarter-barrel HisF-N1 to the $(\beta\alpha)_4$ half-barrel HisF-N1N1 leads to a more soluble and compact protein, while the same quaternary structure is maintained.

Discussion

Quarter-barrel sequences were shaped in the pre-LUCA eon

Using advanced models of phylogenetic analysis, we reconstructed the HisF sequence of an ancestor for Bacteria and Crenarchaeota and deepest ancestors for Cyanobacteria, Firmicutes, and Proteobacteria. The two most important phenomena that preclude reliable sequence reconstruction are horizontal gene transfer and a weak phylogenetic signal. The composition of extant *his* operons indicates that several recombination events and horizontal gene transfer between Bacteria and Archaea have taken place and suggests the absence of a complete operon in the archeal ancestor and presumably in the LUCA. Accordingly, phylogenetic trees deduced from HisG sequences and from the concatenated sequences of five histidine biosynthesis proteins differ in topology.[20] To avoid artefacts, these findings asked for a critical control of sequence selection, which let us ignore all sequences that were incompatible with the canonical tree of life. However, even a data set free of horizontal gene transfer does not guarantee a reliable tree if the embedded phylogenetic signal is weak. To circumvent this problem, HisH sequences were exploited in addition to HisF, as these two proteins, which form an obligate dimer, have most probably coevolved. Based on this carefully selected information, we reconstructed reliable ancestral HisF sequences. Our findings suggest a substantial amount of mutations in the pre-LUCA era, as sequence identity among different quarter-barrels in HisF-LUCA was already reduced to approximately 20%. We therefore have to assume that after the postulated gene fusion events in the genome of a pre-LUCA, numerous modifications have altered the composition and length of the originally identical quarter-barrel sequences. Since then, this value has decreased to approximately 15% identical residues as seen in the cross-comparisons of extant quarter-barrels (Table 1).

HisF-N1 is a model for the ancient quarter-barrel

Our comprehensive computational analysis of extant HisF sequences showed that within all analysed data sets, the conservation of HisF-N1 is higher than for the other three quarter-barrels (Table S1). These findings suggested that HisF-N1 is the closest relative of the ancient quarter-barrel, which must have been a stable and presumably tetrameric protein. In accordance with this prediction, recombinant HisF-N1 could be purified and characterised, whereas HisF-C1 required mutational stabilisation, and the two smaller proteins HisF-N2 and HisF-C2 obviously were degraded following their expression in *E. coli*. Probably, HisF-N2 and HisF-C2 in isolation do not have the minimum size to form an interior hydrophobic core, which could compensate for the entropically unfavourable ordering of the peptide backbone and the side chains upon folding.[27] In accordance with our findings, the characterisation of fragments of tachylectin 2 and the tumor suppressor protein $p16^{INK4}$, which belong to the WD40 and the ankyrin repeat families, respectively, has shown that a minimum of two repeats, corresponding to 100 residues for tachylectin 2 and 66 residues for $p16^{INK4}$, are required for the formation of structured proteins.[28,29]

HisF-N1 and HisF-C1-Ala128Cys+Tyr143His are stabilised by the formation of disulfide bonds. The beneficial effect of disulfide bonds is based on the decrease in the entropy of the unfolded state, which results in an increased free-energy difference to the folded state.[30] Since the cytoplasm is generally reducing, only a few intracellular prokaryotic proteins with disulfide bonds have been identified to date, which, moreover, seem to be only transiently formed and appear to be more important for function than for stability.[31,32] However, the analysis of a number of X-ray structures of proteins from hyperthermophilic archaea suggests that disulfide bonds contribute to their high thermal stability, an assumption that is supported by computational genome analysis.[33,34] Further stabilisation of HisF-N1 and HisF-C1-Ala128Cys+Tyr143His is achieved by the noncovalent association of two dimers (each composed of two disulfide-linked monomers) to the tetramer. In accordance with this finding, protein design studies have shown that oligomerisation is an important mechanism to increase conformational stability.[35–37] Along the same lines, proteins from hyperthermophiles often show a higher association state than their homologues from mesophiles.[38,39]

Model for the evolution of HisF from a HisF-N1-like quarter-barrel

Together, the computational reconstruction of ancient sequences as well as the properties of purified HisF quarter-barrels and the HisF-N1N1 half-barrel lead to the three-step model of $(\beta\alpha)_8$-

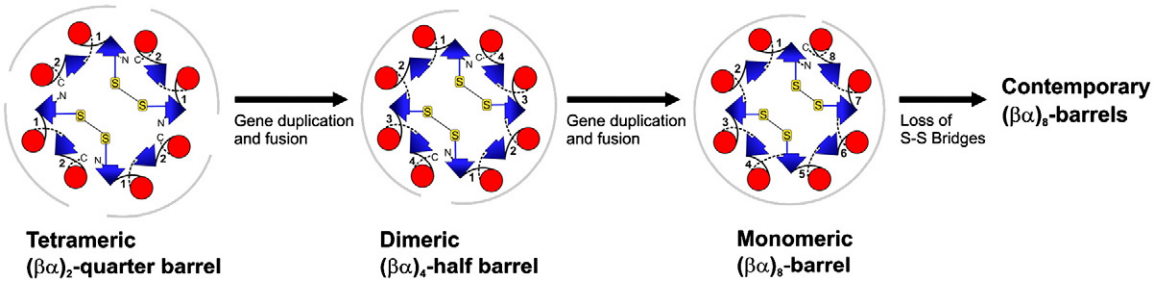

Fig. 4. Three-step model for the evolution of an ancestral $(\beta\alpha)_8$-barrel protein HisF from disulfide-linked $(\beta\alpha)_2$ quarter-barrels *via* $(\beta\alpha)_4$ half-barrels.

barrel evolution outlined in Fig. 4. The lower level of stability and higher level of heterogeneity of HisF-N1 under reducing conditions compared to oxidizing conditions suggest that disulfide bonds played a crucial role for the formation of a stable tertiary and quaternary structure of the original $(\beta\alpha)_2$ quarter-barrel. The highly conserved cysteine residue, which is not important for the function of HisF,[23] might be a remnant of this ancestral situation. In a more general sense, disulfide bonds might have played an important role at the early stages of fold evolution where small protein fragments had to be stabilised to an extent that allowed them to be functional. Following gene duplication and fusion, larger and more stable proteins certainly were less dependent on the stabilising effect of these specific cross-links, most of which were consequently lost (Fig. 4). Most likely, the ancient HisF-LUCA protein no longer depended on this effect, as it contained no other cysteine residues than Cys9 (Fig. S2). Interestingly, the sequence of HisF-LUCA contains a histidine residue at position 143, which is in agreement with the stabilising effect of the Tyr143His mutation for HisF-C1.[12]

Conclusion

The combination of results from independent computational and experimental approaches has led to the most parsimonious model of HisF evolution shown in Fig. 4. The model is reliable, because the information obtained from the two approaches is complementary and the drawn conclusions are mutually supportive. Computational biology suggested the existence of a common predecessor of the HisF quarter-barrels and argued that HisF-N1 is its closest extant relative. It is plausible to assume that the original quarter-barrel was a folded protein, a prediction that was confirmed by the experimental characterisation of HisF-N1. Moreover, it was shown that four HisF-N1 copies assemble to a tetramer being stabilised by disulfide bonds, which would have been impossible to predict by current computational methods. Since it is assumed that HisF is an old $(\beta\alpha)_8$-barrel protein,[40] the drawn scenario might be paradigmatic for the evolution of this ubiquitous and versatile fold. We believe that similar combined computational–experimental approaches such as the one outlined here will provide new insights into the evolution of other folds as well.

Materials and Methods

Cloning of *his*F-N1, *his*F-N1N1, *his*F-C1, and *his*F-C1-A128C + Y143H

The *his*F-N1 gene (corresponding to base pairs 1–222 of *his*F) carrying the Phe23Trp exchange to facilitate purification, concentration determination, and spectroscopic characterisation was amplified by PCR, using pET11c-*his*F-F23W + W156F[16] as template, and the oligonucleotides 5′-AGC**CATATG**CTCGCTAAAAGAATAATCGCG-3′ (newly introduced NdeI restriction site in bold) and 5′-ATA**GGATCC**TCAGTCGATCTGCTCGGCCAC-3′ (newly introduced BamHI restriction site in bold) as 5′ and 3′ primers, respectively.

Using the template pET21a-*his*F-N1-R5A + F23W + E46Q, which had been generated in an attempt to stabilise HisF-N1,[16] two copies of the *his*F-N1 gene were amplified in two different PCRs. In one PCR, the oligonucleotide T7 promotor (Stratagene) was used as 5′ primer, and the oligonucleotide 5′-GGAATA**GCTAGC**CTCGGCCACC-TTTTCGACCAG-3′ (newly introduced NheI restriction site in bold) was used as 3′ primer. In the other PCR, the oligonucleotide 5′-CTA**GCTAGC**GCTAAAGCGAT-AATC-3′ (newly introduced NheI restriction site in bold) was used as 5′ primer, and the oligonucleotide 5′-ATA**G-GATCC**TCAGTCGATCTGCTCGGCCAC-3′ (newly introduced BamHI restriction site in bold) as 3′ primer. Both amplification products were first digested with NheI and then ligated, yielding *his*F-N1N1. Both halves of the resulting HisF-N1N1 polypeptide chain carried the exchanges R5A + F23W + E46Q. Since these substitutions do not influence the properties of the protein, they are not explicitly mentioned in the study.

The *his*F-C1 gene (corresponding to base pairs 366–588 of *his*F) was amplified by PCR, using the plasmid SK+/III P-P[8] as template, and the oligonucleotides 5′-ATA**CA-TATG**CAGGCCGTTGTCGTGGTGGCGATA-3′ (newly introduced NdeI restriction site in bold) and 5′-ATA**G-GATCC**TCATGTGGTTAGTGGCCTCAC-3′ (newly introduced BamHI restriction site in bold) as 5′ and 3′ primers, respectively.

The amplified *his*F-N1, *his*F-N1N1, and *his*F-C1 genes were digested with NdeI and BamHI and ligated into the plasmid pET21a (Stratagene).

For the incorporation of the A128C exchange into *his*F-C1 by conventional PCR, pET21a-*his*F-C1 was used as template, and the oligonucleotides 5′-ATTCATATGCA-

GGCCGTTGTCGTGTGTATAGATGCA-3′ (new codon underlined) and T7-terminator (Stratagene) were used as 5′ and 3′ primer, respectively. Overlap extension PCR[41] was performed to introduce the Y143H mutation into *his*F-C1-A128C. The two megaprimers were generated in two separate PCRs, using the oligonucleotides 5′-TTCATGGTCTTCACCCATTCCGGAAAGAAGAAC-3′ and T7 promotor as 5′ primers, and the oligonucleotides T7 terminator and 5′-GTTCTTCTTTCCGGAATGGGTGAAGACCATGAA-3′ as 3′ primers (new codons underlined), respectively. The resulting amplification products were used in a third PCR, together with the gene flanking oligonucleotides T7 promoter and T7 terminator, to amplify *his*F-C1-A128C + Y143H. This gene was digested with NdeI and BamHI and ligated into the plasmid pET21a (Stratagene).

Heterologous gene expression and purification of recombinant proteins

For expression of the cloned genes, competent *E. coli* BL21(DE3) cells were transformed with the various pET21a constructs. Single colonies were used to inoculate 50 ml of LB medium containing ampicillin (150 μg/ml) and incubated at 37 °C overnight. This culture was used to inoculate 1 l of LB medium containing ampicillin (150 μg/ml), which was incubated at 37 °C until OD_{600} = 0.8 was reached. Expression was then induced by addition of 1 mM IPTG, and incubation was continued overnight. Cells were harvested by centrifugation (Sorvall RC-5B, GS3 rotor, 4000 rpm, 4 °C). The cells were suspended in 50 ml of 50 mM potassium phosphate (pH 7.5), lysed by sonification (Branson Sonifier W-250D; 2×1 min, 50% pulse, 0 °C), and centrifuged again (Sorvall RC-5C, SS34 rotor, 13,000 rpm, 15 min) to separate the soluble from the insoluble fraction of the extract.

All recombinant proteins were purified from the insoluble cell fraction. To this end, the proteins were solubilised by the addition of 6 M guanidinium chloride, which was then removed stepwise by dialysis against 50 mM potassium phosphate as described.[12] The refolded proteins were purified by preparative gel-filtration chromatography (HiLoad 26/60 Superdex 75 prep grade, column volume 320 ml; GE Healthcare), which was performed at 4 °C in 50 mM potassium phosphate and 300 mM KCl (pH 7.5) with an elution rate of 0.5 ml/min. Fractions containing pure recombinant protein (as judged by SDS-PAGE) were pooled and dialysed against 50 mM potassium phosphate (pH 7.5) over night. If required, the purified proteins were concentrated using an Amicon Centrifugal Device (molecular mass cut-off, 5 kDa), dropped into liquid nitrogen, and stored at −80 °C. Protein concentrations were determined by measuring the absorbance at 280 nm, using molar extinctions coefficients that were calculated from the amino acid sequence.[42]

The yields were 77 mg of HisF-N1, 42.5 mg of HisF-N1N1, 3.25 mg of HisF-C1, and 8.4 mg of HisF-C1-A128C + Y143H from 1 l of culture medium.

Analytical methods

Analytical gel-filtration chromatography was performed with a calibrated Superdex 75 column (volume, 100 ml; Amersham). Protein (0.03–0.17 mg) was applied on the column and eluted with a flow rate of 0.5 ml/min in 50 mM potassium phosphate and 300 mM potassium chloride (pH 7.5) at 25 °C. Apparent molecular masses were calculated from the corresponding elution volumes using a calibration curve that was obtained with standard proteins.

Sedimentation equilibrium runs were performed in a Beckman analytical ultracentrifuge (model E) at 24 °C and 16,000 rpm, and followed by measuring the absorbance at 277 nm. The concentration of HisF-N1, HisF-N1N1 and HisF-C1-A128C-Y143H was 0.20 mg/ml in 50 mM potassium phosphate buffer (pH 7.5). The runs were analysed with the meniscus-depletion method.[43] Molecular masses were calculated applying specific volumes of 0.75 ml/g.[44]

Far-UV CD spectra were recorded with a JASCO 815 CD spectrometer (d = 1 mm) at 25 °C in 50 mM potassium phosphate (pH 7.5), using protein concentrations of 0.2 mg/ml for HisF-N1, HisF-C1, and HisF-C1-A128C-Y143H, and 0.14 mg/ml for HisF-N1N1. The secondary-structure content was calculated from the spectra with the program ContinLL.[45–47]

Fluorescence emission spectra (excitation wavelength, 280 nm) were measured at 25 °C with a Cary Eclipse spectrophotometer (Varian) using 5 μM of protein dissolved in 50 mM potassium phosphate (pH 7.5). The generated emission of HisF-N1 and HisF-N1N1 is caused by the single Trp23, which was introduced by site-directed mutagenesis. The emission of HisF-C1 and HisF-C1-A128C + Y143H is caused by the single native Trp156 of HisF.

Unfolding of protein (0.20 mg/ml) induced by urea was followed in 50 mM potassium phosphate (pH 7.5) at 25 °C by the decrease of the fluorescence intensity at 320 nm after excitation at 280 nm, or by monitoring the decrease of the far-UV CD signal at 220 nm. The signals were measured after different time intervals until no further change was observed to ensure that equilibrium was reached. The obtained intensities were normalized and plotted as fractional change of the native signal. The midpoint of unfolding $D_{1/2}$ (molar), which represents the concentration of urea at which 50% of the protein has nonnative structure, served as an operational measure for conformational stability.

Chemical cross-linking of 0.2 mg HisF-N1 or HisF-C1-A128C + Y143H dissolved in 200 μl of 50 mM potassium phosphate (pH 7.5) was performed by incubation with 0.2% glutardialdehyde for 2 min. The reaction was stopped by the addition of 10 mM $NaBH_4$ and analysed by SDS-PAGE using a Tris–tricine gel (20% polyacrylamide).

Sequence comparison

Sequences were compared by means of a Smith–Waterman algorithm[48] (scores: BLOSUM 62, gap opening −10, gap extension −0.5). For each pairwise alignment of sequences S_i and S_k, the number of identical residues ident (S_i,S_k) was determined. A Mann–Whitney rank sum test[25] was used to compare the distributions of ident(S_i,S_k) values determined for sets of quarter-barrel sequences. To compare the similarity of ancient quarter-barrels and extant ones, a one-sample t test[49] was applied. In this case, the distributions were tested against a single ident(S_i,S_k) value resulting from the respective ancient barrels, which served as the expected mean value.

Selecting sequences for reconstruction

The starting point for the compilation of a multiple sequence alignment (MSA) of concatenated HisF and HisH sequences was the DSSP database (content as in autumn 2009)[21] entity related to the PDB entry 1GPW. Sequence stretches missing in HisF from *T. maritima* were

removed. Underrepresented taxa such as the genus *Thermotoga* were supplemented with sequences of closely related species by harvesting sequences with BLAST[50] and the National Center for Biotechnology Information (NCBI) database.[51] MAFFT[52] was used to create MSAs. Sequences branching at nodes incongruent with the "nearly universal trees" of life[22] were removed because it is difficult to distinguish between effects of stochastic noise, paralogous duplications, and true gene transfer events. The final MSA, which was termed MSA_{HisFH}, consisted of 87 concatenated sequences. For sequence reconstruction, the HisF part of MSA_{HisFH} was used. The resulting MSA MSA_{HisF} is given in Fig. S2, which also contains the four reconstructed sequences.

Reconstruction of ancient sequences

For sequence reconstruction and computation of phylogenetic trees, programs of the PhyloBayes 3.0 software suite[19] were utilized. MSA_{HisFH} was analysed under the time-homogeneous CAT model[19] launching eight independent MCMC samplings of length 50,000 to ensure convergence. A consensus tree was deduced from the concatenation of these eight chains. The maximum difference of posterior probabilities of tree bipartitions between any two chains was 0.12, indicating that all eight consensus topologies of the resulting trees were identical. The posterior number of biochemical profile categories was estimated to $CAT_{HisF}=74$.

The nonhomogeneous model CAT + BP[19] was utilized with a fixed number of $CAT_{HisF}=74$ categories of biochemical profiles. Four independent chains were run for 12,000 cycles. Ancestral sequences were deduced from the posterior distribution by means of the programs *ancestralseq* and *mapping* of the nh_PhyloBayes package.[19] We only utilized ancestral sequences related to tree nodes possessing a posterior probability >0.95.

Branch lengths corresponding to the respective subsequences of MSA_{HisF} were estimated under the CAT + BP model. Normalized branch lengths were obtained from the inferred stochastic mappings[19] and for subsets of sites corresponding to each quarter-barrel.

Acknowledgements

We thank Birte Höcker and Steffen Schmidt for insightful comments on the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft (STE 891/4-3).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.03.057

References

1. Lupas, A. N. & Koretke, K. K. (2008). *Evolution of protein folds. In* Computational Structural Biology—Methods and Applications (Schwede, T. & Peitsch, M. C., eds), World Scientific, Hackensack, NJ.
2. Söding, J. & Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, **25**, 837–846.
3. Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E. & Regan, L. (2005). A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr. Opin. Struct. Biol.* **15**, 464–471.
4. Caetano-Anolles, G., Kim, H. S. & Mittenthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl Acad. Sci. USA*, **104**, 9358–9363.
5. Sterner, R. & Höcker, B. (2005). Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055.
6. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198.
7. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.
8. Thoma, R., Schwander, M., Liebl, W., Kirschner, K. & Sterner, R. (1998). A histidine gene cluster of the hyperthermophile *Thermotoga maritima*: sequence analysis and evolutionary significance. *Extremophiles*, **2**, 379–389.
9. Höcker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner, R. (2001). Dissection of a $(\beta\alpha)_8$-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36.
10. Höcker, B., Claren, J. & Sterner, R. (2004). Mimicking enzyme evolution by generating new $(\beta\alpha)_8$-barrels from $(\beta\alpha)_4$-half-barrels. *Proc. Natl Acad. Sci. USA*, **101**, 16448–16453.
11. Höcker, B., Lochner, A., Seitz, T., Claren, J. & Sterner, R. (2009). High-resolution crystal structure of an artificial $(\beta\alpha)_8$-barrel protein designed from identical half-barrels. *Biochemistry*, **48**, 1145–1147.
12. Seitz, T., Bocola, M., Claren, J. & Sterner, R. (2007). Stabilisation of a $(\beta\alpha)_8$-barrel protein designed from identical half barrels. *J. Mol. Biol.* **372**, 114–129.
13. Claren, J., Malisi, C., Höcker, B. & Sterner, R. (2009). Establishing wild-type levels of catalytic activity on natural and artificial $(\beta\alpha)_8$-barrel protein scaffolds. *Proc. Natl Acad. Sci. USA*, **106**, 3704–3709.
14. Gerlt, J. A. & Babbitt, P. C. (2001). Barrels in pieces? *Nat. Struct. Biol.* **8**, 5–7.
15. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765.
16. Richter, M. (2008). Studies on the evolution of the $(\beta\alpha)_8$-barrel fold from $(\beta\alpha)_2$ modules. PhD thesis, University of Regensburg.
17. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
18. Söding, J., Remmert, M. & Biegert, A. (2006). HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res.* **34**, W137–W142.
19. Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N. & Gouy, M. (2008). Parallel adaptations to high temperatures in the Archaean eon. *Nature*, **456**, 942–945.
20. Fondi, M., Emiliani, G., Lio, P., Gribaldo, S. & Fani, R. (2009). The evolution of histidine biosynthesis in archaea: insights into the *his* genes structure and organization in LUCA. *J. Mol. Evol.* **69**, 512–526.

21. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
22. Puigbo, P., Wolf, Y. I. & Koonin, E. V. (2009). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59.
23. Beismann-Driemeyer, S. & Sterner, R. (2001). Imidazole glycerol phosphate synthase from *Thermotoga maritima*. Quaternary structure, steady-state kinetics, and reaction mechanism of the bienzyme complex. *J. Biol. Chem.* **276**, 20387–20396.
24. Klem, T. J. & Davisson, V. J. (1993). Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. *Biochemistry*, **32**, 5177–5186.
25. Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60.
26. Buck, M. (1998). Trifluoroethanol and colleagues: cosolvents come of age. Recent studies with peptides and proteins. *Q. Rev. Biophys.* **31**, 297–355.
27. Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A. & Matthews, C. R. (1999). Identifying the structural boundaries of independent folding domains in the α subunit of tryptophan synthase, a β/α barrel protein. *Protein Sci.* **8**, 1200–1209.
28. Yadid, I. & Tawfik, D. S. (2007). Reconstruction of functional β-propeller lectins via homo-oligomeric assembly of shorter fragments. *J. Mol. Biol.* **365**, 10–17.
29. Zhang, B. & Peng, Z. (2000). A minimum folding unit in the ankyrin repeat protein p16(INK4). *J. Mol. Biol.* **299**, 1121–1132.
30. Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558.
31. Linke, K. & Jakob, U. (2003). Not every disulfide lasts forever: disulfide bond formation as a redox switch. *Antioxid. Redox Signal.* **5**, 425–434.
32. Raines, R. T. (1997). Nature's transitory covalent bond. *Nat. Struct. Biol.* **4**, 424–427.
33. Beeby, M., O'Connor, B. D., Ryttersgaard, C., Boutz, D. R., Perry, L. J. & Yeates, T. O. (2005). The genomics of disulfide bonding and protein stabilization in thermophiles. *PLoS Biol.* **3**, e309.
34. Mallick, P., Boutz, D. R., Eisenberg, D. & Yeates, T. O. (2002). Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc. Natl Acad. Sci. USA*, **99**, 9679–9684.
35. Schwab, T., Skegro, D., Mayans, O. & Sterner, R. (2008). A rationally designed monomeric variant of anthranilate phosphoribosyltransferase from *Sulfolobus solfataricus* is as active as the dimeric wild-type enzyme but less thermostable. *J. Mol. Biol.* **376**, 506–516.
36. Thoma, R., Hennig, M., Sterner, R. & Kirschner, K. (2000). Structure and function of mutationally generated monomers of dimeric phosphoribosylanthranilate isomerase from *Thermotoga maritima*. *Struct. Fold. Des.* **8**, 265–276.
37. Walden, H., Bell, G. S., Russell, R. J., Siebers, B., Hensel, R. & Taylor, G. L. (2001). Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase. *J. Mol. Biol.* **306**, 745–757.
38. Sterner, R. & Liebl, W. (2001). Thermophilic adaptation of proteins. *Crit. Rev. Biochem. Mol. Biol.* **36**, 39–106.
39. Vieille, C. & Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **65**, 1–43.
40. Fani, R., Lio, P. & Lazcano, A. (1995). Molecular evolution of the histidine biosynthetic pathway. *J. Mol. Evol.* **41**, 760–774.
41. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. & Pease, L. R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, **77**, 51–59.
42. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423.
43. Yphantis, D. A. (1964). Equilibrium ultracentrifugation of dilute solutions. *Biochemistry*, **3**, 297–317.
44. Cohn, E. & Edsall, J. (1943). *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions.* Reinhold, New York.
45. van Stokkum, I. H., Spoelder, H. J., Bloemendal, M., van Grondelle, R. & Groen, F. C. (1990). Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal. Biochem.* **191**, 110–118.
46. Whitmore, L. & Wallace, B. A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.* **32**, W668–W673.
47. Whitmore, L. & Wallace, B. A. (2008). Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, **89**, 392–400.
48. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
49. Student (1908). The probable error of a mean. *Biometrika*, **6**, 1–25.
50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
51. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58.
52. Katoh, K., Kuma, K., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518.