Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

De novo Backbone and Sequence Design of an Idealized α/β-barrel Protein: Evidence of Stable Tertiary Structure

F. Offredi[1], F. Dubail[1], P. Kischel[1], K. Sarinski[2], A. S. Stern[3] C. Van de Weerdt[1], J. C. Hoch[3], C. Prosperi[4], J. M. François[1] S. L. Mayo[2] and J. A. Martial[1]
[1]*Laboratoire de Biologie Moléculaire et Génie Génétique Université de Liège, B6, Sart Tilman, Belgium*
[2]*Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA*
[3]*The Rowland Institute for Science, Cambridge, MA 02142 USA*
[4]*Département de Physique Université de Liège, B5, Sart Tilman, Belgium*

Abstract

We have designed, synthesized, and characterized a 216 amino acid residue sequence encoding a putative idealized α/β-barrel protein. The design was elaborated in two steps. First, the idealized backbone was defined with geometric parameters representing our target fold: a central eight parallel-stranded β-sheet surrounded by eight parallel α-helices, connected together with short structural turns on both sides of the barrel. An automated sequence selection algorithm, based on the dead-end elimination theorem, was used to find the optimal amino acid sequence fitting the target structure. A synthetic gene coding for the designed sequence was constructed and the recombinant artificial protein was expressed in bacteria, purified and characterized. Far-UV CD spectra with prominent bands at 222 nm and 208 nm revealed the presence of α-helix secondary structures (50%) in fairly good agreement with the model. A pronounced absorption band in the near-UV CD region, arising from immobilized aromatic side-chains, showed that the artificial protein is folded in solution. Chemical unfolding monitored by tryptophan fluorescence revealed a conformational stability (ΔG_{H20}) of 35kJ/mol. Thermal unfolding monitored by near-UV CD revealed a cooperative transition with an apparent T_m of 65 °C. Moreover, the artificial protein did not exhibit any affinity for the hydrophobic fluorescent probe l-anilinonaphthalene-8-sulfonic acid (ANS), providing additional evidence that the artificial barrel is not in the molten globule state, contrary to previously designed artificial α/β-barrels. Finally, ^{1}H NMR spectra of the folded and unfolded proteins provided evidence for specific interactions in the folded protein. Taken together, the results indicate that the *de novo* designed α/β-barrel protein adopts a stable three-dimensional structure in solution. These encouraging results show that *de novo* design of an idealized protein structure of more than 200 amino acid residues is now possible, from construction of a particular backbone conformation to determination of an amino acid sequence with an automated sequence selection algorithm.

Keywords*:* protein design; backbone parameterization; side-chain modeling; fluorescence; circular dichroism

Abbreviations used: DLS, dynamic light-scattering; Tm, thermal denaturation temperature; AMFP, atomic mean force potential; ORBIT, optimization of rotamers by iterative techniques; DEE, dead-end elimination; GMEC, global minimum energy conformation; PDA, protein design automation; ANS, 1-anilinonaphthalene-8-sulfonic acid; wt, wild-type; IGPS, indole-3-gycerolphosphate synthase; PRAI, phosphorybosylanthranilate isomerase; TIM, triosephosphate isomerase

Introduction

De novo protein design (also referred to as the inverse protein folding problem) is an attractive way to assess current knowledge of the relationships between a protein amino acid sequence and the three-dimensional structure that the polypeptide chain finally adopts after folding in solution. Recent experimental results suggest that protein folding rates and mechanisms are determined largely by native-state topology rather than specific interactions.[1,2] Analysis by molecular dynamics simulation of folding behavior of two peptides having a sequence identity of 15% has shown that the native topology determines, to a large extent, the free energy surface. Folding happens along multiple pathways with a statistical weight that depends on the sequence. The amino acid sequence, and thus the specific interactions between different side-chains determine the most probable folding route.[3] These results support the suggestion that the first effort in protein design should focus on optimizing the stability of a particular topology rather than explicitly designing for kinetic accessibility[4] In any case, the latter task remains very difficult. Because proteins are very complex macromolecules involving thousands of atoms, calculating the stability of all possible conformations is inherently difficult and time-

Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

consuming. Yet it is possible to divide the conformational space of a protein into two distinct conformational spaces of similar complexity[5]: one associated with the backbone conformation (or main chain) and defining the target structure (the fold or topology) and one associated with side-chain conformation. In *de novo* design, it is postulated that these two conformational spaces can be treated separately, so that the complete designing process is reduced to two major steps.[6,7] First, the protein topology, or target structure, must be defined and optimized. This step is backbone selection or backbone design. Then the lowest-energy sequence fitting and stabilizing the defined tertiary structure must be found. This step is called side-chain or sequence design. Ponder & Richards were the first to explore the relationship between these two conformational spaces, using a fixed natural backbone as target structure.[8] They used the concept of rotamers, defined as statistically significant amino-acid side-chain conformations,[9] to represent side-chain flexibility. The side-chain rotamers were recorded in a rotamer library and simple exclusion volume criteria were used to enumerate the allowed sequences for a given template. More recently, the inverse protein folding methodology has been revisited by several groups.[10-18] Mayo and co-workers established a protein design automation (PDA) cycle where artificial peptides were analyzed systematically in order to improve the design method.[12] This strategy shows great promise for solving the second-step problem in protein design, by providing experimental feedback for improving the potential energy function and design methods.[10,11,14,19,20] The combinatorial problem of optimizing side-chain conformations is solved by means of a search algorithm based on the dead-end elimination (DEE) theorem.[21,22] This heuristic method finds the most favorable combination of side-chain rotamers in their optimal conformation for the target structure. Sequences are ranked with an appropriate potential energy function depending on the location of each side-chain in the protein: core, surface, or boundary position.[20,23] This recent progress in sequence selection algorithms and the robustness of such algorithms towards backbone perturbations suggest that they can be used to design sequences for entirely novel and/or idealized backbones.[7,24,25] Most efforts have focused on fixed protein backbones of natural proteins as target conformations for the first design step.[6,10] Different approaches to *de novo* protein design have emerged. One is to take the backbone of a known protein and create a new protein core[14,18,26]; another is to start with a natural backbone but to place new side-chains at all positions.[10] One team successfully created a totally artificial protein by designing both the backbone conformation and the sequence.[15,16] Yet the large majority of sequence designs have been based on fixed high-resolution backbones of relatively small natural proteins. Only an exceptional few have taken backbone flexibility into account.[13,15,16,25] A small, idealized, 38 residue helical peptide[27] and a larger, 73 residue three-helix bundle protein structure[28] have been created with success, but designing larger idealized proteins seems more challenging. Some effort has been put into the *de novo* design of larger proteins with a repetitive structure, mainly the parallel $(\alpha/\beta)_8$-barrel (TTM-barrel). To date, none of the attempts to design a TIM-barrel structure *de novo* has yielded a structure with all the properties of natural native proteins.[29-32] In the best cases, the resulting protein behaves like a molten globule, and no structure has yet been solved.[33,34] In the present work, we have designed an idealized, 216 residue α/β-barrel backbone, using geometric parameters such as distances and angles between secondary structures to describe the barrel's topology characterized by 4-fold symmetry. The relative positioning of secondary structures was done with the help of short, conserved structural motifs.[35] The idealized target fold was used to find, by means of an automated sequence selection algorithm, the best rotameric sequence stabilizing this structure.[10,24] We report here the first biophysical characterizations of this protein designed from scratch and demonstrate a stable tertiary structure. Our results support the idea that in order to obtain a parallel $(\alpha/\beta)_8$-barrel, it is crucial to take side-chain packing specificity into account in an "idealized" backbone conformation defined from first principles.

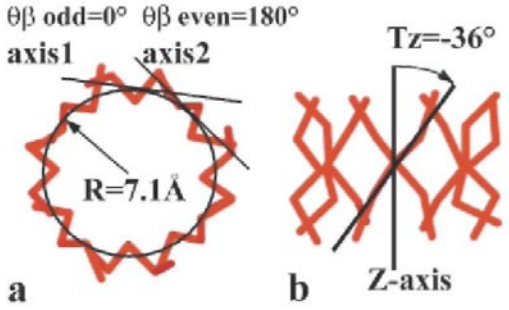
Results

Protein design strategy

In this study, we designed an α/β-barrel (the TIM-barrel topology) from scratch in two steps. In the first step, we used simple parameters to define an idealized artificial backbone representing the α/β-barrel topology. We used five geometric parameters to model a β-sheet with 4-fold symmetry (Figure 1). The system was subjected to 300 steps of gradient conjugate energy minimization. The eight α-helix barrel surrounding the central β-sheet was built with five additional geometric parameters (Figure 2). Some of these parameters were adjusted with the help of short structural motifs called αβ1 and αβ3 turns (see Materials and Methods). The geometric parameter values and definitions are summarized in Table 1. The hierarchical process used to build the idealized backbone and the lowest-energy sequence compatible with the idealized backbone are shown in Figure 3. In the second step, the side-chain sequence and conformations were chosen with the help of an automated selection algorithm.[10]
The residues were classified as occupying core, surface, or boundary positions according to the distances of their C^{α} and C^{β} atoms with respect to a surface calculated from the C^{α} atom positions in the backbone. We designed

each of the corresponding regions separately, using different rotamer libraries and different potential energy functions to score the sequences. The final sequence is shown in Figure 4 with the secondary structures and the energy profile.

Figure 1. *Parameters used to build the β-sheet scaffold. (a) Top view of the β-sheet scaffold (i.e. the C^{α} trace). The scaffold consists of eight strands. R is the radius of the barrel. $\theta\beta_{odd}$ and $\theta\beta_{even}$ are the rotation angles of odd and even strands about axis 1 and axis 2, respectively. (b) Side-view of the β-sheet scaffold. Tz is the angle between the z-axis and the β-strand axis.*



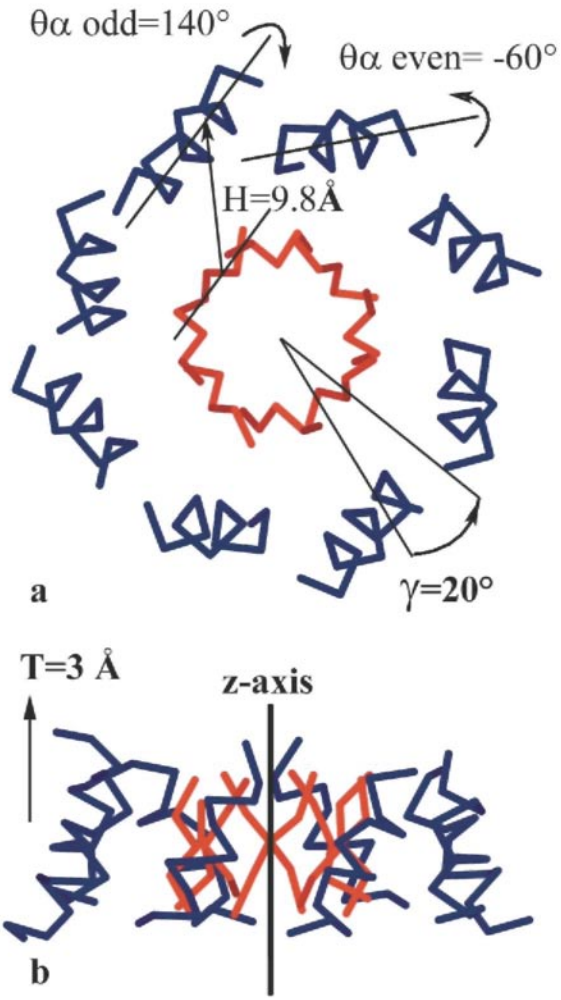
Protein model assessment

The model of the artificial α/β-barrel was assessed with the atomic non-local environment assessment (ANOLEA) server.[36,37] The method is based on a statistical atomic mean force potential (AMFP) involving short-range and non-local interactions between heavy atoms of the structure to be evaluated. The ANOLEA program is able to assess the global quality of a protein's 3D structure, to observe local error in a structure, and to give information on topology on the basis of the energy profile. The energy profile of a natural α/β-barrel was determined. We chose indole-3-glycerolpho-sphate synthase (PDB code lIGS) as a natural pro-energy profile of the artificial barrel displays the same pattern typical of an α/β-barrel. The energy profile of the artificial protein was divided into four topologically equivalent sub-units (1-54, 55-108, 109-162, 163-216) and was superposed in order to identify high-energy zones (Figure 4). It appears that these high-energy zones are located predominantly in βα loops and involve surface positions. tein representative of the topology of the α/β-barrel. The energy profile of a natural α/β-barrel displays a characteristic curve (Figure 5) with eight deep minima corresponding to the eight β-strands of the barrel.

Table 1. *Summary of the parameters used to describe the idealized α/β barrel backbone*

Parameter	Description	Value
R(Å)	Equatorial radius	7.1
N	Number of strands	8
Tz (deg.)	Angle between strand axis and z axis (i.e. barrel axis)	-36
$\theta\beta_{odd}$ (deg.)	Rotation angle of odd strand on its major axis	0
θβ?$_{even}$ (deg.)	Rotation angle of even strand on its major axis	180
H (Å)	Radius offset, distance between helix axis and sheet plane	9.8
y (deg.)	Helix angle offset, rotation of helix barrel about z axis	20
T(Å)	Helix axis shift, translation of helix barrel about z axis	3
$\theta\alpha_{odd}$ (deg.)	Rotation angle of odd helix on its major axis	140
$\theta\alpha_{even}$ (deg.)	Rotation angle of even helix on its major axis	-60

Figure 2. *Description of the parameters used to construct the α-helix scaffold around the β-sheet barrel. (a) $\theta\alpha_{odd}$ and $\theta\alpha_{even}$ are the rotation angles of odd and even helices about their axes. The offset radius H is the distance between the α-helix axis and the β-sheet plane at the equatorial plane. Angle y is the offset angle defining rotation of the α-helix scaffold about the β-sheet barrel, or the shear between the β-strand barrel and the α-helix barrel. A zero value for offset angle y means that the helix axis and the strand axis are colinear with the barrel center. (b) T is the helix axis shift defining the relative displacement of the α-helix scaffold relative to the β-sheet barrel.*



Gene synthesis, protein production, and purification,

We used a fast method called recursive PCR[38] to synthesize the gene encoding the designed amino acid sequence. The gene was cloned into the pET-11d expression vector and sequenced completely. The protein was overproduced as inclusion bodies, dissolved in urea, and bound to an ion-exchange column. The bound protein was refolded and eluted from the column, then purified by size-exclusion chromatography to eliminate the aggregated form of the protein (see Materials and Methods).

Biophysical characterization

We first used dynamic light-scattering (DLS) to show that at concentrations less than 1 mg/ml, the artificial protein was not aggregated. We observed a monodispersed particle size distribution centered around 10 nm. This low concentration of protein was sufficient for a first biophysical characterization. CD in the far UV (Figure 6) revealed that the protein in solution has a high percentage of secondary structures. The determined α-helix content[39] was 50%, in good agreement with the protein model. Near-UV CD spectra (Figure 7) showed pronounced absorption bands in the aromatic region, particularly in the absorption region of phenylalanine, with prominent positive vibronic bands 6 nm apart.[40] These results suggest that the single tryptophan, five tyrosine, and 17 phenylalanine residues are immobilized in a rigid tertiary structure. Thermal unfolding of the protein was

monitored by recording the CD signal at 265 nm (Figure 8) rather than in the far-UV region as is traditionally the case. The reason is that the intensity difference after thermal unfolding was small, even at 208 nm or 222 nm. Moreover, the general shape of the far-UV CD spectrum remained unchanged as the temperature was increased. At 265 nm the signal was stronger and it disappeared completely after unfolding. At this wavelength we observed a single-step unfolding curve. Thermal melting was cooperative with an inflexion point at 65 °C. We could not derive thermodynamic parameters from the experiment because thermal unfolding was not reversible: the dissolved protein precipitated irreversibly beyond 85 °C. To estimate the stability of the protein, we monitored equilibrium chemical unfolding by measuring tryptophan average emission wavelength fluorescence and far-UV CD at 222 nm. Guanidine hydrochloride (GdnHCl) was used to unfold the protein because urea did not shift the tryptophan fluorescence emission peak, and thus appeared unable to unfold the protein. The C_m values derived from fluorescence and far-UV CD are 3.9 M and 3.97M, respectively (Figure 9). Chemical unfolding was reversible and we determined a free energy of unfolding of 35 (± 3)kJ/mol when the reaction was monitored by fluorescence. We estimated the free energy of unfolding in the presence of 850 mM of NaCl (data not shown) by monitoring the emission of fluorescence. The data were fit to a two-state model and we calculated a ΔG_{H2O} of 19.2 kJ/mol. When monitoring CD at 222nm, in the same sample, we observed a sudden loss in ellipticity at very low concentrations in denaturant, up to 0.2 M, followed by a gain of ellipticity up to 1 M in GdnHCl. This observation suggests a rearrangement of helices at low concentration of denaturant, and thus shows the presence of an intermediate not observed by fluorescence. The best fit was estimated in this case by excluding the values ranging from 0 M to 1 M in GdnHCl, resulting in a free energy of unfolding of 20 (± 2.5) kj/mol. The use of a two-state model probably underestimates the free energy of unfolding. Our next aim was to estimate the packing quality of the protein's hydrophobic core. In other words, we wanted to check whether the protein was in the molten-globule state as is often the case for artificial proteins, especially designed α/β-barrels.[32,33] The molten globule is a stable intermediate stabilized by mild denaturing conditions. This intermediate has a secondary structure and is highly compact but its side-chains are not tightly packed.[41] It can be detected by means of a hydrophobic probe emitting fluorescence upon binding. The absence of ANS binding, and thus tight packing of side-chains in the protein core, has become a criterion of success in protein *de novo* design.[42] In our experiment (Figure 10), the probe did not bind to the artificial protein after refolding and purification on an ion-exchange column. If a chaotropic agent such as GdnHCl was present at low concentration, we observed the appearance of a strong ANS fluorescence signal suggestive of a loosely packed structure. When the agent's concentration was increased to 2 M, the signal disappeared. A similar pattern has been observed with other proteins. The unfolded and folded forms of the artificial protein could be distinguished by means of ^{1}H NMR measurements (Figure 11(a) and (b)). In proteins, there are many hydrogen nuclei in slightly different environments and with different chemical shifts, resulting in many overlapping proton nuclear magnetic resonances. The slow motions (long correlation times, τ_c) associated with proteins lead to relatively broad lines. The result is a spectrum with a broad envelope, even when high magnetic fields are used. However, the NMR spectrum of the native form is influenced by the tertiary structure of the protein. In a folded protein, freedom of motion is restricted. The hydrogen nuclei are in different environments, so that there is a wide range of chemical shifts and poor resolution of the resonance lines. In an unfolded protein, there is (supposedly free) segmental rotation of side-chains. Hydrogen nuclei that were not equivalent because of slightly different environments become equivalent, and the resonance lines are sharper. Hence, sharp lines are indicative of the absence of a tertiary structure (see Figure 11(b)). In Figure 11(a), the least shielded protons (>8ppm) are those of histidine (C2), and of the indole group of tryptophan. The aromatic region including the hydrogen atoms of tryptophan, tyrosine, phenylalanine, and the C4 hydrogen atom of histidine appears at $\delta = 7$ ppm. Figure 11(b) shows a few sharp lines (6 ppm and 7.3 ppm) in this area. The C^α hydrogen atoms appear at higher field values and generally display a broad envelope that may be obscured partially by the solvent peak. This is visible in Figure 11(a), but unfortunately a few wide NMR lines due to the solvent and purification column can be seen. The methyl region usually appears in the region corresponding to the highest field values ($\delta = 1$ ppm) and corresponds to the methyl groups of the aliphatic side-chains of valine, leucine, and iso-leucine (Figure 11(a)). Our ^{1}H NMR data confirm the presence of a tertiary structure in the *de novo* designed protein.

Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

Figure 3. *A drawing of the de novo-designed idealized backbone (A, B, C) and side-chain ensemble (D). The β-sheet barrel (red) was built first with idealized geometric parameters (A), then the surrounding α-helix barrel (blue) was constructed around the β-sheet barrel with appropriate parameters. Short structural motifs (C, gray) αβl, αβ3 and βABα were used to connect both barrels. A very fast and fully automated selection algorithm based on the dead-end elimination theorem was used to find the best rotameric sequence (D, the lowest-energy sequence, represented in green) compatible with the main chain defined above as a scaffold.*

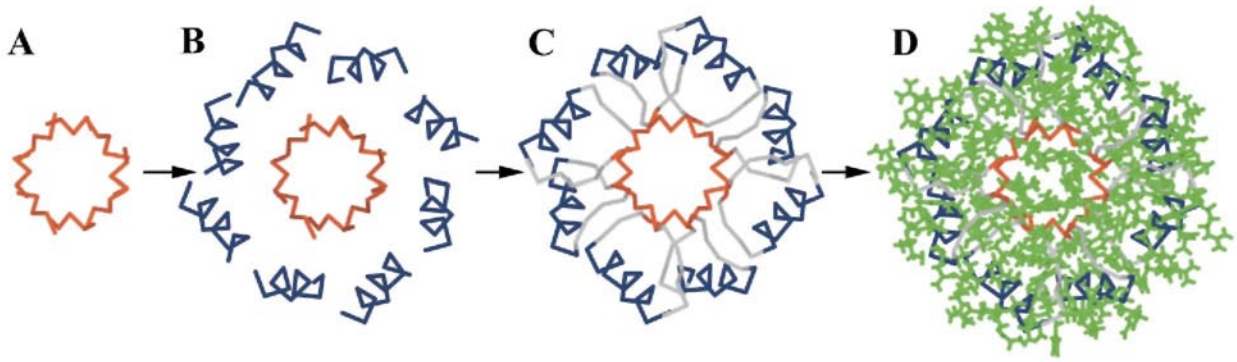
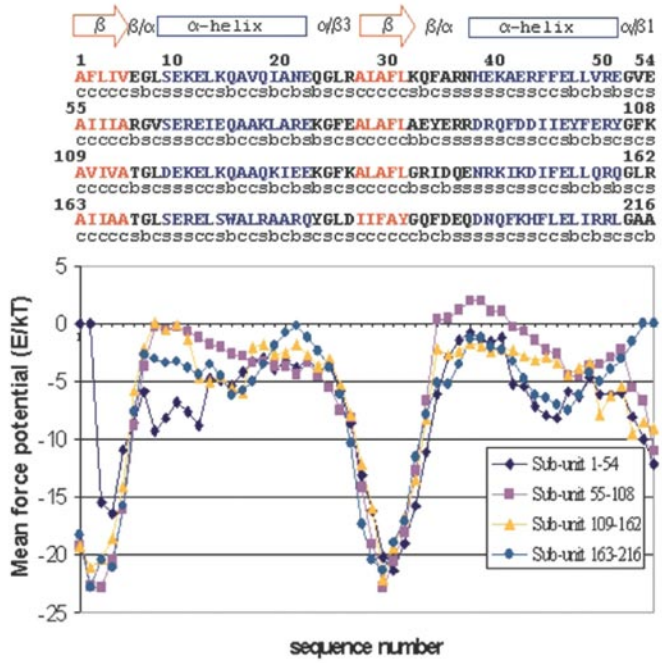


Figure 4. *Amino acid sequence, secondary structure, solvent-accessibility of residues, and energy profile of the artificial protein. The designed sequence is displayed as an alignment of four subunits (1-54, 55-108,109-162 and 163-216). Each subunit is composed of two β-strands (in red), two α-helices (in blue), two β/α loops, one αβl loop, and one αβ3 loop. The solvent-accessibility of each residue computed by the resclass subroutine of the ORBIT program is indicated under the amino acid sequence (c, core residue; b, boundary residue; and s, solvent-exposed residue). The energy profile of the modeled structure is shown. The eight deep minima corresponding to the β-strand region of the structure are typical of the $(\alpha/\beta)_8$-barrel topology. High-energy regions are located in helices and loops and are solvent-exposed in the model.*



Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

Figure 5. *Energy profiles computed with ANOLEA. The blue curve is the energy profile of indole-3-glycerolphosphate synthase (IIGS) chosen as a natural-protein reference to assess the overall quality of the artificial protein model. The energy profile of the artificial protein is colored red. The designed protein displays the typical signature of the α/β-barrel topology: eight deep, narrow minima corresponding to the eight β-strands in the protein core. All along the sequence, the energy appears slightly higher for the artificial structure than for the natural one.*

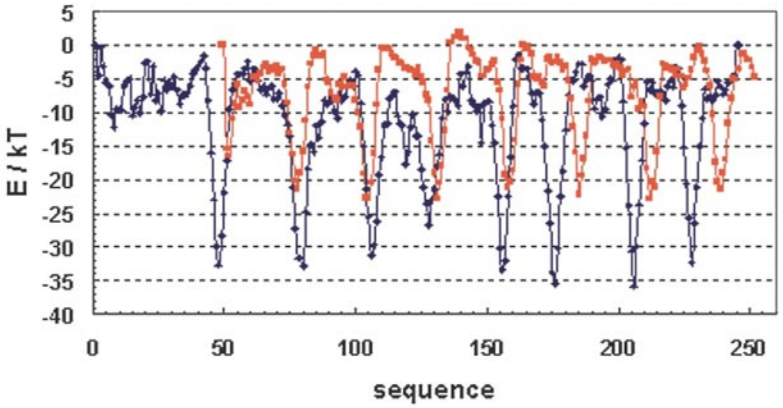


Figure 6. *Far UV-CD spectra showing the high α-helix content of the artificial protein, estimated at 53% from the ellipticity at 208 nm, in good agreement with the model. The protein was prepared in 50 mM borate, 850 mM NaCl at pH 8.5, its concentration being 10 µM. The spectrum was recorded at 25 °C.*

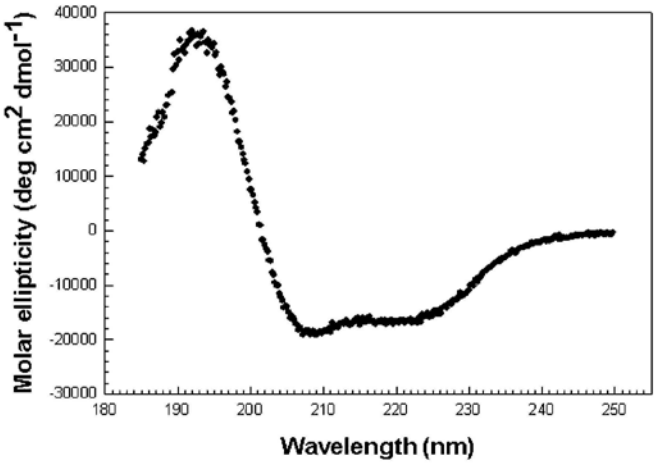
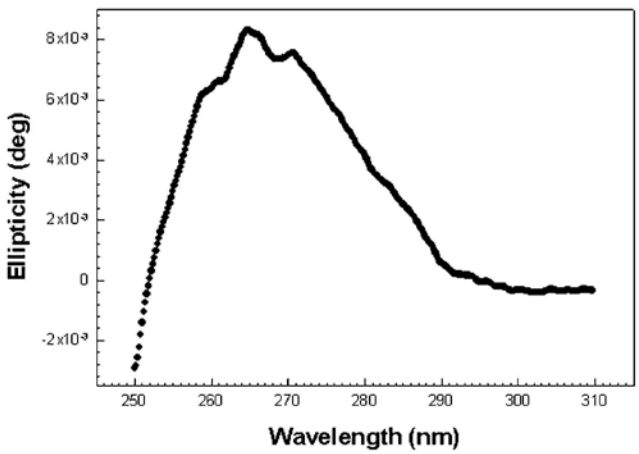


Figure 7. *Near UV-CD spectra showing a strong positive signal arising from the immobilized aromatic side-chains attesting the presence of a tertiary structure. The protein concentration was 0.4 mg/ml (16 µM) prepared in 50 mM borate, 850 mM NaCl at pH 8.5 and the spectrum was recorded at 25 °C.*



Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

Figure 8. *Equilibrium thermal unfolding reaction monitored by CD at 265 nm.*

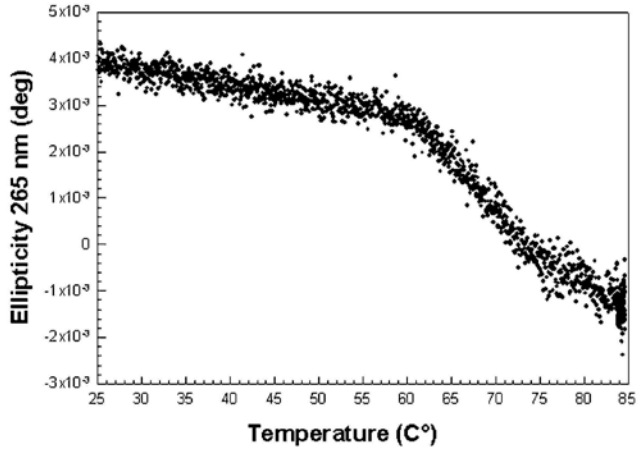
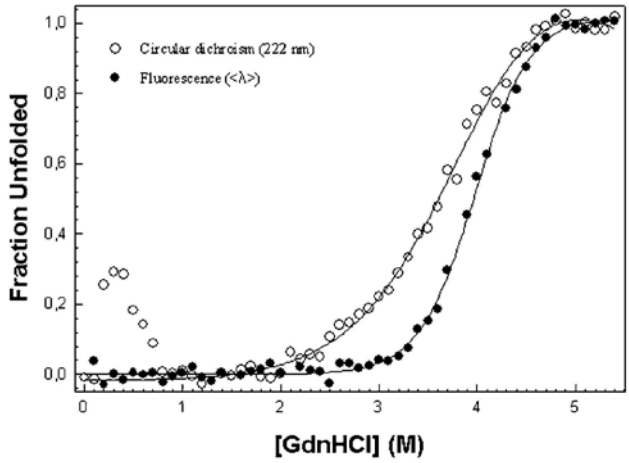
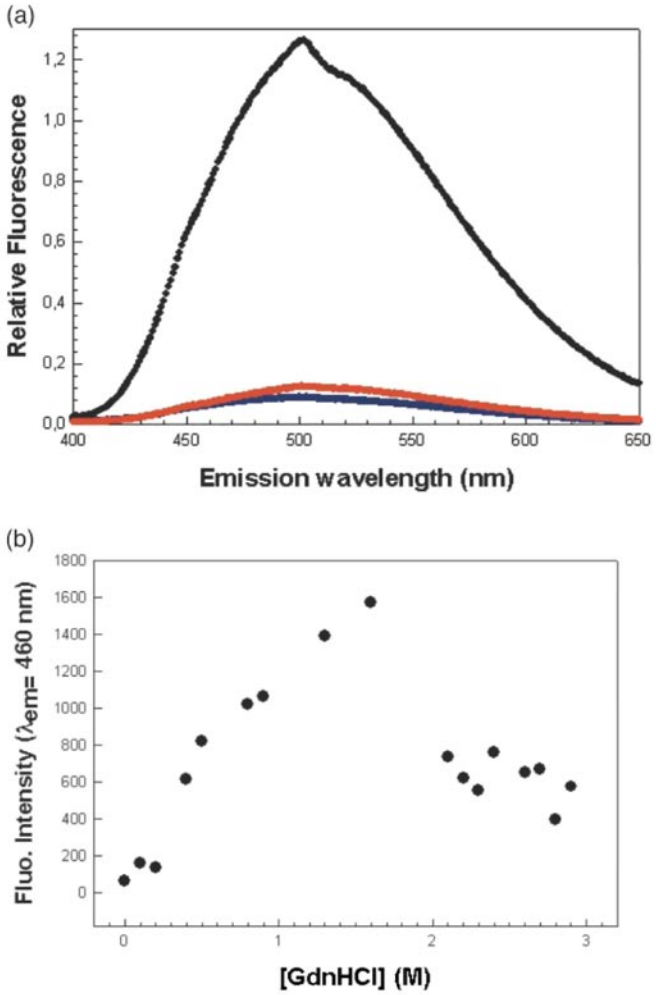


Figure 9. *Guanidine hydrochloride equilibrium unfolding reaction monitored by the tryptophan emission fluorescence (average emission wavelength, filled circles) and far UV-CD (222 nm, open circles).* $C_{mU\text{-}N}$ *= 3.97 and 3.9 for unfolding curve monitored by fluorescence and CD, respectively. The* ΔG_{H2O} *of unfolding was estimated to 35 kj/mol by fitting the data to a two-state model when the unfolding is monitored by fluorescence.*



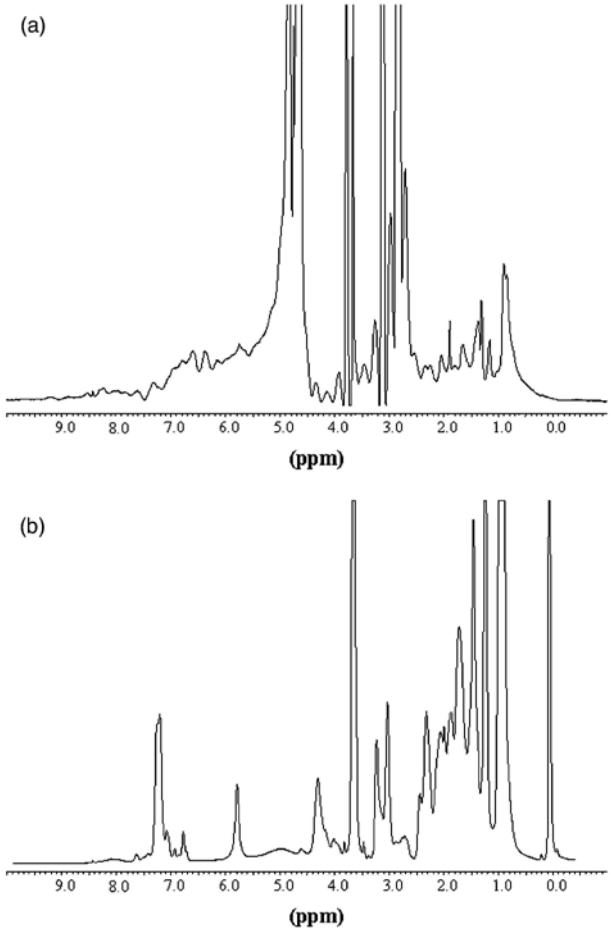
Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

***Figure 10**. (a) Fluorescence emission spectra of ANS with the protein in 0 M GdnHCl (blue), 1M GdnHCl (black), and 2M GdnHCl (red) in the presence of 850 mM NaCl. (b) Variation of fluorescence intensity of ANS as a function of the increasing concentration of the denaturant. The maximum of intensity is obtain around 1.5 M and reflects the exposure of hydrophobic surface to solvent.*



Published in: Journal of Molecular Biology (2003), vol. 325, iss.1, pp. 163-174
Status: Postprint (Author's version)

Figure 11. *(a) ^{1}H NMR spectra of the folded artificial barrel. The protein concentration is 12 mg/ml in 0.1 M NaCl, 50 mM Hepes (pH 7). (b) ^{1}H NMR spectra of the unfolded artificial barrel. Absorption bands at 3 ppm and 6 ppm correspond to impurities. The unfolded condition was obtained after a quick dialysis of the protein denatured in 6 M GdnHCl against 50 mM Tris-HCl (pH 8.5). ^{1}H NMR spectra were recorded with a Bruker DRX 400 MHz spectrometer. All assays were performed at 25 °C. All programs were from the Bruker library. Proton NMR spectra with water presaturation were obtained with a spectral width of 4 kHz for 16 K frequency and time-domain data points.*



Discussion

The present attempt to design an idealized artificial α/β-barrel protein from scratch has yielded a protein that is stable, non-aggregated, and well packed in solution. The protein displays an α-helix content close to that of the model (50% as determined by far-UV CD), and several lines of evidence suggest that it also has a tertiary structure. Firstly, the near-UV CD spectrum shows a strong positive signal, indicating that the aromatic side-chains are buried in a rigid, asymmetric environment characteristic of a fully folded protein. The NMR data support this view. Furthermore, unlike many designed proteins,[6] ours does not behave like a molten globule in ANS binding experiments. We have thus determined that the designed protein has native-like properties rather than molten-globule-like properties. Furthermore, we have monitored the protein's thermal and chemical unfolding at equilibrium. From the melting curve, we deduced an apparent T_m of 65 °C, but were unable to determine a free energy of thermal unfolding because the protein precipitated irreversibly when the temperature exceeded 85 °C. The artificial protein is more than marginally stable and sufficiently soluble for biophysical characterization. Our data support the view that such efforts to optimize side-chain packing may be crucial for designing an idealized α/β-barrel topology from scratch. This preliminary biophysical characterization is not sufficient to prove that we have succeeded in creating the first idealized α/β-barrel, but it does suggest that we have made significant progress in the right direction. The major obstacle to determining the protein's structure is its relatively low solubility, as both crystallization and 2D NMR require high solubility. We have postulated that the two conformational spaces (backbone conformation and side-chains conformation) can be treated

independently in the design process. In our approach, topology is considered first: an idealized backbone is designed, starting with simple geometric parameters describing relative angles and distances between regular secondary structures. Then we seek a set of side-chains, using a sequence selection algorithm to find the optimal amino-acid sequence and side-chain conformation fitting the target backbone. To reduce the complexity of the combinatorial problem, we treated core, boundary, and surface residues separately. Special care was taken to avoid placing too many voluminous residues (tryptophan, phenylalanine) or small residues (alanine) in the protein core. This was done by performing three successive rounds of rotamer placement in the core, with an energy minimization step (resulting in a slight relaxation of the backbone) after each of the first two rounds. Side-chain placement at boundary and surface positions was done without backbone perturbation. Computational procedures for selecting protein sequences from fixed natural targets yield sequences very similar to *wt* sequences.[14,17,19,25] This is especially true for buried positions, where packing constraints are highest. Biophysically, our protein core appears to be well designed, possessing the tight side-chain packing characteristic of native-like structures. On the other hand, computational procedures for surface design often yield sequences remarkably different from *wt* sequences. This may be due to the fact that the protein surface takes part in interactions with other molecules and could mean that functional properties should be taken into account in design procedures.[17] Previous computational protein design results suggest that helix propensity is a key factor in sequence design for surface helix positions.[43,44] In other words, the present sequence was not optimized to avoid self-aggregation. We hope to solve the solubility and stability problem by directed evolution of surface positions, using a folding reporter protein[45] to select the more soluble variants from a library and testing them for improved stability. Our results are in good agreement with the critical assessment of the model by ANOLEA. The high-energy zones on the energy profile are indeed located at surface positions. A combined rational and combinatorial protein engineering approach has been used successfully to redesign the enzyme IGPS to display PRAI-like catalytic activity.[46] We aim to combine rational *in silico* design and *in vitro* surface-directed evolution to improve the solubility of our artificial protein.

Materials and Methods

Backbone design

An idealized backbone representing the TIM-barrel topology (a central β-sheet of eight parallel strands surrounded by eight parallel α-helices) was built, taking into account several geometric parameters reflecting the relative positions of secondary structures.[47] Figure 1 illustrates the geometric parameters used to construct the central β-sheet barrel. We first considered the scaffold (i.e. the backbone C^{α} trace) of the hyperboloid shaped β-sheet barrel constituting the core of the protein.

The geometric properties of the 4-fold symmetric β-bar-rel, and the packing of the residues inside the β-barrel have been well described,[48-52] as has the design of an idealized β-sheet with 4-fold symmetry about the barrel axis.[53] The BIOGRAF program (Molecular Simulations Incorporated, San Diego, CA) was used to build a similar idealized β-sheet structure: the scaffold of the protein core was designed using the five geometric parameters listed in Table 1. Parameters *R*, *N* and *T*z were as described.53 Two rotation angles $\theta\beta_{odd}$ and $\theta\beta_{even}$, describing the rotation of the β-strand about its axis, were employed to create a 4-fold symmetry about the barrel axis as described by Lasters *et al.*[53] These parameters were set to 0° and 180°, respectively. The β-strands were matched onto the sheet scaffold and the resulting structure was subjected to 300 steps of conjugate gradient minimization using the DREIDING force-field,[54] in order to improve the hydrogen bond network between adjacent β-strands and reduce steric clashes between atoms. Five additional geometric parameters were used to build around the core an outer barrel composed of eight helices, the helix axes and β-strand axes being parallel. The parameters used to define the β-sheet structure and the surrounding α-helix barrel are summarized in Figure 2. The offset radius *H* and the offset angle *y* were determined by fitting natural barrels onto the idealized β-sheet scaffold. For the offset radius *H*, a distance of 10 Å was found as the average distance between the helix axes and the strand axes. This value is in good agreement with previous observations of α-helix packing against a β-sheet.[47] The offset angle *y* determined at the same time was set at 20°. The last three parameters were set by adjusting short structural motifs called αβl and αβ3 turns[35] or αGβ and αGBAβ55 at the junctions between helices and strands. An αβl connection always involves an even β-strand, and most of the time, it immediately follows an αβ3 connection, suggesting a 4-fold symmetry at the level of αβ turns. These turns are recurrent loops found in all $(\alpha/\beta)_8$-barrel proteins. They show conserved structural and sequence features, such as a tight turn of glycine residues with positive Φ angles and a characteristic hydrophobic pattern in the sequence.[35] This strategy enabled us to choose among the huge number of possibilities when defining the spatial positions of helices relative to the β-sheet barrel. The parameters used were the helix axis shift *T*, describing translation of the α-helix barrel with regard to the β-barrel along the major axis of the former, and rotation angles $\theta\alpha_{odd}$ and $\theta\alpha_{even}$, describing the rotation of odd and even helices about their major axes. The values of these three parameters were

adjusted by trial and error until the correct helix orientations were reached. The parameters $\theta\alpha_{odd} = 140°$, $\theta\alpha_{even} = -60°$,and $T = 3$ A were found to allow α-helices and β-strands to be connected by turns. Helices with idealized conformations (Φ=-57°, ψ=-47° and ω = 180°) were matched onto this scaffold. Capping-box[55] and βABα motifs[56] were used to connect the opposite side of the barrel. The various secondary structures (helices, turns, and strands) were connected so as to yield an idealized α/β-barrel backbone, thus finalizing the first step of our *de novo* design procedure.

Side-chain design

An automated side-chain selection algorithm[24] was used to find the optimal combination of side-chains in the optimal conformation compatible with the idealized backbone defined in the first step. The positions to be designed were classified as core, surface or boundary positions by means of an algorithm considering the distances of C^{α} atoms and C^{β} atoms to a surface having a dot density of 10 $Å^{-2}$, computed using the C^{α} atoms only[10,57] The surface area calculation uses the Connolly algorithm with a probe radius of 8 Å.[58] The distances considered in classifying the residue positions are the distance along the C^{α}-C^{β} vector from the C^{α} atom to the surface (*d*l) and the distance from the C^{β} atom to the nearest surface point (*d2*). If *d*1 is greater than 5Å and *d*2 is greater than 2 A, the residue is classified as a core residue. If the sum of *d*1 and *d*2 is less than *2.7*Å, the residue is classified as a surface residue. If the residue does not belong to either of these two categories, it is classified as a boundary residue. The 92 core residues were divided into two regions: residues pointing into the barrel (14 residues) and the rest (78 residues). Side-chain design was done independently in these two regions using a backbone-dependent rotamer library[59] including A, V, L, I, W, F, and Y as non-polar side-chains. To decrease the large number of aromatic side-chains found in the first round, the procedure of seeking an ensemble of rotamers for the core design was carried out twice more, with a 50 step conjugate gradient minimization between each two successive core design runs. To choose the 39 boundary residues, we used a similar procedure and a backbone-dependent rotamer library containing both hydrophobic side-chains and the charged and non-charged polar side-chains N, D, Q, E, R, K, T, S, and H. The 83 surface residues were chosen using only polar and charged side-chains. Cysteine and proline were not considered in the design, in order to avoid disulfide-bridge formation and *cis-trans* isomeriza-tion that could slow the folding process. The DEE theorem[22,60] implemented in the ORBIT program (optimization of rotamers by iterative technique) was used to solve the combinatorial search problem. The DEE theorem identifies rotamers that cannot be members of the global minimum energy conformation (GMEC). The sequences were ranked with a potential energy function depending on residue location: for core residues, it combined van der Waals potential and an atomic solvation potential.[11,23] For surface residues, it contained electrostatic and hydrogen bond potentials.[20,44] Boundary residues were chosen using a mixed surface and core potential energy function.

Energy profiles

The energy profiles were performed on the artificial structure and on a natural α/β-barrel structure using the knowledge-based mean force potential (MFP) at atomic level implemented in the ANOLEA server.[36,37] An average energy window of five residues was used for each residue in the profile.

Oligonucleotide sequence

Oligonucleotides were designed to overlap with each other over approximately 15 bases and to have a T_m between 52°C and 54°C. The lengths of the inner primers ranged from 54 to 60 bases. The two outer oligonucleotides were 25 and 27 bases long, respectively. One contained an *Nco*I restriction site and the other a *Bam*HI restriction site. Six unique restriction sites were included in the nucleotide sequence of the artificial gene. Preferred codons of *Escherichia coli*[61] were used to construct the gene (GCG Analysis Package). The oligo - nucleotides purchased from Eurogentec (Seraing, Belgium) have the following sequences (5'-3'):

f0: 27-ggcatgcc atggcgtttc tgattgtgg, **f1**: 54-atggcgtttc tgattgtgga aggtctgagc gaaaaagaac tgaaacaggc ggtg, **f2**: 58-cgcgcgattg cgtttctgaa acagtttgcc cgcaaccatg aaaaagcgga acgttttt, **f3**: 58-gcgtggaagc gattattatt gcccgcggcg tgagcgaacg tgaaattgaa caggcggc, **f4**: 58-tcgaagcgtt agccttctta gcggaatatg aacgtcgtga tcgtcagttt gatgatat, **f5**: 58-ggctttaaag cggtgattgt ggcgacaggc ctggatgaaa aggagttaaa gcaagccg, **f6**: 57-gtttcaaagc cttagccttc ttaggccgta ttgaccagga aaatcgtaac atcaacg, **f7**: 55-gggccttcgt gccatcattg cggccacggg tttaagtgag cgcgagctga gctgg, **f8**: 58-ggcctggata ttatttttgc atatggccag tttgatgaac aggataacca gtttaaac, **r1**: 58-agaaacgcaa tcgcgcgcag gccctgttcg ttcgcaatct gcaccgcctg tttcagtt, **r2**: 58-taataatcgc ttccacgcct tcacgcacta gtagttcaaa aaaacgttcc gctttttc, **r3**: 59-aagaaggcta acgcttcgaa acccttctcc cgggccagtt tcgccgcctg ttcaatttc, **r4**: 60-aatcaccgct ttaaagccat agcgttcaaa atattcaata atatcatcaa actgacgatc, **r5**: 57-aggctaaggc tttgaaaccc ttctcttcaa ttttttgcgc ggcttgcttt aactcat, **r6**: 60-gatggcacga aggccctgac gctgaaggag ctcgaagata tcgttgatgt tacgattttc, **r7**: 58-caaaaataat atccaggcca tactgacgcg

ccgcacgcag cgcccagctc agctcgcg, **r8**: 57-ctacgccgcg cccagacgac gaatcagttc cagaaaatgt ttaaactggt tatcctg, **r0**: 25-cgcggatcct acgccgcgcc cagac

Gene construction

The designed sequence was synthesized by recursive PCR[38] using 16 inner overlapping oligonucleotides, two outer oligonucleotides (25 and 27 bases), and Pwo DNA polymerase (Boehringer). The amount of each outer primer was 25 pmol and that of each inner primer was 2.5 pmol in a final volume of 100 µl. The dNTP concentration was 0.2 mM and the reaction buffer for the polymerase was used in appropriate concentration. Two amplification steps were performed as followed. In step 1, the following cycle was repeated five times: denatura-tion at 92 °C for one minute, annealing at 52 °C for two minutes, primer extension at 72 °C for one minute. In the second step, the following cycle was repeated 20 times: denaturation at 94 °C for one minute, annealing at 58 °C for two minutes, primer extension at 72 °C for one minute. The synthetic gene was inserted into the pET-11d expression vector (Stratagene) at the Ncol and *Bam* HI restriction sites to yield pET-lld-OctaV.

DNA sequencing

Double-stranded plasmid pET-11d-OctaV was purified using a standard protocol[62] and *E. coli* RR1 strain was used as a host for plasmid amplification. Plasmid dideoxy DNA sequencing was performed using the Amersham Thermo Sequenase kit with bacteriophage T7 promoter and terminator oligonucleotides labeled with fluorescein dye. The sequences were analyzed with an ALF sequencing unit (Pharmacia).

Protein expression and purification

E. coli BL21(DE3) bacteria were transformed with the pET-11d-OctaV expression vector and the protein was overproduced as inclusion bodies without the need for induction with isopropyl-β-D-thiogalactopyranoside. The inclusion bodies were collected by centrifugation, washed three times with buffer, and dissolved in 6 M urea, 50 mM Tris-HCl (pH 8.5). The solution containing the dissolved inclusion bodies was centrifuged before purification. Refolding was carried out using a modified version of a protocol used initially to refold a fusion protein immobilized on a column.[63] The mixture was loaded onto an ion-exchange column HiTrap Q (Pharmacia) and urea was removed by a downward gradient at a flow-rate of 2 ml min^{-1} The purified, refolded protein was finally eluted with 1 M NaCl. The soluble protein was separated from the aggregated fraction by gel-filtration on a Sephadex G100 column (C26-100 Pharmacia) at 4 °C in the same buffer. The different fractions were analyzed by DLS and the non-aggregated fractions were saved for further biophysical characterization.

Circular dichroism (CD) measurements

CD spectra were recorded on a Jobin-Yvon CD6 spectrometer at pH 8.5 in 50 mM borate, 850 mM NaCl. For far-UV CD spectra, a 2 mm pathlength cell was used with a one second integration time and a 1 nm bandwidth. The spectra recorded from 190 nm to 250 nm with an increment of 0.2 nm are averages of nine accumulation scans. The sample was maintained at 25 °C in a circulating waterbath controlled by a thermoelectric unit. The protein concentration was 0.25 mg/ml. For near-UV CD spectra, a 1 cm pathlength cell was used with a one second integration time and a 1 nm bandwidth. The spectra recorded from 250 nm to 310 nm with an increment of 0.2 nm are averages of nine scans with a protein concentration of 0.4 mg/ml. All spectra are corrected for blank absorption. Prior to CD spectrum acquisition, the protein concentration was determined by the Bio-Rad protein assay using bovine serum albumin as standard. The α-helix percentage was calculated as $f_\alpha = 100((-[\theta]_{208} - 4000)/29000)$.[39]

Heat denaturation

The thermal melting curve was monitored by CD at 265 nm. The protein concentration was 1.25 mg/ml in 50 mM Tris-HCl, 1 M NaCl at pH 8.5. The temperature of the sample was increased from 20 °C to 85 °C over a period of 2.5 hours. Data were collected every 15 seconds with an integration time of 2.5 seconds.

Chemical denaturation

Fluorescence measurements were performed with an Aminco SLM 8100 fluorimeter. The sample was maintained at 25 °C in a circulating waterbath controlled by a thermoelectric unit. The protein concentration was 20 µM (0.5 µg/µl) in 50 mM Tris-HCl (pH 8.5), 50 mM NaCl in each sample and for all fluorescence measurements and CD measurements. For the unfolding experiment, the excitation wavelength was 291 nm and

the fluorescence emission spectra were recorded at each dénaturant concentration in the wavelength range 320-360 nm. The average emission wavelength, ⟨λ⟩, was calculated for each emission spectrum using the equation:[64]

$$\langle\lambda\rangle = \frac{\sum_i Fi\lambda i}{\sum_i Fi}$$

F is the fluorescence intensity and λ is the wavelength. Each sample was incubated with GdnHCl at 0.1 M increments. The normalized average emission wavelength

was plotted as a function of the concentration of GdnHCl. Unfolding was monitored with the same sample by far-UV CD at 222 nm during an accumulation time of two seconds, repeated 20 times, for each concentration of denaturant with a 2 mm pathlength cell. The data were fit to a two-state model for the estimation of the free energy of unfolding in water at 25 °C.[65] The data ranging from 0 M to 1 M in GdnHCl were not included in the fitting of the unfolding curve followed by CD at 222 nm. Chemical unfolding of the protein in the presence of a high concentration of salt was monitored by recording the wavelength of maximum fluorescence emission of tryptophan upon excitation at 280 nm. Each sample containing protein at 1 μM (25 μg/ml) in 50 mM borate, 850 mM NaCl at pH 8.5 was run in triplicate, in the presence of from 0 M to 5 M GdnHCl by 0.25 M increments. The normalized wavelength of maximum fluorescence emission of tryptophan was plotted as a function of the concentration of GdnHCl. The data were fit to a two-state model for the estimation of the free energy of unfolding in water at 25 °C. The reversibility of the unfolding reaction was verified prior to the fluorescence measurements in a rena-turation test, as follows: the artificial protein (1 μM) was first denatured in 6 M GdnHCl in 50 mM borate, 850 mM NaCl at pH 8.5 for 16 hours, then renaturation was induced by dialysis against the buffer used for the fluorescence measurements.

ANS binding measurements

1-Anilinonaphthalene-8-sulfonic acid (ANS) was purchased from Sigma. The extrinsic fluorescence measurements were performed with the same fluorimeter as that used for the intrinsic fluorescence measurements. The protein concentration was 1 μM (25 μg/ml) in 50 mM borate, 850 mM NaCl at pH 8.5 and the concentration of ANS was 150 μM with 0 M, 1 M, or 2 M GdnHCl. The solutions were left overnight for equilibration. The excitation wavelength was 390 nm and the emission of fluorescence was monitored from 400 nm to 650 nm. The concentration of ANS was calculated from the measurement of the absorbance at 350 nm using an extinction coefficient of 5000 cm^{-1} M^{-1}.[66] The ANS fluorescence was measured under the following conditions: protein concentration 20 μM (0.5 μg/μl) in 50 mM Tris-HCl (pH 8.5), 50mM NaCl with the concentration of GdnHCl ranging from 0 M to 3 M, and addition of ANS at the final concentration of 2 mM.

Light-scattering

DLS was measured with a Brookhaven apparatus consisting of a BI-200 goniometer, a BI-2030 digital correlator, and a LEXEL Ar ion laser operating at 488 nm wavelength and with a Photocor apparatus with a He-Ne laser operating at 633 nm. All measurements were performed at a scattering angle of 90° and the temperature was maintained at 25 °C. All samples containing the protein were centrifuged for 20 minutes at l0000g prior to data acquisition. Distribution functions of the translational diffusion coefficient D_t and of the hydro-dynamic radius were derived from the autocorrelation functions using the program CONTIN[67] and the cumulant analysis.

Nuclear magnetic resonance (NMR) measurements

^{1}H NMR spectra were recorded with a Bruker DRX 400 MHz spectrometer in 50 mM Hepes (pH 7), 100 mM NaCl. Under these conditions, highly concentrated protein solution (12 mg/ml) was stable enough for recording a proper ^{1}H NMR spectrum during the time-course of the NMR experiment (six hours). The final protein concentration was 35 mg/ml for the non-folded form of the protein obtained after dialysis of the protein denatured in 6 M GdnHCl against 50 mM Tris-HCl (pH 8.5). All assays were performed at 25 °C. All programs were from the Bruker library. Proton NMR spectra with water presaturation were obtained with a spectral width of 4 kHz for 16 K frequency and time-domain data points.

Acknowledgements

We are grateful to D. Schaak and P. Osterhout of the Rowland Institute, B. Gordon and A. Street of Caltech for their very stimulating scientific interest and to A. Matagne, M. Galleni and M. Muller for critical reading of the manuscript. We thank A. Stern of the Rowland Institute for the backbone construction and N. Otthiers for peptide sequencing. F.O. is the recipient of a doctoral fellowship from the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA). This work was supported, in part, by European Space Agency grant number 12987/98/NL/VJ(IC).

References

1. Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, 405, 39-42.

2. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, 96,11305-11310.

3. Ferrara, P. & Caflisch, A. (2001). Native topology or specific interactions: what is more important for protein folding? *J. Mol. Biol.* 306, 837-850.

4. Kim, D. E., GU, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA*, 95, 4982-4986.

5. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu. Rev. Biochem.* 66, 549-579.

6. DeGrado, W. E, Summa, C. M., Pavone, V., Nastri, F. & Lombardi, A. (1999). *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* 68, 779-819.

7. DeGrado, W. F. (1997). Proteins from scratch. *Science*, 278, 80-81.

8. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193, 775-791.

9. Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* 125, 357-386.

10. Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, 278, 82-87.

11. Dahiyat, B. I. (1999). *In silico* design for protein stabilization. *Curr. Opin. Biotechnol.* 10, 387-390.

12. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* 5, 895-903.

13. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *ƒ. Mol. Biol.* 290, 305-318.

14. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA*, 94, 10172-10177.

15. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, 282, 1462-1467.

16. Harbury, P. B., Tidor, B. & Kim, P. S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl Acad. Sci. USA*, 92, 8408-8412.

17. Wernisch, L., Hery, S. & Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.* 301, 713-736.

18. Desjarlais, J. R. & Handel, T. M. (1995). New strategies in protein design. *Curr. Opin. Biotechnol.* 6, 460-466.

19. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* 5, 470-475.

20. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Struct. Fold Des. 7*, R105-R109.

21. Desmet, J., De Mayer, M., Hazez, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356, 539-542.

22. Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. ƒ*. 66, 1335-1340.

23. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* 9, 509-513.

24. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). *De novo* protein design: towards fully automated sequence selection (see comments). *ƒ. Mol. Biol.* 273, 789-796.

25. Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* 6,1701-1707.

26. Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.* 6, 1167-1178.

27. Fezoui, Y., Connolly, P. J. & Osterhout, J. J. (1997). Solution structure of alpha t alpha, a helical hairpin peptide of *de novo* design. *Protein Sci.* 6, 1869-1877.

28. Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H. & DeGrado, W. F. (1999). Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc. Natl Acad. Sci. USA*, 96, 5486-5491.

29. Beauregard, M. *et al.* (1991). Spectroscopic investigation of structure in octarellin (a *de novo* protein designed to adopt the alpha/beta-barrel packing). *Protein Eng.* 4, 745-749.

30. Goraj, K., Renard, A. & Martial, J. A. (1990). Synthesis, purification and initial structural characterization of octarellin, a *de novo* polypeptide modelled on the alpha/beta-barrel proteins. *Protein Eng.* 3, 259-266.

31. Tanaka, T., Hayashi, M., Kimura, H., Oobatake, M. & Nakamura, H. (1994). *De novo* design and creation of a stable artificial protein. *Biophys. Chem.* 50, 47-61.

32. Tanaka, T., Kimura, H., Hayashi, M., Fujiyoshi, Y., Fukuhara, K. & Nakamura, H. (1994). Characteristics of a *de novo* designed protein. *Protein Sci.* 3, 419-427.

33. Houbrechts, A. *et al.* (1995). Second-generation octa-rellins: two new *de novo* $(beta/alpha)_8$ polypeptides designed for investigating the influence of beta-residue packing on the alpha/beta-barrel structure stability. *Protein Eng.* 8, 249-259.

34. Tanaka, T., Kuroda, Y., Kimura, H., Kidokoro, S. & Nakamura, H. (1994). Cooperative deformation of a *de novo* designed protein. *Protein Eng. 7,* 969-976.

35. Scheerlinck, J. P. *et al*. (1992). Recurrent alpha beta loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins: Struct. Funct. Genet.* 12, 299-313.

36. Melo, F. & Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *ƒ. Mol. Biol.* 277, 1141-1152.

37. Melo, F & Feytmans, E. (1997). Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.* 267, 207-222.

38. Prodromou, C. & Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis. *Protein Eng.* 5, 827-829.

39. Greenfield, N. & Fasman, G. D. (1969). Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*, 8, 4108-4116.

40. Strickland, E. H. (1974). Aromatic contributions to circular dichroism spectra of proteins. *CRC Crit. Rev. Bioche.m* 2, 113-175.

41. Ptitsyn, O. (1996). How molten is the molten globule? *Nature Struct. Biol.* 3, 488-490.

42. Betz, S. F., Raleigh, D. P. & DeGrado, W. F. (1993). *De novo* protein design: from molten globules to nativelike states. *Curr. Opin. Struct. Biol.* 3, 601-610.

43. Strop, P., Marinescu, A. M. & Mayo, S. L. (2000). Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci.* 9, 1391-1394.

44. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* 6, 1333-1337.

45. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnol.* 17, 691-695.

46. Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2000). Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold. *Nature*, 403, 617-622.

47. Cohen, F. E., Sternberg, M. J. & Taylor, W. R. (1982). Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* 156, 821-862.

48. Murzin, A. G., Lesk, A. M. & Chothia, C. (1994). Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* 236, 1369-1381.

49. Murzin, A. G., Lesk, A. M. & Chothia, C. (1994). Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* 236, 1382-1400.

50. Lasters, I., Wodak, S. J., Alard, P. & van Cutsem, E. (1988). Structural principles of parallel beta-barrels in proteins. *Proc. Natl Acad. Sci. USA*, 85, 3338-3342.

51. Wodak, S. J., Lasters, I., Pio, F. & Claessens, M. (1990). Basic design features of the parallel alpha/ beta barrel, a ubiquitous protein-folding motif. *Biochem. Soc. Symp.* 57, 99-121.

52. Lesk, A. M., Branden, C. I. & Chothia, C. (1989). Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. *Proteins: Struct. Funct. Genet.* 5,139-148.

53. Lasters, I., Wodak, S. J. & Pio, F (1990). The design of idealized alpha/beta-barrels: analysis of beta-sheet closure requirements. *Proteins: Struct. Funct. Genet. 7*, 249-256.

54. Mayo, S. L., Olafson, B. D. & Goddard, W. A. I. (1990). Dreiding: a generic force field for molecular simulations. *J. Phys. Chem.* 94, 8897-8909.

55. Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci.* 7, 21-38.

56. Wintjens, R., Wodak, S. J. & Rooman, M. (1998). Typical interaction patterns in alpha/beta and beta/alpha turn motifs. *Protein Eng.* 11, 505-522.

57. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des.* 3, 253-258.

58. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221, 709-713.

59. Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* 230, 543-574.

60. De Maeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des.* 2, 53-66.

61. Nakamura, Y., Gojobori, T. & Ikemura, T. (1998). Codon usage tabulated from the international DNA sequence databases. *Nucl. Acids Res.* 26, 334.

62. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edit., vol. 2, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

63. Stempfer, G., Holl-Neugebauer, B. & Rudolph, R. (1996). Improved refolding of an immobilized fusion protein. *Nature Biotechnol.* 14, 329-334.

64. Royer, C. A., Mann, C. J. & Matthews, C. R. (1993). Resolution of the fluorescence equilibrium unfolding profile of trp aporepressor using single tryptophan mutants. *Protein Sci.* 2, 1844-1852.

65. Fersht, A. (1998). *Structure and Mechanism in Protein Science*, Freeman, W.H. and company, New York.

66. De Filippis, V., de Laureto, P. P., Toniutti, N. & Fontana, A. (1996). Acid-induced molten globule state of a fully active mutant of human interleukin-6. *Biochemistry*, 35, 11503-11511.

67. Provencher, S. W. (1982). CONTIN: a general purpose constrained regularization program for inverting noisy linear algebraic and intergral equations. *Comput. Phys. Commun.* 27, 229.