Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure prediction tools

Maximiliano Figueroa[a,*,1,2], Mike Sleutel[b,1], Marylene Vandevenne[a], Gregory Parvizi[a], Sophie Attout[a], Olivier Jacquin[a], Julie Vandenameele[c], Axel W. Fischer[d], Christian Damblon[e], Erik Goormaghtigh[f], Marie Valerio-Lepiniec[g], Agathe Urvoas[g], Dominique Durand[g], Els Pardon[b,h], Jan Steyaert[b,h], Philippe Minard[g], Dominique Maes[b], Jens Meiler[d], André Matagne[c], Joseph A. Martial[a], Cécile Van de Weerdt[a,*]

[a] GIGA-Research, Molecular Biomimetics and Protein Engineering, University of Liège, Liège, Belgium
[b] Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[c] Laboratoire d'Enzymologie et Repliement des Protéines, Centre for Protein Engineering, University of Liège, Liège, Belgium
[d] Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN, United States
[e] Department of Chemistry, Univeristy of Liège, Belgium
[f] Laboratory for the Structure and Function of Biological Membranes, Center for Structural Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium
[g] Institute for Integrative Biology of the Cell (I2BC), UMT 9198, CEA, CNRS, Université Paris-Sud, Orsay, France
[h] Structural Biology Research Center, VIB, Pleinlaan 2, 1050 Brussels, Belgium

ARTICLE INFO

Article history:
Received 15 February 2016
Received in revised form 19 April 2016
Accepted 12 May 2016
Available online 12 May 2016

Keywords:
De novo design
Artificial proteins
Protein design
Molecular modeling

ABSTRACT

Despite impressive successes in protein design, designing a well-folded protein of more 100 amino acids *de novo* remains a formidable challenge. Exploiting the promising biophysical features of the artificial protein Octarellin V, we improved this protein by directed evolution, thus creating a more stable and soluble protein: Octarellin V.1. Next, we obtained crystals of Octarellin V.1 in complex with crystallization chaperons and determined the tertiary structure. The experimental structure of Octarellin V.1 differs from its *in silico* design: the (αβα) sandwich architecture bears some resemblance to a Rossman-like fold instead of the intended TIM-barrel fold. This surprising result gave us a unique and attractive opportunity to test the state of the art in protein structure prediction, using this artificial protein free of any natural selection. We tested 13 automated webservers for protein structure prediction and found none of them to predict the actual structure. More than 50% of them predicted a TIM-barrel fold, i.e. the structure we set out to design more than 10 years ago. In addition, local software runs that are human operated can sample a structure similar to the experimental one but fail in selecting it, suggesting that the scoring and ranking functions should be improved. We propose that artificial proteins could be used as tools to test the accuracy of protein structure prediction algorithms, because their lack of evolutionary pressure and unique sequences features.

1. Introduction

A longstanding dogma in structural biology states that the tertiary structure of a protein is largely determined by its primary structure (Dobson, 2003). It is also widely accepted that in evolution structure is better conserved than sequence, i.e. quite diverging sequences fold into a similar tertiary structure at the level of arrangement of secondary structure elements. Solving the protein folding problem, i.e., predicting a protein's tertiary structure from its primary structure *de novo* is considered the "holy grail" of computational structural biology. Conversely, an "inverse protein folding problem" can be defined as, given a three-dimensional structure often in the form of arrangement of secondary structure element, the design of a sequence that folds into the desired fold (Pabo, 1983). Besides its importance as a fundamental question in biology, solving the inverse folding problem paves the way to engineering proteins with custom structures

[*] Corresponding authors.
E-mail addresses: maxifigueroa@udec.cl (M. Figueroa), c.vandeweerdt@ulg.ac.be (C. Van de Weerdt).
[1] These authors contributed equally to this work.
[2] Current address: Biochemistry and Molecular Biology Department, University of Concepcion, Chile.

http://dx.doi.org/10.1016/j.jsb.2016.05.004
1047-8477/© 2016 Elsevier Inc. All rights reserved.

and functions. In recent years, principles have been emerging and great successes have been achieved in the design of small artificial proteins, some with specific catalytic activities (Fleishman et al., 2011; Koga et al., 2012; Kuhlman et al., 2003; Röthlisberger et al., 2008). These breakthroughs are partly due to improvements of dedicated algorithms, which can now explore vast regions of conformational space with improved energy functions and in reasonable time (Carbonell and Trosset, 2015). Due to the close relation of the protein folding problem with its inverse cousin, these improvements are also illustrated by progress in protein structure prediction as illustrated by CASP events over the last decade (Kryshtafovych et al., 2014).

As an alternative to *de novo* protein design, directed molecular evolution has been successfully developed to improve the structural and functional properties of natural enzymes (Tao and Cornish, 2002). This approach, which exploits a simple iterative Darwinian optimization process, has led to major improvements of properties such as catalytic activity (Reetz, 2007), stability (Eijsink et al., 2005) and solubility (Waldo, 2003). Not surprisingly, directed evolution has also emerged as the best way to optimize the properties of *de novo* designed enzymes (Khersonsky et al., 2010; Ward, 2008) or even generate new catalytic functions, when combined with computational methods (Chaput et al., 2008). This "black box" approach remains the most effective way to break through the existing limitations of *in silico* design.

As negative results are greatly underreported in the field, it is hard to determine the maximum protein size achievable with *de novo* design. Considering the fact that for a given length of "n" amino acids of artificial sequence protein, we have then 20^n possible sequences. This creates a very large space to explore in order to find the sequence(s) which can fold as the desired target. The ability of a software to explore and then select these sequences with high accuracy is affected with each amino acids added to a polypeptidic chain, not only for the new 20 possible amino acids that could belong to the new position, but also for all the possible interactions, geometry and effects at local and global level in the protein structure that have to be tested. Then, the size of a *de novo* designed protein really matters. The most convincing success of *de novo* protein design – the TOP7 fold – has only 106 residues (Kuhlman et al., 2003). However, *de novo* construction of a stable, soluble single-domain protein of more than two hundred amino acids is still a challenge. The few successes reported so far in the construction of large artificial proteins often involved assembly of multiple copies of the same motif, each not exceeding 40 amino acids in length (Parmeggiani et al., 2008; Urvoas et al., 2010). Among the last group, clearly protrudes from the rest the work of Huang et al. (2016), where they clearly succeeded in the design, production and characterization of an artificial TIM barrel protein of 184 amino acids, taking advantage of the structural internal symmetry of the protein, repeating four times the same motif. Other approaches involve recombination of larger protein fragments with a limited redesign of residues at the interface, an approach that has been applied successfully to $(\beta\alpha)_8$ barrel proteins (Eisenbeis et al., 2012; Fortenberry et al., 2011). Both approaches, although very valuable, ultimately limit the structural diversity that can be achieved with *de novo* designed proteins as large portions of existing protein are reused as templates. The goal of our ongoing Octarellin project is to design a well-structured single-domain protein exceeding the (arbitrary yet appealing) 200-amino-acid threshold without considering any internal symmetry and with a $(\beta\alpha)_8$ fold.

Octarellins are artificial proteins, more than 200 amino acids long, designed to adopt the $(\beta\alpha)_8$ fold characteristic of the archetypal TIM barrel. Work in our lab, based on various approaches, has yielded several generations of Octarellins (Beauregard et al., 1991; Figueroa et al., 2013; Goraj et al., 1990; Houbrechts et al., 1995), but solubility and structural stability issues have prevented us from determining the exact structure of any of them. Although the secondary structure of one of the previous version, Octarellin V, described in 2003 (Offredi et al., 2003), seemed compatible with the *in silico* model, this protein failed to meet the technical requirements for NMR spectroscopy and X-ray diffraction.

In the present work, we have used directed evolution to optimize the artificial protein Octarellin V to improve solubility and stability. The optimized protein is called Octarellin V.1. We have crystallized this protein with the help of different crystallization chaperons and have determined its tertiary structure. As it turns out, the experimental X-ray structure deviates from our idealized $(\beta\alpha)_8$ design. This unexpected result has led us to take a close look at the state of the art in automated protein structure prediction, using, in CASP fashion, the primary structure of Octarellin V.1 as sole boundary condition for several automated protein structure prediction servers. The results demonstrate the shortcomings of existing automated servers, as more than a half of them predicted a $(\beta\alpha)_8$ structure, similar to the designed protein, while none of the them could give us the real fold of the protein.

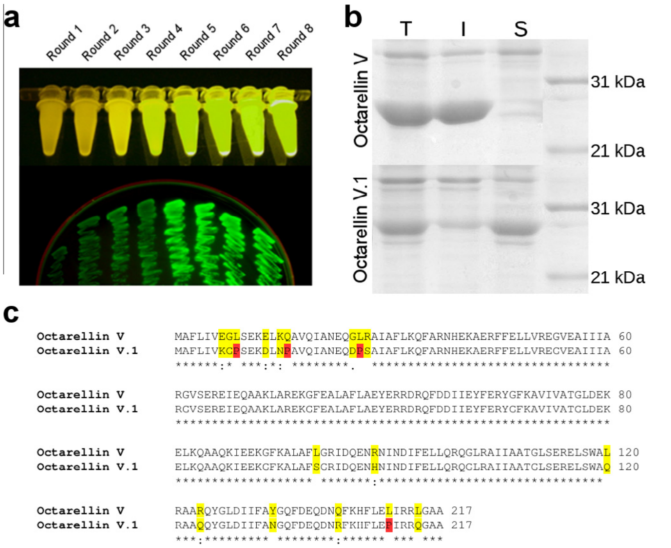
2. Results

2.1. Selection of a soluble variant of Octarellin V by directed evolution

More than ten years ago, in our attempt to address the inverse folding problem, we designed the artificial protein Octarellin V (Offredi et al., 2003). This protein displayed promising features as it was not a molten globule and its secondary structure content was compatible with the *in silico* design. However, the protein was expressed in inclusion bodies and both stability and solubility were unsatisfactory. To improve these properties and make the protein amenable for further characterization, we have performed eight consecutive rounds of error-prone-PCR-based directed evolution, using as retaining criterion the solubility of the protein inside bacteria (see Section 4). The resulting chosen variant, dubbed Octarellin V.1, displays 16 mutations located mainly in the N- and C-terminal regions (93% sequence identity; Fig. 1).

2.2. Octarellin V.1 is more stable and better folded than its parent protein Octarellin V

As directed evolution can alter the structure of an artificial protein, Octarellin V.1 was biophysically characterized in order to compare it with both Octarellin V and the *in silico* design. Far-UV circular dichroism (CD) spectroscopy analysis of Octarellin V.1 revealed ~32%, and ~22% of helical and β-strand content, respectively (Figueroa et al., n.d.). These values are identical within error limit to those determined by infrared spectroscopy (~30% and ~16%, respectively) and furthermore they are not significantly different from the values obtained for Octarellin V, hence suggesting that the directed evolution process did not cause any significant changes at the secondary structure level. At the tertiary structure level, the Octarellin V.1 looks well folded NMR spectroscopy also supports the presence of tertiary structure (Fig. 2), although we conclude from the 153 out of a possible 217 signals in the 2D-HSQC spectrum recorded at pH 7.0 that a portion of the protein is either unstructured or highly mobile. Moreover, small-angle X-ray scattering (SAXS, Fig. 2) revealed a minor difference between Octarellin V and V.1 proteins in high q range ($q > 0.28$ Å^{-1}), indicating changes among short distances smaller than about 20 Å (Fig. 2a). In addition, the biophysical characterization of Octarellin V.1 showed a thermostable protein with cooperative unfolding (Figueroa et al., n.d.).



```
Octarellin V     MAFLIVEGLSEKELKQAVQIANEQGLRAIAFLKQFARNHEKAERFFELLVREGVEAIIIA 60
Octarellin V.1   MAFLIVKGPSEKDLNPAVQIANEQDPSAIAFLKQFARNHEKAERFFELLVREGVEAIIIA 60
                 ******:* ***:*: ********.  *********************************

Octarellin V     RGVSEREIEQAAKLAREKGFEALAFLAEYERRDRQFDDIIEYFERYGFKAVIVATGLDEK 80
Octarellin V.1   RGVSEREIEQAAKLAREKGFEALAFLAEYERRDRQFDDIIEYFERYGFKAVIVATGLDEK 80
                 ************************************************************

Octarellin V     ELKQAAQKIEEKGFKALAFLGRIDQENRNINDIFELLQRQGLRAIIAATGLSERELSWAL 120
Octarellin V.1   ELKQAAQKIEEKGFKALAFSGRIDQENHNINDIFELLQRQGLRAIIAATGLSERELSWAQ 120
                 ******************* *******:*******************************

Octarellin V     RAARQYGLDIIFAYGQFDEQDNQFKHFLELIRRLGAA 217
Octarellin V.1   RAAQQYGLDIIFANGQFDEQDNRFKHFLEPIRRQGAA 217
                 ***:********* ********:****** *** ***
```

Fig. 1. Obtaining Octarellin V.1 by directed evolution. a) GFP-based screening for soluble mutants was used through 8 rounds of directed evolution; the best clone of each round is shown. b) SDS-PAGE analysis showing that the 16 mutations led to a soluble protein (T: total fraction; I: insoluble fraction; S: soluble fraction). c) Sequence alignment of Octarellin V and Octarellin V.1, highlighting in yellow the mutations after directed evolution.

All these results confirm the validity of directed evolution as a tool for obtaining a more stable and soluble protein without significantly altering the tertiary structure. Interestingly, their SAXS curves were found to deviate strongly from the theoretical one (Fig. 2a, solid gray curve), indicating that neither Octarellin V.1 nor Octarellin V are folded as expected from the *de novo* design model.

2.3. The structure of Octarellin V.1 reveals a mismatch with its de novo design

The apparent structural flexibility, combined with the lack of promising nucleation hits, motivated us to use chaperons to crystallize the optimized protein. To minimize the risk of misinterpreting results, we chose two different crystallization chaperons and compare the obtained structures of Octarellin V.1.

We thus selected two kinds of tailor-made binders boasting excellent track records: nanobodies (Domanska et al., 2011; Korotkov et al., 2009; Rasmussen et al., 2011) (also called VHH antibody fragments) and αRep (Ferrandez et al., 2014; Guellouz et al., 2013; Urvoas et al., 2010) proteins. The former, which are well known, are single-domain antibody fragments derived from a *Camelidae* species; the latter constitute a relatively new family of artificial protein binders based on natural HEAT repeat proteins (Guellouz et al., 2013; Urvoas et al., 2010). We obtained seven monoclonal nanobodies (Nb) after a llama immunization with 1 mg of Octarellin V.1 (see Section 4) and were able to isolate stable complexes with all of them, using size exclusion chromatography. From a library of 1.7×10^9 clones, we selected four independent αRep proteins (with a varying number of repeats), all of which successfully formed complexes with Octarellin V.1. Isothermal titration calorimetry confirmed 1:1 binding stoichiometry for all complexes, with dissociation constants in the nanomolar range for the nanobodies and in the micromolar range for the αRep chaperons (Figueroa et al., n.d.). Crystals of one nanobody-Octarellin V.1 complex were obtained, in two different polymorphs, $P2_1$ and $I4_122$, diffracting at 1.95 and 3.20 Å, respectively. In addition, one αRep-Octarellin V.1 complex was successfully crystallized in the $P2_1$ space group, diffracting at a maximum resolution of 2.22 Å.

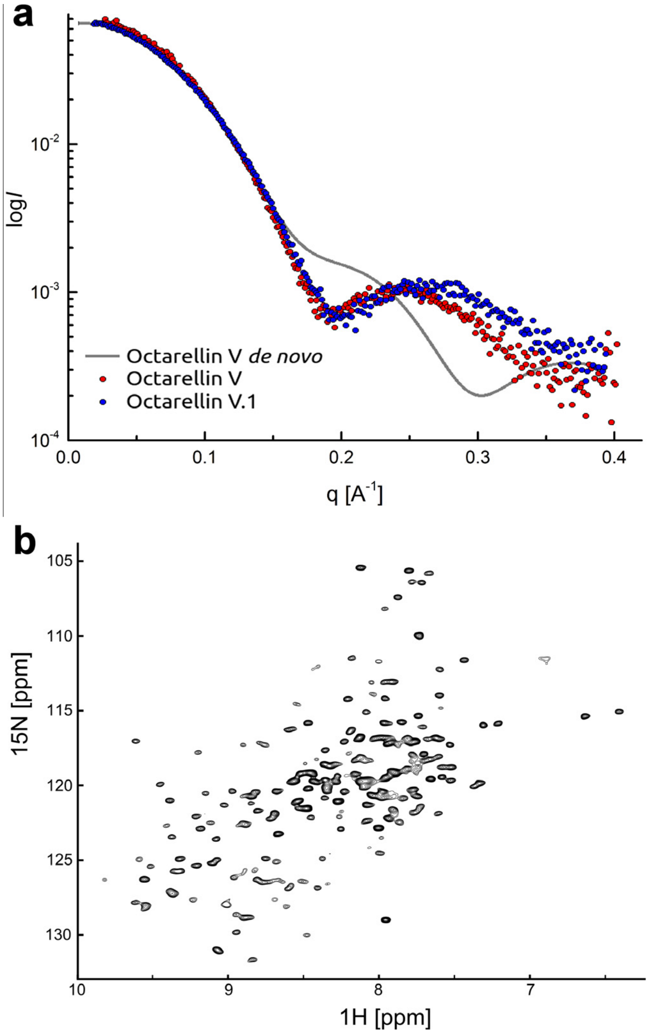


Fig. 2. Tertiary structure analysis by SAXS and 2D-NMR. a) SAXS curves of Octarellins V (red) and V.1 (blue). They are almost superimposed, with differences in the high-q range. From these curves, the radius of gyration (Rg) was estimated at 20.33 ± 0.3391 Å for Octarellin V.1 and 20.198 ± 0.834 Å for Octarellin V. The expected SAXS curve for the *in silico* design is shown as well (green), denoting no correspondence after $q > 0.15$ Å^{-1} with the experimental data (i.e. for distances smaller than 40 Å), so the protein is expected to adopt a fold differing from the intended one. b) 2D-HSQC NMR spectrum. Spectra were recorded at 65 °C in 50 mM phosphate buffer, pH 7.0. The presence of 153 signals and their broad distribution support the presence of a well-defined tertiary structure, but with some non-structured regions.

The structure of the protein complexes were determined (Fig. 3). Using the information of the nanobody and αRep proteins, the molecular replacement technique was used to solve the phase problem (see Section 4 and Table 1 for X-ray diffraction statistics). The first crystal for the nanobody – Octarellin V.1 complex ($P2_1$) showed two protein complexes per unit cell, whereas the second crystal ($I4_122$) showed only one. In both cases the structure of the Octarellin is not complete, missing information in a segment after the first strand between Pro9 and Arg37; the information for the C-term is missing as well, having not electron density between Phe192 and Ala217. The structure obtained from the αRep-Octarellin V.1 complex is also incomplete. In this case, the information between the amino acids Gly8 – P26, Phe35 – His39, Gly62 – Glu65, Ala87 – Asp97, Ala114 – Gln123, Lys135 – Asn147, Phe192 – Glu199, and Arg213 – Ala217 is missing. The superimposition of the three Octarellin V.1 structures displays a full agreement in the position of each secondary structure

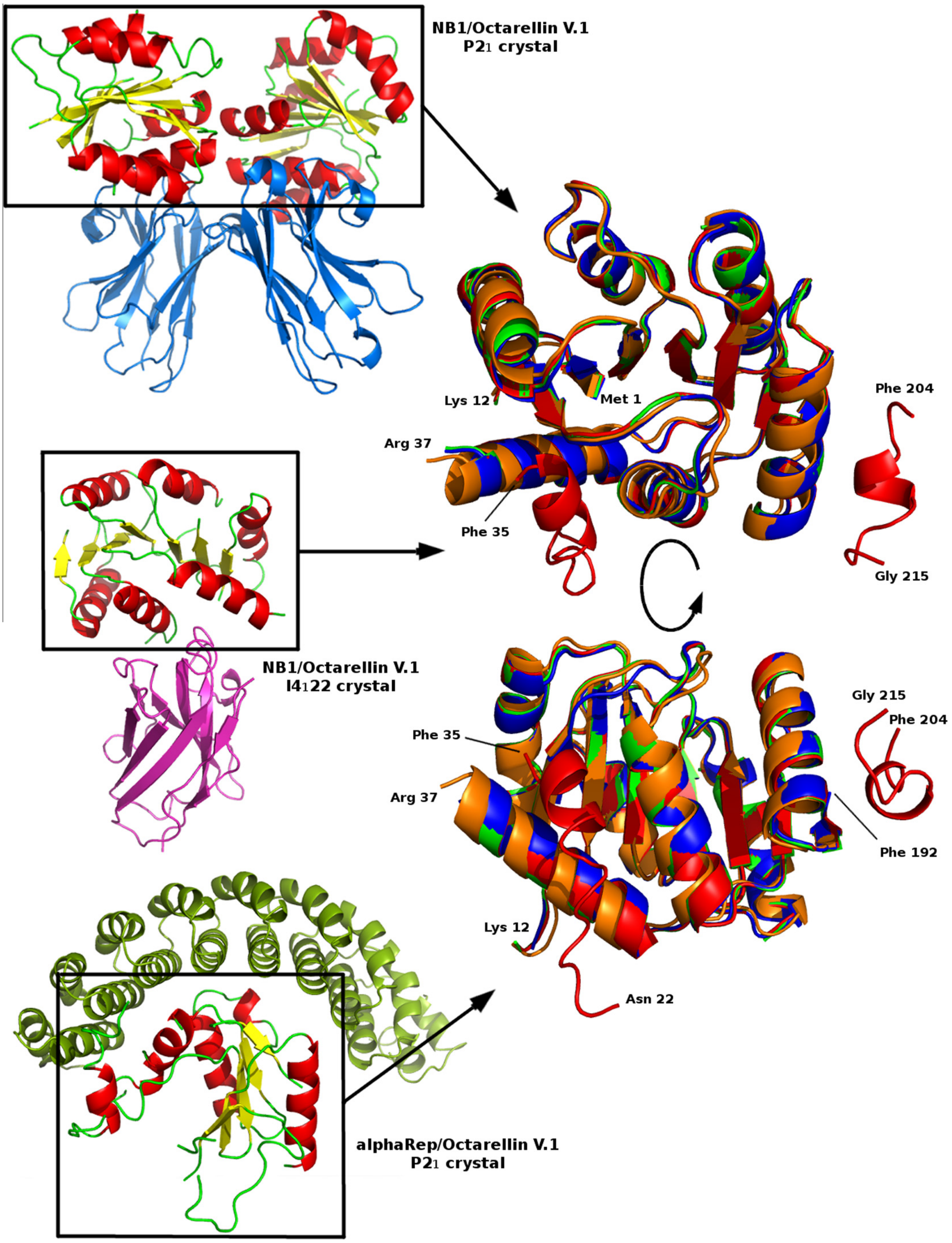


Fig. 3. X-ray structure of Octarellin V.1. The Octarellin V.1 has been co-crystallized as complex using a nanobody (NB1) and also an αRep. Two different crystals have been obtained for the complex NB1/Octarellin V.1 (spaces groups P21 and I4122). The crystal P21 has an unit cell consists of 2 nanobodies (light blue) and 2 Octarellins V.1 (whose secondary structure elements (SSE) are in color: red (helices), yellow (strands) and non-ordered sections (green). The crystal belonging to the I4122 space group displays only one complex per unit cell (nanobody in magenta and Octarellin V.1 colored by its SSE). The complex aRep/Octarellin V.1 has been crystallized in the space group P21 and it display only one complex per unit cell (αRep in dark green and Octarellin V.1 colored by its SSE). The structure of Octarellin V.1 in each crystal is not complete, missing some regions. However, the superposition of all the structures (right panel) shows the agreement of the overall fold between them, but also a complementarity: some regions that are not present in one crystal can be observed in other crystal (blue and green, Octarellin V.1 from NB1/Octarellin V.1 P21 crystal; orange, NB1/Octarellin V.1 I4122 crystal; red, αRep/Octarellin P21 crystal).

elements. Moreover, there is a complementarity between the structures: some segments not observed using the nanobody are observed using αRep and vice versa. Using this complementarity, and modeling the final missing parts, we built the final model of Octarellin V.1 (Figs. 3 and 4a). These structures haven been deposited into the Protein Data Bank under the identifiers 5BOP and 4ZV6.

From our consensus X-ray model (Fig. 4a), we conclude that Octarellin V.1 does have a 3-layer (αβα) sandwich architecture, but with a topology reminiscent of the Rossmann fold (Rao and Rossmann, 1973), rather than the expected TIM-barrel fold. All four relevant hits [score >2.0] obtained when searching against the CATH domain database have the *3.40.50 Rossmann fold* topology. The fold of Octarellin V.1 differs, however, from the canonical Ross-

Table 1
Data collection and refinement statistics (molecular replacement).

	Nanobody – Octarellin V.1	Nanobody – Octarellin V.1	αRep – OctarellinV.1
Data collection			
Space group	P21	I4122	P21
Cell dimensions			
a, *b*, *c* (Å)	54.92, 62.86, 95.11	100.03, 100.03, 158.42	73.13, 42.02, 85.19
α, β, γ (°)	90, 96.217, 90	90, 90, 90	90, 106.35, 90
Resolution (Å)	49.67–1.95 (2.00–1.95)	19.85–3.20 (3.28–3.20)	47.10–2.22 (2.28–2.22)
R_{merge}	3.9 (181)%	13.6 (64.1)%	8.5 (118.2)%
$I/\sigma I$	8.51 (0.90)	18.8 (3.1)	16.15 (0.75)
Completeness (%)	97.5 (96.6)	99.3 (99.5)	98.8 (98.9)
Redundancy	2.3 (2.2)	4.3 (3.9)	3.7 (3.9)
Refinement			
Resolution (Å)			
No. reflections	210180 (14392)/89920 (6596)	54584 (3721)/12606 (948)	90942 (6881)/24372 (1771)
$R_{\text{work}}/R_{\text{free}}$	19.21/23.49	26.81/29.96	23.25/28.68
No. atoms	4725	2192	3404
Protein	4574	2192	3404
Ligand/ion	0	0	0
Water	151	0	0
B-factors	44.31 (14.83–122.09)	62.35 (7.90–151.23)	66.51 (29.15–130.46)
Protein	44.5	62.35 (7.90–151.23)	66.51 (29.15–130.46)
Water	37.5	–	–
R.m.s. deviations			
Bond lengths (Å)	0.011	0.006	0.009
Bond angles (°)	1.227	1.019	1.226

Values for the highest resolution shell are given in between brackets. CC1/2 values were used as a guide for selecting the highest usable resolution shell (Karplus and Diederichs, 2012).

mann fold in the connectivity of the secondary structure elements. This could mean that Octarellin V.1 displays a new fold. Whereas the archetypal Rossmann topology consists of two domains (S6||S5||S4||S1||S2||S3) with all β-strands having parallel contacts (||). The central β-sheet in Octarellin V.1 has an architecture of (S8||S7||S3||S4||S5xS1xS6) with two antiparallel β-strand contacts (x), suggesting that "*Rossmann-like* fold" is a more appropriate designation. To check that the protein has the same structure in solution and that this result is not an artifact produced by interaction with the binders, we performed a SAXS analysis, comparing the experimental data with the theoretical SAXS curve obtained from the 3D coordinates of the protein (Supporting Fig. 1). The computed SAXS curve showed an almost perfect fit with the experimental data, indicating that the structure of the protein in solution is the same as that observed in the protein complexes.

Analyzing the experimental structure and comparing it with its *de novo* design (Fig. 4a), we expected the first β-strand to be flanked by S2 and S8, but this is not the case. Rather, S1 is neighbored by S5 and S6. Most surprisingly, the orientation of S1 is anti-parallel to the sheet. The significant structural discrepancy between the computational TIM-barrel design and the experimental model is readily apparent from the *global distance test – total score* (GDT_TS = 28.7) and *local-global alignment* (LGA = 26.8) (Keedy et al., 2009; Zemla, 2003) values obtained upon comparing the two structures. Octarellin V.1 clearly deviates from our design at fold level. This raises the question: does the design also fail at the secondary structure level? Largely, the answer is no (Supporting Fig. 2). Broadly speaking, Octarellin V.1 has seven of the eight predicted β-strands (the expected S2 is a loop in our X-ray model) and six confirmed helices out of the expected eight (the expected H1 is mostly a loop, and we could observe only a fraction of the residues that should constitute H8). DSSP analysis of the two structures revealed that roughly 74% of the designed secondary structure features are present in the final structure, with the obvious reservation that H8 still awaits experimental verification. Analyzing the (βα) motifs, we observed five of the expected eight, and the correlation at structural level is quite good in those motifs (Supporting Fig. 3). Moreover, the region of the protein between amino acids Glu55 and Phe134, appears to be well designed (Fig. 5c), having an almost perfect structural alignment between the *de novo* design and the actual structures. All these analyses suggest that our protein was essentially well designed at secondary structure level, but that it lacks the interactions among secondary structure elements required to get the intended tertiary structure and thus displays a fold differing from the *in silico* model.

2.4. Can automated protein structure prediction tools predict the structure of Octarellin V.1?

An interesting question is whether this global structural mismatch could have been predicted with current structure prediction algorithms. To answer this question, we submitted the sequence of Octarellin V.1 to the 13 top-ranking protein structure prediction servers from the CASP10 event (Huang et al., 2014; Tai et al., 2014) (www.predictioncenter.org). We obtained 49 models (21 *de novo*, 25 threading, and 3 homology modeling models). These models were analyzed (LGA (Keedy et al., 2009; Zemla, 2003)), clustered (STRALCP (Zemla et al., 2007)), and ranked according to their structural proximity to the target X-ray model (by decreasing normalized GDT_TS value; see Fig. 5a and Supplementary Table 1). The best-scoring model (model1_PMS; N GDT_TS = 30.5) performs only marginally better than our *de novo* design (Supplementary Table 1, Octarellin V *de novo* design values; N GDT_TS = 28.7). We detect no obvious correlation between the rank of a model and the corresponding method of structure prediction, i.e. *de novo* modeling and threading are equally good/poor at predicting the target structure. The three models produced by homology modeling score towards the lower end of the ranking. This is logical, given the extremely low sequence identity to any known (natural or artificial) protein, as finding a meaningful starting model is highly improbable. Clustering demonstrates that the models fall essentially into two groups (Fig. 5b): TIM barrels (and modifications thereof; 27 structures) and others (a heterogeneous group, some of which are indeed Rossmann-like folds, see for example the Raptor hits: RaptorX and model_1_RaptorDom; 22 structures). Surprisingly, nine out of the top ten models, according to the

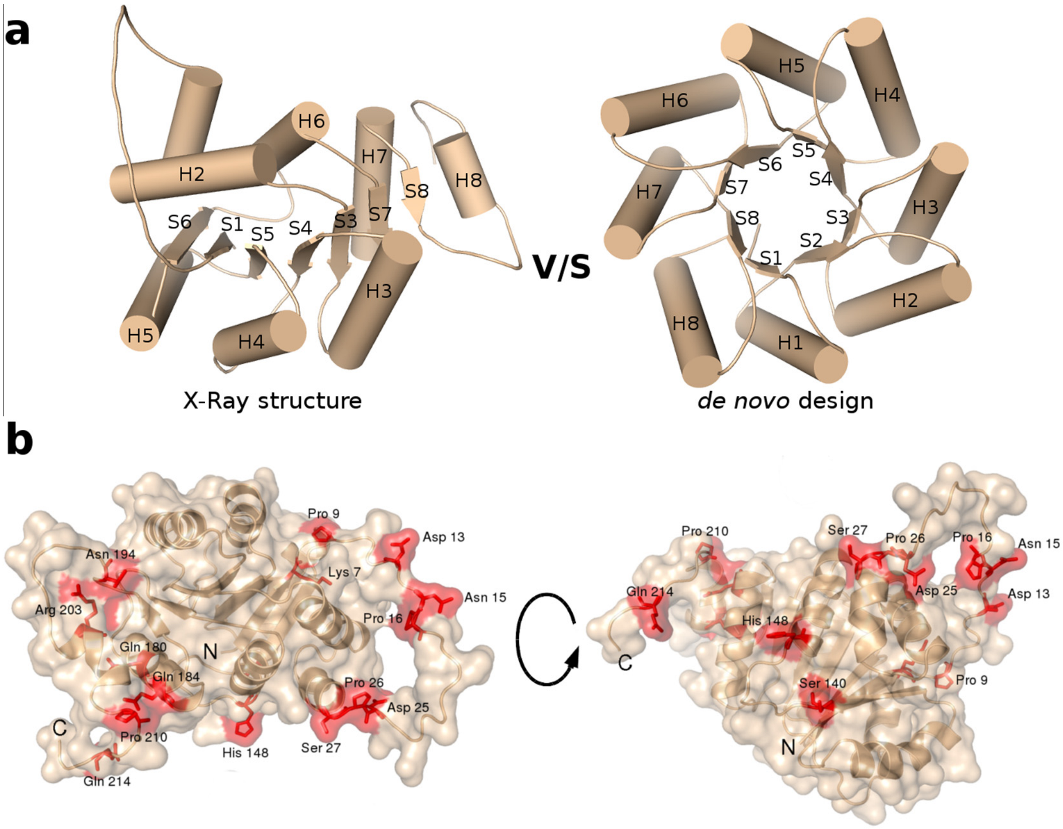


Fig. 4. Structure analysis. a) Comparison of the *de novo* design model with the actual structure of Octarellin V.1. The different folds are shown, and also the expected secondary structure elements; note that the expected strand 2 (S2), helix 1 (H1), and the majority of the helix 8 (H8) are missing in the actual structure. b) Structural location of the mutations obtained by directed evolution. The sixteen mutations obtained after the directed evolution process are highlighted in the final model produced from the X-ray data. Most of the mutations are present in the surface of the protein.

GDT_TS ranking, are TIM(-like)-barrels, an interesting result since the design goal was a TIM-barrel. LGA analysis using the design structure as a reference shows an improvement in the overall result ($<\text{GDT\_TS}>_{\text{X-ray}} = 20 \pm 5$, $<\text{GDT\_TS}>_{\text{Design}} = 29 \pm 14$), the TIM-barrel models being the strongest climbers (see Fig. 5 and Supplementary Table 2). It thus seems that the majority of tested servers (8 out of 13) gravitate more towards our theoretical design than towards the experimentally obtained model. Interestingly, the apparently well designed region of the protein, between amino acids Glu55 and Phe134, is also well predicted by different webservers (Fig. 5c). This region is the main contributor to the fact that TIM barrel models are present at the top of the GDT_TS ranking.

2.5. At which stage could automated structure prediction algorithms fail? Structure prediction of Octarellin V.1 with BCL::Fold

Protein structure prediction algorithms fail for one of two reasons: failure to create the correct topology (sampling) or difficulties in identifying the correct topology by superior energy (scoring). As the precise reason for failure is difficult to assess from automated on-line methods, where only a limited set of models is returned to the user, we created 20,000 models of Octarellin V.1 with the folding algorithm BCL::Fold (Karakaş et al., 2012). BCL::Fold is particularly suited for this test of sampling efficiency, as it assembles secondary structure elements in space, thereby directly enumerating the vast space of possible protein topologies. Encouragingly, BCL::Fold did not sample the $(\beta\alpha)_8$ TIM-barrel fold for Octarellin V.1 but suggested a wide variety of folds with the $(\alpha\beta\alpha)$ sandwich architecture being the most frequent among a number of more complex architectures. The most accurate topology sampled has a RMSD100 value of 7.3 Å, a GDT_TS score of 45.4, for secondary structure elements only, and a N GDT_TS = 30.8 when loops have been added. Because of ambiguous secondary structure predictions, S2 as well as H1, H5 and H8 were missing from this model. All remaining α-helices (H2, H3, H4, H5, H7) agree in location with the consensus X-ray model (Fig. 4a, Supporting Fig. 4). The central β-sheet has a S8||S4||S5||S3||S7||S1||S6 topology, i.e. it is all parallel with β-strands S1, S6, and S8 in identical positions with Octarellin V.1. Further, the contacts between S4||S5 and S3||S7 are correctly predicted (Supporting Fig. 4). The study demonstrates that at least in the case of BCL::Fold current *de novo* structure prediction algorithms fail to sample the Octarellin V.1 fold although they sample topologies that reproduce the majority of the fold's features.

3. Discussion

Protein design, known as the inverse folding problem, involves screening the vastness of protein sequence space to identify candidate sequences, which can fold into a predefined tertiary structure.

Computational design enables us to engineer protein folds not yet observed in nature thereby increasing space of protein template structures available for engineering. This paradigm has led to impressive advances, particularly in designing proteins of around 100 amino acids or less (Koga et al., 2012; Kuhlman et al., 2003), and most recently with a full TIM barrel protein thanks to the internal structural symmetry displayed by this fold (Huang et al., 2016). The sTIM of Huang et al. has been designed using an approach different to that used by us for Octarellin V: they have considered a quarter of the protein and then, thanks to the

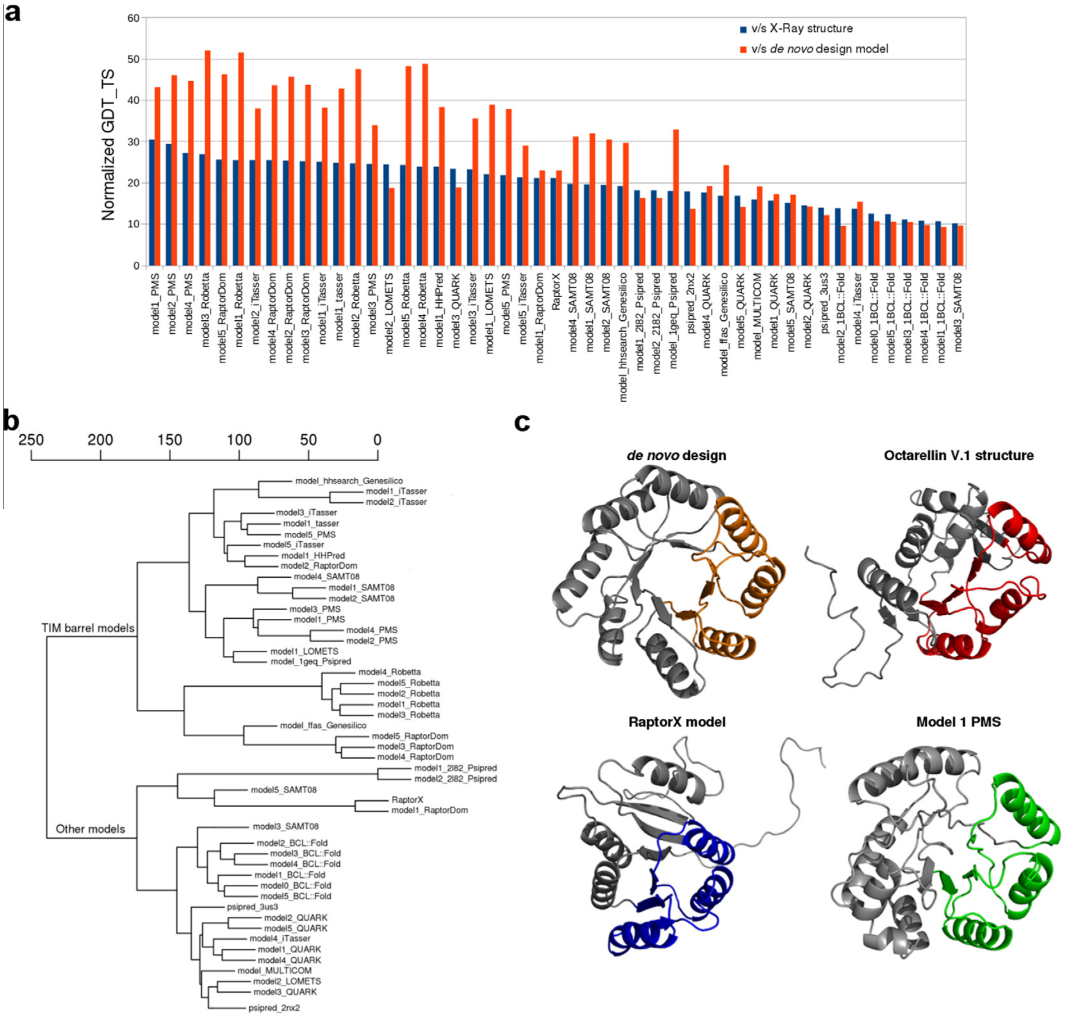


Fig. 5. Automated structure prediction tools evaluated using Octarrelin V.1. a) GDT_TS evaluation of the different models. The 49 models were analyzed against the experimental structure of the Octarellin V.1. They were ranked according to their GDT_TS values (blue bars). A comparison of the models with the *de novo* design model (TIM barrel) is also presented (red bars), showing an increase in the GDT_TS values. b) Dendrogram of the structural clustering of all of the models created from the Octarellin V.1 sequence, using a pairwise euclidean distance protocol. Two main branches are distinguished: models with a TIM-barrel-based fold, and those with other kinds of fold. c) A portion of Octarellin V.1 appears as a "well designed region" (red region) because it matches the *de novo* design (orange region) and because it was well predicted in several models (examples: Raptor X model, blue region, and model 1 from the PMS server, green region).

symmetry, replicated it in the space to get the full protein, instead of a full idealized backbone as we described for our protein back in 2003. Protein design is the most stringent test to validate current knowledge about the protein sequence-structure-function relationship. In particular, large proteins are involved in countless cellular processes, and the capacity to design proteins of any size and structure, and subsequently any function, will be crucial to synthetic biology. A major advance has been accomplished by Brunette et al. (2015) and Koga et al. (2012), who have determined some rules to design proteins and then successfully applied them to design large artificial proteins.

The Octarellin project began many years ago with this central hypothesis: before attempting to engineer a large new fold not existent in nature, it must be possible to *de novo* engineer a large protein fold that is well represented in nature – i.e. it is known that a large and diverse set of sequences exist that adopt this fold. The $(\beta\alpha)_8$ or TIM-barrel fold became focus of the project as it is one of around ten superfolds (Orengo et al., 1994) and is present in a multitude of enzymes with a wide range of functions. This choice offers 1) an opportunity to test protein design protocols on a fold with more than 200 amino acids and 2) the prospect of using a well-designed TIM-barrel protein as an artificial scaffold to create synthetic enzymes or other functional proteins.

In the present work, we describe the structure of Octarellin V.1, which was derived from the *in silico* design Octarellin V through a process of directed evolution. Although the design strategies used for Octarellin V are somewhat outdated given the considerable advances in the field over the last decade, its native-like properties and promising biophysical features made it the most promising starting point among Octarellins designs so far.

Octarellin V.1 shows 16 mutations (93% identity to the parental Octarellin V). Its increased solubility and stability demonstrate the validity of applying directed evolution to an artificial protein to improve these features without changing its structure. Most of the mutations are located at the surface of the experimentally determined structure (Fig. 4b). Interestingly, Octarellin V is proline free, but four prolines appeared in the course of the directed evolution process (L9P, Q16P, L26P and L210P). This is more than the number expected to arise through random mutagenesis. It strongly suggests that the prolines introduced may be at least partly responsible for the improved solubility and stability of Octarellin V.1. This notion is supported as many studies on natural proteins having shown that prolines play a role in both protein solubility (Pande et al., 2005; Steward et al., 2002) and protein stability (Eijsink et al., 2004; Hardy et al., 1993; Jaenicke, 2000; Mainfroid et al., 1996; Vanhove et al., 1996).

Determination of the tertiary structure of Octarellin V.1 was straightforward, thanks to the use of crystallization chaperons. Using nanobodies and αRep proteins, we created stable protein complexes with improved chances of crystallization. Determining the tertiary structure was also facilitated by using the structure of the crystallization chaperons for molecular replacement in order to solve the phase problem. The structure of Octarellin V.1 in the complexes obtained is not complete, but the structures are complementary, showing always the same fold. Experimental SAXS analysis on Octarellin V.1 yields a perfect match with the SAXS curve computed from the tertiary coordinates of Octarellin V.1 (Supporting Fig. 1) confirming that the protein structure has not been altered by the crystallization chaperons. Moreover, the tertiary structure information may explain the low *m* values observed experimentally. These are significantly lower than predicted (i.e. $-21.4\ \mathrm{kJ\ mol^{-1}\ M^{-1}}$ and $-10.4\ \mathrm{kJ\ mol^{-1}\ M^{-1}}$, for urea and GdmCl respectively) from the size of the protein (Myers et al., 1995). Such discrepancy might be the result of unfolding being not fully cooperative (i.e. not two-state), which is inconsistent with our observation that fluorescence and CD data coincide in the presence of urea (and also with the absence of ANS binding in the presence of denaturant, data not shown). Alternatively, the low experimental *m* values suggests that less surface becomes solvent accessible upon Octarellin unfolding. This is in good agreement with data indicating that the native state is actually partially disordered, specifically at the C-term and between S1 and H2 regions. The number of observed peaks in the 2D-HSQC NMR experiment support this observation as well. The observed number of residues in each obtained crystal is consistent as well with the low *m* value: around 160 amino acids are observed in each crystal structure. Considering that both crystallization helpers are stabilizing the Octarellin V.1, it is expected to observe more residues than those observed in solution in the NMR experiment (Fig. 2b).

While the alternating pattern of β-strands and α-helices is largely present and also assembles in the typical βαβ motif, the overall arrangement deviates considerably from a $(\beta\alpha)_8$-barrel. The original vision governing the design of Octarellin was that of a protein with the shortest possible sequence conducive to folding into an idealized TIM barrel with minimalistic loop segments. This design is not particularly forgiving – there is very little margin for error, even at the level of secondary structure: most natural TIMs have at least eight (βα)-units forming a self-closing barrel although rare instances of $(\beta\alpha)_7$, and even $(\beta\alpha)_6$ that form ¾-barrels are known (Ochoa-Leyva et al., 2013).

Nevertheless, Octarellin V and Octarellin V.1 represent an interesting mode of ‘failure’ in protein design that has been observed before: Bharat et al (Bharat et al., 2008) discuss also an attempted $(\beta\alpha)_8$-barrel design by fragment recombination that also succeeded on the secondary structure and βαβ motif level, but failed to achieve the target $(\beta\alpha)_8$-topology. In both cases secondary structure and local tertiary structural features largely agree with the original design. However, the target $(\beta\alpha)_8$-barrel fold is energetically frustrated pushing the protein to adopt an alternative topology that is destabilized and less soluble when compared to typical naturally occurring proteins (Eisenbeis et al., 2012). Arguably, these are cases of being ‘almost right’ or having created a so-called ‘hopeful monster’ (Bharat et al., 2008) where a potentially small number of frustrations prevent the protein from adopting the target $(\beta\alpha)_8$-barrel fold.

In this case, we have here an original tool not just for understanding where the *in silico* design algorithm failed, but also to test *de novo* folding algorithms. If *de novo* folding algorithms are sufficiently accurate to recognize sequences as unlikely to fold into $(\beta\alpha)_8$-barrel fold, unsuccessful designs could be discarded prior to experimental validation. In our case this experiment failed for the most part – a large number of automated algorithms predicted the $(\beta\alpha)_8$-barrel fold. This result is more intriguing if we consider a recent work where it is demonstrated the evolutionary relationship between flavodoxin and TIM-barrel folds (Farías-Rico et al., 2014), considering that the actual structure of Octarellin V.1 is related with the flavodoxin fold. This raises an important question: are protein structure prediction algorithms, even if classified as ‘*de novo*’, biased towards folds over-represented in the PDB such as $(\beta\alpha)_8$-barrels? This question is particularly interesting in the present scenario as the bias cannot be related to the respective proportions of entries for the various folds: there are over two times more Rossmann fold structures (CATH ID: 3.40.50-10922 entries) than $(\beta\alpha)_8$-barrels (CATH ID: 3.20.20-4277 entries) in the Protein Data Bank (PDB).

In the represent case, a more likely type of bias might be linked to the fact that few so-called ‘*de novo*’ folding algorithms are actually *de novo* in the sense that no templates from the PDB are used to construct the model. For example, Rosetta (Gront et al., 2011) assembles the tertiary structure from fragments consisting of up to nine residues. This sequence is long enough to span part of a β-strand, the connecting loop and a turn of a subsequent α-helix and vice versa. Consequently, local sequence similarity to a $(\beta\alpha)_8$-barrel structure in the PDB results in an abundance of fragments from such barrels creating a bias in predicted structures towards a $(\beta\alpha)_8$-barrel fold. Similarly, the methods I-TASSER (Roy et al., 2010) and QUARK (Zhang, 2014) also assemble the tertiary structure from fragments, which are chosen based on sequence and secondary structure similarity. In a strict sense these methods perform comparative modeling combining a large number of templates. We argue that the initial secondary structure prediction step that most servers perform to guide the fold selection process in conjunction with local sequence motifs seen on $(\beta\alpha)_8$-barrel can mislead *de novo* folding algorithms if they reuse fragments from the PDB. This type of bias can even be observed when some of the eight βα motifs cannot be confidently predicted by secondary structure prediction algorithms. Neither psipred (Buchan et al., 2013) nor jpred3 (Cole et al., 2008) predicted a $(\beta\alpha)_n$ sequence, let alone hitting the n = 8 mark.

None of the tested protein structure prediction servers could even remotely predict the structure we observe experimentally. This is reflected in the average GDT_TS value we obtained, i.e. 16, as compared to 28 for the mean value of the most recent CASP10 event (Tai et al., 2014). This demonstrates that Octarellin V.1 is indeed a particularly difficult target for protein structure prediction, perhaps in part for lack of clear homologous at sequence level. Had we performed a blind test, it is more than likely we would have concluded that Octarellin V.1 is a TIM-barrel. This case study thus highlights critical shortcomings in current structural prediction approaches. When contrasting these results obtained by automated servers with a human-guided prediction algorithm that abstains from using fragments from the PDB (BCL::Fold), the correct αβα-fold is frequently sampled with α-helices in the correct location but the topology of the β-sheet deviating form that observed in the experimental models. This finding indicates that in the present case the sampling problem was not yet completely solved – the correct fold was not sampled, only similar ones. The remarkable consensus of different protein structure prediction methods on a topology that significantly deviates from the experimentally determined structure makes Octarellin V.1 an intriguing benchmark protein for protein structure prediction groups to tackle current shortcomings of the algorithms.

In conclusion (Fig. 6), we believe that the number of registered protein folds could be drastically increased by reporting failed protein designs and depositing their experimental structure into the Protein Data Bank. This way, new possible conformations could be studied and those data could help to improve the

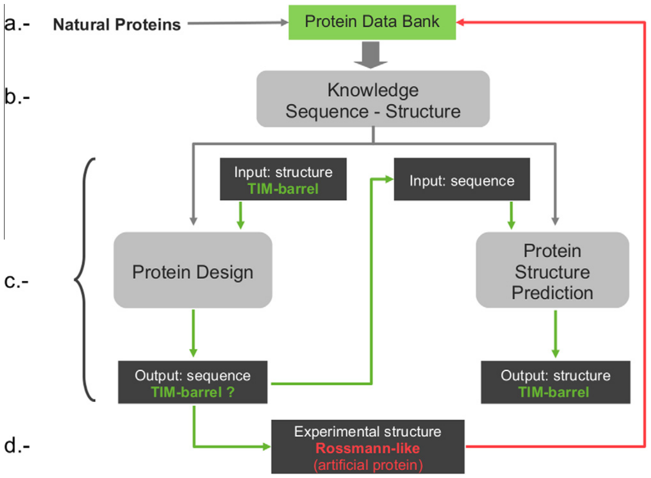


Fig. 6. Combination of protein design with protein structure prediction techniques could help to increase the number of novel structures into the Protein Data Bank. (a) The Protein Data Bank (PDB) mainly consists of natural proteins, and the structural information obtained is translated into knowledge used to understand the sequence-structure relationship (b). This knowledge either used to predict protein structures, using only the sequences of the protein targets, or to design proteins using a backbone as target (c). This last process is completely *in silico*. In our case, we have combined the two processes, using an idealized TIM-barrel backbone to obtain a sequence capable to fold as the target; the obtained sequence has then been used as 'input' to predict its structure, in order to corroborate the design. As a matter of fact, the predicted structure was in most of the cases a TIM-barrel. However, once the experimental structure was solved (d), the actual structure differed from the design and structure predictions. In general, we believe the failed designs can now feed the PDB with new structures/folds enlarging the spectrum of possible conformations, which ones can be taken in account in future algorithms for protein design or protein structure predictions.

parametrization of software which are used to predict protein structure and/or design new proteins.

From this perspective, it seems fundamental that researchers report all their designed proteins, both the successful and failed ones, and deposit their experimentally-determined structure into the PDB. Protein structure prediction based on *de novo* techniques could definitely benefits from the increase of the number of known sequence-fold relationships and from the subsequent improvement of future algorithms.

4. Materials and methods

4.1. Directed evolution by error-prone PCR

Random mutagenesis of the Octarellin gene was performed by error-prone PCR with the *GeneMorph II random Mutagenesis Kit* (Stratagene, USA), performing 20 cycles of amplification with 100 ng target DNA sequence. The Octarellin variants were cloned in fusion with the GFP gene in a modified pET21 vector. We used the GFP-based screening methods described by Waldo (2003) to analyze the libraries. After 8 cycles, we identified a clone with a soluble phenotype inside the bacteria, and we have denominated it Octarellin V.1.

4.2. Cloning, production, and purification of Octarellin

Octarellin V and V.1 were subcloned into the pJB122 expression vector (given by Jonathan Blackburn). Octarellin V was purified from inclusion bodies as it was previously reported (Offredi et al., 2003). Octarellin V.1 was produced in BL21(DE3) at 37 °C after 4 hours of induction (IPTG 1 mM). The protein was purified by ionic exchange and size exclusion chromatographies.

4.3. NMR measurement

Using Octarellin V.1 at 0.2 mM in 50 mM phosphate, pH 7.0, the NMR spectra were measured at 500 MHz 1H frequencies on Bruker Avance spectrometer equipped with a TCI cryoprobe. NMR experiments were performed at 338 °K. The NMR data were processed with TopSpin 2.1 software (Bruker).

4.4. SAXS

SAXS experiments were carried out using the Nanostar instrument (Bruker, Karlsruhe, Germany), with X-rays generated by a rotating anode (Cu Kα, wavelength $\lambda = 1.54$ Å). The scattered X-rays were collected at 20 °C using a 2D position sensitive detector (Vantec) positioned at 710 mm from the sample. SAXS data were averaged and background subtracted using the program package PRIMUS (Konarev et al., 2003). Data were collected in a 50 mM phosphate buffer pH 8.0. The calculated curve was obtained by using the program Crysol (Konarev et al., 2003) from the crystal structure by adding missing residues using the program Modeller (Eswar et al., 2001).

4.5. αRep selection, production and purification

The selection was done as previously described using αRep library 2.1 (Guellouz et al., 2013; Tiouajni et al., 2014) using 40 mg/mL of Octarellin V.1 coated on a micro-titer ELISA plate. Positive αRep genes were sub-cloned in the pQE81L vector (Qiagen) for αRep production and purification.

The protein were produced in *Escherichia coli* BL21(DE3). The purification was performed in two chromatographic steps (IMAC and size exclusion chromatography).

4.6. Nanobody production and purification

The *in vivo* nanobody production was performed by llama immunization (six times) with 1 mg in total of purified Octarellin V.1 over a period of 6 weeks, and the nanobodies were obtained according to Pardon et al. (2014). The *in vitro* production of nanobodies was performed in *E. coli* WK6(Su^-) in TB media supplemented with ampicillin. The purification protocol was the same than used to purify αRep proteins.

4.7. Protein crystallization

The seven complexes nanobody/Octarellin V.1 (10 and 5 mg/mL) plus one complex αRep/Octarellin V.1 (21 and 10 mg/mL) were submitted to crystallization screening using hanging drop vapor diffusion technique with the commercial screening INDEX, Crystal, and Crystal II (Hampton Research); and by sitting drop technique using the commercial screenings Clear Strategy I and II, JSCG Plus (Molecular Dimensions), and JBSC Basic HTSL (Jena Bioscience) assisted by Crystal Phoenix robot (Art Robbins). All the screenings were performed at 18 °C.

4.8. Structure resolution

All diffraction data were collected at 100 K using synchrotron radiation, and data sets of diffracting crystals were processed with the XDS suite (Kabsch, 2010) using Xscale for scaling and merging of the reflections. Initial data quality was assessed in phenix.xtriage (Adams et al., 2010). Phase information was obtained by molecular replacement with the PHASER program (McCoy et al., 2007) in the CCP4 software package (Winn et al., 2011). The Nanoboy/Octarellin V.1 data sets were phased using a CDR-loops truncated nanobody (PDB: 1MVF) as search model. For the phasing of the αRep/Octarellin V.1 dataset we used a single repeat of an artificial alpha helicoidal repeat protein (PDB: 3LTJ). The Phenix.AutoBuild program (Adams et al., 2010) was used for automated model building. Model building was finalized by manual building cycles in Coot (Emsley and Cowtan, 2004), alternated with refinement using Phenix.Refine (Adams et al., 2010). Temperature factors of the P21 and I4122 Nanobody/Octarellin V.1 structures were refined through TLS refinement using 16 and 9 groups, respectively. The amount of TLS groups used during refinement was determined by the TLSMD server (Painter and Merritt, 2006). The temperature factors of the αRep/Octarellin V.1 structure were isotropically refined. The obtained models were validated with the Molprobity server (Davis et al., 2007). All structure figures were prepared in PyMOL (http://www.pymol.org/). Data collection and processing statistics are summarized in Table 1.

4.9. Tertiary structure prediction using webservers

The webservers:

- Robetta (Bradley et al., 2005) (http://www.robetta.org),
- iTasser (Roy et al., 2010) (http://zhanglab.ccmb.med.umich.edu/I-TASSER/),
- PMS (Joo et al., 2014) (http://lee.kias.re.kr/~protein/wiki/doku.php?id=model:submit),
- RaptorX (Källberg et al., 2012) (http://raptorx.uchicago.edu/StructurePrediction/predict/),
- GeneSilico (Kurowski and Bujnicki, 2003) (https://genesilico.pl/meta2/),
- HHPred (Söding et al., 2005) (http://toolkit.tuebingen.mpg.de/hhpred),
- Lomets (Wu and Zhang, 2007) (http://zhanglab.ccmb.med.umich.edu/LOMETS/),
- Quark (Xu and Zhang, 2012) (http://zhanglab.ccmb.med.umich.edu/QUARK/),
- SamT-08 (Karplus, 2009) (https://compbio.soe.ucsc.edu/SAM_T08/T08-query.html),
- PsiPred (Buchan et al., 2013) (http://bioinf.cs.ucl.ac.uk/psipred/),
- Multicom (Cheng, 2008) (http://casp.rnet.missouri.edu/multicom_3d.html),
- Tasser (Zhou and Skolnick, 2009) (http://cssb.biology.gatech.edu/skolnick/webservice/TASSER/index.html)
- BCL::Fold (Karakaş et al., 2012) (http://www.meilerlab.org/index.php/servers/show?s_id=12),

were used to model the 3D structure of the Octarellin V.1 using the default parameters. The best five models (when it was possible) were considered by further analysis. In total, 49 models were considered in this study.

4.10. Models comparison and clustering

The 49 models produced by the protein structure prediction webservers were compared structurally with the actual structure of Octarellin V.1 or the *in silico* designed Octarellin V using the LGA approximation (Zemla, 2003) (http://proteinmodel.org/AS2TS/LGA_list/lga_pdblist.html), and the models were ranked according with their GDT_TS value and normalized GDT_TS value. The normalized GDT_TS value is obtained through the normalization of the GDT_TS by the number of amino acids modeled. All the models were clustered using the STRALCP tool (Zemla et al., 2007) (http://proteinmodel.org/AS2TS/STRALCP/index.html) and a dendrogram with this information was created to well visualize the relationship among the different models.

Acknowledgments

The present work was supported in part by the European Space Agency under contract No. ESA AO-2004-070. M.F. acknowledges Wallonie-Bruxelles International (WBI) and Belgian Science Policy (BELSPO) for postdoctoral fellowships. We also thank Piriscucha D'Esneux for streak seeding assistance, Jordane Delbecq for technical assistance and Aline Lausberg for her comments regarding the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2016.05.004.

References

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C., Zwart, P.H., 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221. http://dx.doi.org/10.1107/S0907444909052925.

Beauregard, M., Goraj, K., Goffin, V., Heremans, K., Goormaghtigh, E., Ruysschaert, J.-M., Martial, J.A., 1991. Spectroscopic investigation of structure in octarellin (a de novo protein designed to adopt the α/β-barred packing). Protein Eng. 4, 745–749. http://dx.doi.org/10.1093/protein/4.7.745.

Bharat, T.A.M., Eisenbeis, S., Zeth, K., Höcker, B., 2008. A βα-barrel built by the combination of fragments from different folds. Proc. Natl. Acad. Sci. 105, 9942–9947. http://dx.doi.org/10.1073/pnas.0802202105.

Bradley, P., Misura, K.M.S., Baker, D., 2005. Toward high-resolution de novo structure prediction for small proteins. Science 309, 1868–1871. http://dx.doi.org/10.1126/science.1113801.

Brunette, T.J., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D.C., Tsutakawa, S.E., Hura, G.L., Tainer, J.A., Baker, D., 2015. Exploring the repeat protein universe through computational protein design. Nature. http://dx.doi.org/10.1038/nature16162, advance online publication.

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., Jones, D.T., 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 41, W349–W357. http://dx.doi.org/10.1093/nar/gkt381.

Carbonell, P., Trosset, J.-Y., 2015. Computational protein design methods for synthetic biology. Methods Mol. Biol. Clifton NJ 1244, 3–21. http://dx.doi.org/10.1007/978-1-4939-1878-2_1.

Chaput, J.C., Woodbury, N.W., Stearns, L.A., Williams, B.A., 2008. Creating protein biocatalysts as tools for future industrial applications. Expert Opin. Biol. Ther. 8, 1087–1098. http://dx.doi.org/10.1517/14712598.8.8.1087.

Cheng, J., 2008. A multi-template combination algorithm for protein comparative modeling. BMC Struct. Biol. 8, 18. http://dx.doi.org/10.1186/1472-6807-8-18.

Cole, C., Barber, J.D., Barton, G.J., 2008. The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 36, W197–W201. http://dx.doi.org/10.1093/nar/gkn238.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., Richardson, D.C., 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 35, W375–W383. http://dx.doi.org/10.1093/nar/gkm216.

Dobson, C.M., 2003. Protein folding and misfolding. Nature 426, 884–890.

Domanska, K., Vanderhaegen, S., Srinivasan, V., Pardon, E., Dupeux, F., Marquez, J.A., Giorgetti, S., Stoppini, M., Wyns, L., Bellotti, V., Steyaert, J., 2011. Atomic structure of a nanobody-trapped domain-swapped dimer of an amyloidogenic β2-microglobulin variant. Proc. Natl. Acad. Sci. 108, 1314–1319. http://dx.doi.org/10.1073/pnas.1008560108.

Eijsink, V.G.H., Bjørk, A., Gåseidnes, S., Sirevåg, R., Synstad, B., van den Burg, B., Vriend, G., 2004. Rational engineering of enzyme stability. J. Biotechnol. 113, 105–120. http://dx.doi.org/10.1016/j.jbiotec.2004.03.026.

Eijsink, V.G.H., Gåseidnes, S., Borchert, T.V., van den Burg, B., 2005. Directed evolution of enzyme stability. Biomol. Eng. 22, 21–30. http://dx.doi.org/10.1016/j.bioeng.2004.12.003.

Eisenbeis, S., Proffitt, W., Coles, M., Truffault, V., Shanmugaratnam, S., Meiler, J., Höcker, B., 2012. Potential of fragment recombination for rational design of proteins. J. Am. Chem. Soc. 134, 4019–4022. http://dx.doi.org/10.1021/ja211657k.

Emsley, P., Cowtan, K., 2004. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132. http://dx.doi.org/10.1107/S0907444904019158.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M., Pieper, U., Sali, A., 2001. Comparative protein structure modeling using MODELLER. In: Current Protocols in Protein Science. John Wiley & Sons, Inc.

Farías-Rico, J.A., Schmidt, S., Höcker, B., 2014. Evolutionary relationship of two ancient protein superfolds. Nat. Chem. Biol. 10, 710–715. http://dx.doi.org/10.1038/nchembio.1579.

Ferrandez, Y., Dezi, M., Bosco, M., Urvoas, A., Valerio-Lepiniec, M., Bon, C.L., Giusti, F., Broutin, I., Durand, G., Polidori, A., Popot, J.-L., Picard, M., Minard, P., 2014. Amphipol-mediated screening of molecular orthoses specific for membrane protein targets. J. Membr. Biol. 247, 925–940. http://dx.doi.org/10.1007/s00232-014-9707-3.

Figueroa, M., Oliveira, N., Lejeune, A., Kaufmann, K.W., Dorr, B.M., Matagne, A., Martial, J.A., Meiler, J., Van de Weerdt, C., 2013. Octarellin VI: using rosetta to design a putative artificial (β/α)8 protein. PLoS One 8, e71858. http://dx.doi.org/10.1371/journal.pone.0071858.

Figueroa, M., Vandenameele, Julie, Goormaghtigh, Erik, Valerio-Lapiniec, Marie, Minard, Philippe, Matagne, André, Van de Weerdt, Cécile, n.d. Biophysical characterization of the artificial protein Octarellin V. 1 and binding test with its X-ray helpers. Data in Brief.

Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.-M., Wilson, I.A., Baker, D., 2011. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332, 816–821. http://dx.doi.org/10.1126/science.1202617.

Fortenberry, C., Bowman, E.A., Proffitt, W., Dorr, B., Combs, S., Harp, J., Mizoue, L., Meiler, J., 2011. Exploring symmetry as an avenue to the computational design of large protein domains. J. Am. Chem. Soc. 133, 18026–18029. http://dx.doi.org/10.1021/ja2051217.

Goraj, K., Renard, A., Martial, J.A., 1990. Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the α/β-barrel proteins. Protein Eng. 3, 259–266. http://dx.doi.org/10.1093/protein/3.4.259.

Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E.M., Baker, D., 2011. Generalized fragment picking in Rosetta: design, protocols and applications. PLoS One 6, e23294. http://dx.doi.org/10.1371/journal.pone.0023294.

Guellouz, A., Valerio-Lepiniec, M., Urvoas, A., Chevrel, A., Graille, M., Fourati-Kammoun, Z., Desmadril, M., van Tilbeurgh, H., Minard, P., 2013. Selection of specific protein binders for pre-defined targets from an optimized library of artificial helicoidal repeat proteins (alphaRep). PLoS One 8, e71512. http://dx.doi.org/10.1371/journal.pone.0071512.

Hardy, F., Vriend, G., Veltman, O.R., van der Vinne, B., Venema, G., Eijsink, V.G.H., 1993. Stabilization of *Bacillus stearothermophilus* neutral protease by introduction of prolines. FEBS Lett. 317, 89–92. http://dx.doi.org/10.1016/0014-5793(93)81497-N.

Houbrechts, A., Moreau, B., Abagyan, R., Mainfroid, V., Préaux, G., Lamproye, A., Poncin, A., Goormaghtigh, E., Ruysschaert, J.-M., Martial, J.A., Goraj, K., 1995. Second-generation octarellins: two new de novo (β/α)8 polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability. Protein Eng. 8, 249–259. http://dx.doi.org/10.1093/protein/8.3.249.

Huang, P.-S., Feldmeier, K., Parmeggiani, F., Fernandez Velasco, D.A., Höcker, B., Baker, D., 2016. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat. Chem. Biol. 12, 29–34. http://dx.doi.org/10.1038/nchembio.1966.

Huang, Y.J., Mao, B., Aramini, J.M., Montelione, G.T., 2014. Assessment of template-based protein structure predictions in CASP10. Proteins Struct. Funct. Bioinf. 82, 43–56. http://dx.doi.org/10.1002/prot.24488.

Jaenicke, R., 2000. Stability and stabilization of globular proteins in solution. J. Biotechnol. 79, 193–203. http://dx.doi.org/10.1016/S0168-1656(00)00236-4.

Joo, K., Lee, J., Sim, S., Lee, S.Y., Lee, K., Heo, S., Lee, I.-H., Lee, S.J., Lee, J., 2014. Protein structure modeling for CASP10 by multiple layers of global optimization. Proteins Struct. Funct. Bioinf. 82, 188–195. http://dx.doi.org/10.1002/prot.24397.

Kabsch, W., 2010. XDS. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132. http://dx.doi.org/10.1107/S0907444909047337.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., Xu, J., 2012. Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. 7, 1511–1522. http://dx.doi.org/10.1038/nprot.2012.085.

Karakaş, M., Woetzel, N., Staritzbichler, R., Alexander, N., Weiner, B.E., Meiler, J., 2012. BCL: fold – de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS One 7, e49240. http://dx.doi.org/10.1371/journal.pone.0049240.

Karplus, K., 2009. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res. 37, W492–W497. http://dx.doi.org/10.1093/nar/gkp403.

Karplus, P.A., Diederichs, K., 2012. Linking crystallographic model and data quality. Science 336, 1030–1033. http://dx.doi.org/10.1126/science.1218231.

Keedy, D.A., Williams, C.J., Headd, J.J., Arendall, W.B., Chen, V.B., Kapral, G.J., Gillespie, R.A., Block, J.N., Zemla, A., Richardson, D.C., Richardson, J.S., 2009. The other 90% of the protein: assessment beyond the Cαs for CASP8 template-based and high-accuracy models. Proteins Struct. Funct. Bioinf. 77, 29–49. http://dx.doi.org/10.1002/prot.22551.

Khersonsky, O., Röthlisberger, D., Dym, O., Albeck, S., Jackson, C.J., Baker, D., Tawfik, D.S., 2010. Evolutionary optimization of computationally designed enzymes: kemp eliminases of the KE07 series. J. Mol. Biol. 396, 1025–1042. http://dx.doi.org/10.1016/j.jmb.2009.12.031.

Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T.B., Montelione, G.T., Baker, D., 2012. Principles for designing ideal protein structures. Nature 491, 222–227. http://dx.doi.org/10.1038/nature11600.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., Svergun, D.I., 2003. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J. Appl. Crystallogr. 36, 1277–1282. http://dx.doi.org/10.1107/S0021889803012779.

Korotkov, K.V., Pardon, E., Steyaert, J., Hol, W.G.J., 2009. Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody. Structure 17, 255–265. http://dx.doi.org/10.1016/j.str.2008.11.011.

Kryshtafovych, A., Fidelis, K., Moult, J., 2014. CASP10 results compared to those of previous CASP experiments. Proteins Struct. Funct. Bioinf. 82, 164–174. http://dx.doi.org/10.1002/prot.24448.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368. http://dx.doi.org/10.1126/science.1089427.

Kurowski, M.A., Bujnicki, J.M., 2003. GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31, 3305–3307. http://dx.doi.org/10.1093/nar/gkg557.

Mainfroid, V., Mande, S.C., Hol, W.G.J., Martial, J.A., Goraj, K., 1996. Stabilization of human triosephosphate isomerase by improvement of the stability of individual α-helices in dimeric as well as monomeric forms of the protein. Biochemistry (Mosc.) 35, 4110–4117. http://dx.doi.org/10.1021/bi952692n.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., Read, R.J., 2007. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674. http://dx.doi.org/10.1107/S0021889807021206.

Myers, J.K., Nick Pace, C., Martin Scholtz, J., 1995. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci. 4, 2138–2148. http://dx.doi.org/10.1002/pro.5560041020.

Ochoa-Leyva, A., Montero-Morán, G., Saab-Rincón, G., Brieba, L.G., Soberón, X., 2013. Alternative splice variants in TIM barrel proteins from human genome correlate with the structural and evolutionary modularity of this versatile protein fold. PLoS One 8, e70582. http://dx.doi.org/10.1371/journal.pone.0070582.

Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A., Van de Weerdt, C., Hoch, J., Prosperi, C., François, J., Mayo, S., Martial, J., 2003. De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure. J. Mol. Biol. 325, 163–174. http://dx.doi.org/10.1016/S0022-2836(02)01206-8.

Orengo, C.A., Jones, D.T., Thornton, J.M., 1994. Protein superfamilies and domain superfolds. Nature 372, 631–634. http://dx.doi.org/10.1038/372631a0.

Pabo, C., 1983. Molecular technology: designing proteins and peptides. Nature 301, 200. http://dx.doi.org/10.1038/301200a0.

Painter, J., Merritt, E.A., 2006. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr. Sect. D 62, 439–450. http://dx.doi.org/10.1107/S0907444906005270.

Pande, A., Annunziata, O., Asherie, N., Ogun, O., Benedek, G.B., Pande, J., 2005. Decrease in protein solubility and cataract formation caused by the Pro23 to Thr mutation in human γD-crystallin. Biochemistry (Mosc.) 44, 2491–2500. http://dx.doi.org/10.1021/bi0479611.

Pardon, E., Laeremans, T., Triest, S., Rasmussen, S.G.F., Wohlkönig, A., Ruf, A., Muyldermans, S., Hol, W.G.J., Kobilka, B.K., Steyaert, J., 2014. A general protocol for the generation of nanobodies for structural biology. Nat. Protoc. 9, 674–693. http://dx.doi.org/10.1038/nprot.2014.039.

Parmeggiani, F., Pellarin, R., Larsen, A.P., Varadamsetty, G., Stumpp, M.T., Zerbe, O., Caflisch, A., Plückthun, A., 2008. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J. Mol. Biol. 376, 1282–1304. http://dx.doi.org/10.1016/j.jmb.2007.12.014.

Rao, S.T., Rossmann, M.G., 1973. Comparison of super-secondary structures in proteins. J. Mol. Biol. 76, 241–256. http://dx.doi.org/10.1016/0022-2836(73)90388-4.

Rasmussen, S.G.F., Choi, H.-J., Fung, J.J., Pardon, E., Casarosa, P., Chae, P.S., DeVree, B.T., Rosenbaum, D.M., Thian, F.S., Kobilka, T.S., Schnapp, A., Konetzki, I., Sunahara, R.K., Gellman, S.H., Pautsch, A., Steyaert, J., Weis, W.I., Kobilka, B.K., 2011. Structure of a nanobody-stabilized active state of the β2 adrenoceptor. Nature 469, 175–180. http://dx.doi.org/10.1038/nature09648.

Reetz, M.T., 2007. Controlling the selectivity and stability of proteins by new strategies in directed evolution: the case of organocatalytic enzymes. Ernst Schering Found. Symp. Proc., 321–340

Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., Albeck, S., Houk, K.N., Tawfik, D.S., Baker, D., 2008. Kemp elimination catalysts by computational enzyme design. Nature 453, 190–195. http://dx.doi.org/10.1038/nature06879.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725–738.

Söding, J., Biegert, A., Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248. http://dx.doi.org/10.1093/nar/gki408.

Steward, A., Adhya, S., Clarke, J., 2002. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. J. Mol. Biol. 318, 935–940. http://dx.doi.org/10.1016/S0022-2836(02)00184-5.

Tai, C.-H., Bai, H., Taylor, T.J., Lee, B., 2014. Assessment of template-free modeling in CASP10 and ROLL. Proteins Struct. Funct. Bioinf. 82, 57–83. http://dx.doi.org/10.1002/prot.24470.

Tao, H., Cornish, V.W., 2002. Milestones in directed enzyme evolution. Curr. Opin. Chem. Biol. 6, 858–864. http://dx.doi.org/10.1016/S1367-5931(02)00396-4.

Tiouajni, M., Durand, D., Blondeau, K., Graille, M., Urvoas, A., Valerio-Lepiniec, M., Guellouz, A., Aumont-Nicaise, M., Minard, P., van Tilbeurgh, H., 2014. Structural and functional analysis of the fibronectin-binding protein FNE from *Streptococcus equi* spp. equi. FEBS J. 281, 5513–5531. http://dx.doi.org/10.1111/febs.13092.

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D.C., van Tilbeurgh, H., Desmadril, M., Minard, P., 2010. Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats. J. Mol. Biol. 404, 307–327. http://dx.doi.org/10.1016/j.jmb.2010.09.048.

Vanhove, M., Raquet, X., Palzkill, T., Pain, R.H., Frère, J.-M., 1996. The rate-limiting step in the folding of the cis-Pro167Thr mutant of TEM-1 β-lactamase is the trans to cis isomerization of a non-proline peptide bond. Proteins Struct. Funct. Bioinf. 25, 104–111. http://dx.doi.org/10.1002/(SICI)1097-0134(199605)25:1<104::AID-PROT8>3.0.CO;2-J.

Waldo, G.S., 2003. Genetic screens and directed evolution for protein solubility. Curr. Opin. Chem. Biol. 7, 33–38. http://dx.doi.org/10.1016/S1367-5931(02)00017-0.

Ward, T.R., 2008. Artificial enzymes made to order: combination of computational design and directed evolution. Angew. Chem. Int. Ed. 47, 7802–7803. http://dx.doi.org/10.1002/anie.200802865.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., Wilson, K.S., 2011. Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235–242. http://dx.doi.org/10.1107/S0907444910045749.

Wu, S., Zhang, Y., 2007. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35, 3375–3382. http://dx.doi.org/10.1093/nar/gkm251.

Xu, D., Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins Struct. Funct. Bioinf. 80, 1715–1735. http://dx.doi.org/10.1002/prot.24065.

Zemla, A., 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31, 3370–3374. http://dx.doi.org/10.1093/nar/gkg571.

Zemla, A., Geisbrecht, B., Smith, J., Lam, M., Kirkpatrick, B., Wagner, M., Slezak, T., Zhou, C.E., 2007. STRALCP—structure alignment-based clustering of proteins. Nucleic Acids Res. 35, e150. http://dx.doi.org/10.1093/nar/gkm1049.

Zhang, Y., 2014. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins Struct. Funct. Bioinf. 82, 175–187. http://dx.doi.org/10.1002/prot.24341.

Zhou, H., Skolnick, J., 2009. Protein structure prediction by Pro-Sp3-TASSER. Biophys. J. 96, 2119–2127. http://dx.doi.org/10.1016/j.bpj.2008.12.3898.