Journal of Structural Biology 195 (2016) 19–30

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/

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.).