Woolfson, D. N., Baker, E. G., & Bartlett, G. J. (2017). How do miniproteins fold? *Science*, *357*(6347), 133-134. https://doi.org/10.1126/science.aan6864

Protein folding and design: miniproteins step up

Derek N Woolfson[1,2,3], Emily G Baker[1], and Gail J Bartlett[1]

[1]School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK

[2]School of Biochemistry, University of Bristol, Medical Sciences Building, University Walk, Bristol BS8 1TD, UK

[3]Bristol BioDesign Institute, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

Massively high-throughput design, construction and analysis of short protein sequences opens new possibilities for probing sequence-to-structure/stability relationships in proteins.

One aspect of the long-standing protein-folding problem (*1*) asks: *how does the one-dimensional amino-acid sequence, or primary structure, of a protein chain determine and maintain its three-dimensional folded state*? This is an important question for several reasons: First, it is fundamental science that tests our understanding of how multiple non-covalent interactions conspire to assemble and stabilise complicated and fascinating biomolecular structures. Second, many natural proteins remain refractory to full experimental scrutiny and, despite the considerable success of homology modelling, improved methods are needed for predicting protein structure and function from sequence. Third, a better understanding of protein folding and stability will lead to more successful *de novo* protein designs, allowing us to delve into the dark matter of protein space (*2*); that is, to design entirely new protein structures not presented to us by biology. On page XXX of this issue, Rocklin and a team led by David Baker in Seattle describe parallel protein design on a massive and unprecedented scale [REF]. This extremely impressive study delivers thousands of variants of foregoing and newly designed so-called miniproteins to address the problem of what sequences specify and stabilise these structures. It opens considerable possibilities for protein folding and design in the future.

Miniproteins are polypeptides of up to 40 – 50 residues with stable tertiary structures, or folds, which comprise a limited number of secondary structure elements (*e.g.*, α helices and β strands). By contrast, their larger relatives have hundreds of amino acids often arranged in complex three-dimensional structures. Thus, miniproteins simplify the protein-folding problem, and potentially allow in-depth examinations of sequence-to-structure/stability relationships without complications from larger protein contexts. Unfortunately, only a few miniproteins that are stable without covalent cross-links or stabilising metal ions are currently available for such studies (*3*). Perhaps this is about to change?

In the new work, Baker's team cleverly combine high-throughput DNA synthesis and cloning (*4, 5*) with methods for selecting stably folded proteins (*6-8*). They implement the latter by expressing libraries of miniproteins on the surface of yeast; tagging the displayed proteins with a fluorescent dye; and discriminating between stable and unstable folds through their ability to resist or succumb to protease treatment, respectively. Proteins that survive are rescued by FACS (fluorescent assisted cell sorting) and then identified by deep sequencing. However, the team's experiments give more than a Boolean yes/no measure of protein resilience: they provide a semi-quantitative measure of stability.

To demonstrate the approach, the authors first apply their method to many variants of a small number of known miniproteins. With the method established, attention is turned to four classes of *de novo* miniproteins, which they design computationally using Rosetta (*9*);

namely, ααα, βαββ, αββα and ββαββ folds, where each Greek letter represents an α helix or a β strand in the peptide string. To cover swaths of sequence space, the team create diverse libraries with minimal sequence identity between members.

Iterative rounds of protease selection and stability scoring are used, with different hypotheses being tested, and tweaks to the design methods and protocols being introduced at each stage. The value of these tweaks is apparent from the improved success rate—*i.e*., the proportion of stable proteins in the starting library—which reaches a staggering 87% for one of the targets. Interestingly, however, both the initial and final design success rates depended critically on the fold being targeted, with the ααα fold proving "easiest" to optimise and the αββα fold the most difficult.

After analyses of the many thousands of sequences tested across these and other miniprotein folds, several key sequence and structural features emerge from this impressive study. First, a long-established basic tenet of protein folding and design shines through: namely, the importance of burying nonpolar surfaces. This is not surprising, but Rocklin and colleagues quantify this showing that stable variants require >30 Å^2/residue of buried hydrocarbon. Second, the initial computational designs based on fragments that most closely matched known fragments from the Protein Data Bank of known protein structures fare the best in the selection studies and give the most stable sequences. This could be a consequence of using Rosetta to achieve the design frameworks, as it is a fragment-based design approach. In future, it will be interesting to see how starting points from parametric and other design approaches perform (*10-12*) .

Interestingly and importantly, one relationship not included or tweaked during the iterative process—it simply emerges from the analysis of the final dataset, and is present upon re-examination of the foregoing datasets—is the importance of having charge side chains at the termini of the α helices that oppose the terminal partial charges of the helices. This concurs with studies of model peptides that form free-standing α helices in solution, where it is attributed to local capping effects of the helices rather than stabilising any helix macrodipole *per se* (*13*).

For such an impressive and expansive piece of work it seems churlish to ask: *what hasn't this study done?* Nonetheless, there are gaps to fill and more steps to take. Notably—and although many of sequences for the target designs have been characterised by circular dichroism spectroscopy, size-exclusion chromatography, thermal and chemical denaturation, and a small number of structures have been verified by nuclear magnetic resonance spectroscopy—more high-resolution structural details would be welcome, for instance from X-ray crystallography.

Such structures would allow the garnered sequence-to-stability correlations to be rationalised in terms of specific non-covalent interactions that likely underlie them. For example, the study points to stabilizing roles for aromatic residues at surface-exposed sites of α helices and β strands in miniproteins, which hint at non-covalent interactions particular to this class of side chain. On this point, another monomeric miniprotein, PPα, with a compact polyproline II helix-turn-α helix structure, has recently been designed, characterised and interrogated in detail (*14*). A key determinant of PPα's stability comes from intimate CH-π interactions between tyrosine residues of the α helix and proline residues of the buttressing polyproline II helix. Studying the role and interplay of these and other non-covalent interactions will be critical for completing the cycle from computational design, experimental testing, and iteration of the design methods.

Have these studies solved the problem-folding problem? In a word, no. However, Rocklin *et al*. have taken high-throughput, data-driven protein design, selection and optimization to

new heights. There remains plenty to be done to define more sequence-to-structure relationships or rules for protein folding, and then to understand these in terms of underlying non-covalent interactions. Nonetheless, this work on miniproteins brings us closer to solving these aspects of the protein-folding problem. Moreover, a combination of high-throughput studies of sequence-to-structure/stability relationships described by Rocklin *et al.* and drilled-down, fully quantitative examinations of the non-covalent interactions could well furnish us with such an understanding. In turn, this will facilitate better engineering of natural proteins and steps in the *de novo* exploration of the dark matter of protein structures.

References:

1. K. A. Dill, J. L. MacCallum, The Protein-Folding Problem, 50 Years On. *Science* **338**, 1042-1046 (2012).
2. W. R. Taylor, V. Chelliah, S. M. Hollup, J. T. MacDonald, I. Jonassen, Probing the "Dark Matter" of Protein Fold Space. *Structure* **17**, 1244-1252 (2009).
3. S. H. Gellman, D. N. Woolfson, Mini-proteins Trp the light fantastic. *Nat Struct Biol* **9**, 408-410 (2002).
4. S. Kosuri, G. M. Church, Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**, 499-507 (2014).
5. M. G. F. Sun, M. H. Seo, S. Nim, C. Corbi-Verge, P. M. Kim, Protein engineering by highly parallel screening of computationally designed variants. *Sci Adv* **2**, (2016).
6. P. Kristensen, G. Winter, Proteolytic selection for protein folding using filamentous bacteriophages. *Fold Des* **3**, 321-328 (1998).
7. V. Sieber, A. Pluckthun, F. X. Schmid, Selecting proteins with improved stability by a phage-based method. *Nat Biotechnol* **16**, 955-960 (1998).
8. M. D. Finucane, M. Tuna, J. H. Lees, D. N. Woolfson, Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry-Us* **38**, 11604-11612 (1999).
9. R. Das, D. Baker, Macromolecular modeling with Rosetta. *Annu Rev Biochem* **77**, 363-382 (2008).
10. P. S. Huang *et al.*, High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481-485 (2014).
11. A. R. Thomson *et al.*, Computational design of water-soluble alpha-helical barrels. *Science* **346**, 485-488 (2014).
12. T. J. Brunette *et al.*, Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580-584 (2015).
13. E. G. Baker *et al.*, Local and macroscopic electrostatic interactions in single alpha-helices. *Nat Chem Biol* **11**, 221-U292 (2015).
14. E. G. Baker *et al.*, Engineering protein stability with atomic precision in a monomeric miniprotein. *Nat Chem Biol*, (2017). DOI: 10.1038/nchembio.2380

Acknowledgements

DNW and EGB are supported by a BBSRC/ERASynBio grant (BB/M005615/1); DNW and GJB are supported by the ERC (340764); and DNW holds a Royal Society Wolfson Research Merit Award (WM140008).

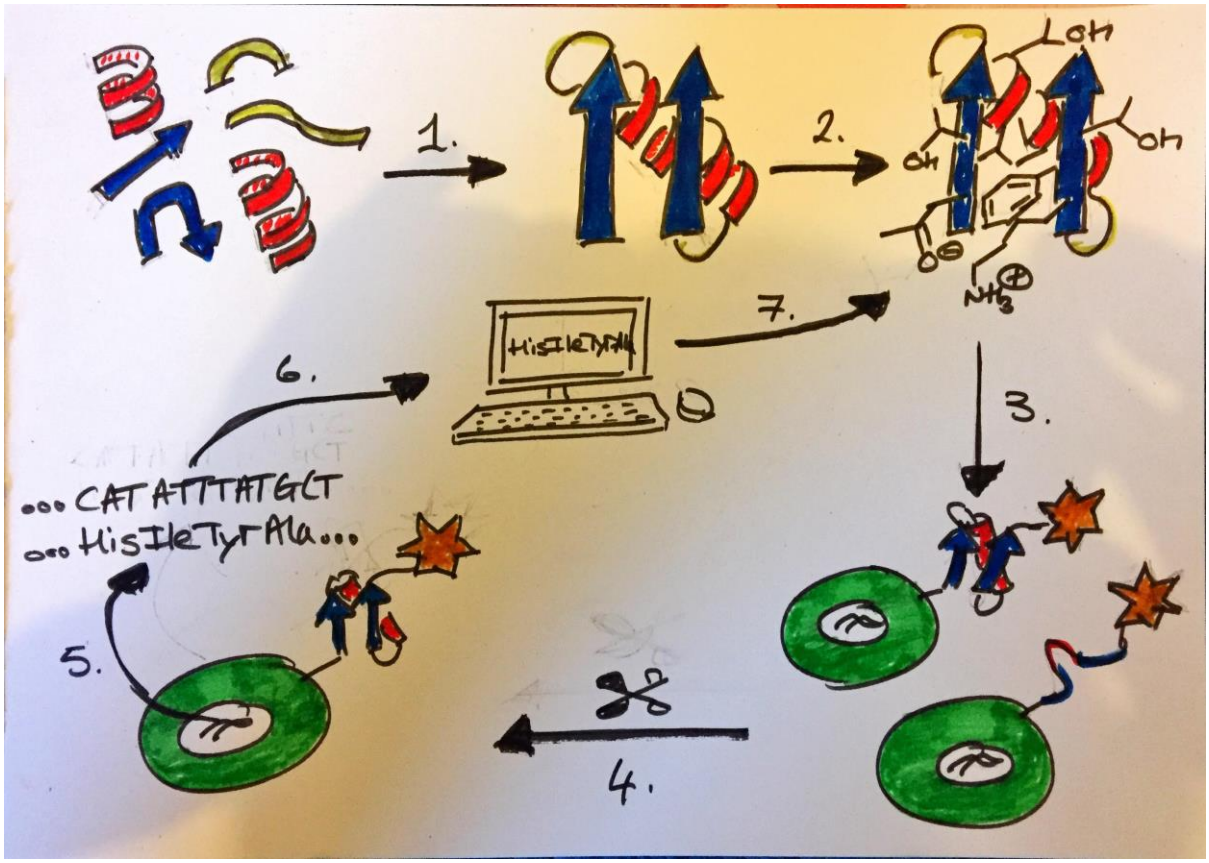

Figure: A new cycle for protein design. 1. Miniprotein structures are designed *in silico* using a fragment-based approach in Rosetta. 2. Libraries of sequences are generated to best fit these structures. 3. The libraries are realized experimentally *via* high-throughput DNA synthesis and cloning, and the resulting proteins are expressed on the surface of yeast. 4. Stable variants are selected based on resistance to treatment with protease. 5&6. The sequences returned are analysed to glean sequence-to-stability relationships, which are fed back into the design cycle (7).