Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

De novo protein design: how do we expand into the universe of possible protein structures?

Derek N Woolfson[1,2,3], Gail J Bartlett[1], Antony J Burton[1], Jack W Heal[1], Ai Niitsu[1], Andrew R Thomson[1] and Christopher W Wood[1,2]

Protein scientists are paving the way to a new phase in protein design and engineering. Approaches and methods are being developed that could allow the design of proteins beyond the confines of natural protein structures. This possibility of designing entirely new proteins opens new questions: What do we build? How do we build into protein-structure space where there are few, if any, natural structures to guide us? To what uses can the resulting proteins be put? And, what, if anything, does this pursuit tell us about how natural proteins fold, function and evolve? We describe the origins of this emerging area of *fully de novo protein design*, how it could be developed, where it might lead, and what challenges lie ahead.

Addresses
[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] BrisSynBio, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

Corresponding author: Woolfson, Derek N (d.n.woolfson@bristol.ac.uk)

Current Opinion in Structural Biology 2015, **33**:16–26

This review comes from a themed issue on **Engineering and design**

Edited by **Sarel J Fleishman** and **Andreas Plückthun**

For a complete overview see the Issue and the Editorial

Available online 18th June 2015

http://dx.doi.org/10.1016/j.sbi.2015.05.009

Introduction and scope of this review

Why study protein structures beyond those presented to us by nature?

It is now clearly established that the number of *protein folds* evolved through and used by biology is limited, and might comprise 1000–10000 different types [1–4]. (Herein, a protein fold is defined as the arrangement of secondary structure elements (SSEs) relative to each other in space.) That natural protein structures are limited stands to reason from the following argument.

Nature could not have explored all of the possible protein sequences or structures over the course of evolution. Many calculations have attempted to illustrate this, but all have caveats [5]. Even with many orders of magnitude shaved off such calculations, which attempt to enumerate the possible permutations of sequences and SSEs, we would be left with a mind-boggling number of molecules compared with say the estimated number of atoms available in the observable universe. In short, natural proteins potentially represent a tiny amount of the possible sequence and fold space. Thus, it is unlikely that nature evolved proteins by sampling this space exhaustively, and that this process was directed in some way; indeed, modern proteins likely arose through assembly and concatenation of smaller fragments [6,7]. On this basis, the vast majority of the possible protein sequences and structures have not been tested by evolution. However, some of these could be evaluated by *de novo* protein design, and potentially provide solutions to new protein-structure/function targets.

In the context of *protein redesign*, natural proteins do provide an extremely powerful toolkit and starting points for engineering new attributes and functions into protein architectures [8–12]. However, in terms of genuinely *de novo* protein design they put up borders between the known and immediately accessible protein world, and what might be out there to explore; rather like the perimeters of the cities in Logan's Run [13] or The Truman Show [14]. For the majority of this review, we focus on two related questions: First, if natural protein structures are not the only possibilities, what other protein folds are there? Second, how can we access these computationally and experimentally? We refer to such proteins that may only be accessible through design as *fully de novo proteins*. That said, we do this in the context of what has been achieved in *de novo* design thus far, and many of these designs should still be considered as being close to observed natural protein folds.

Current estimates of the number of natural protein structures

By the end of 2014, the RCSB Protein Data Bank (PDB) held >105,000 high-resolution protein structures [15]. The most widely used protein-structure classification systems, CATH and SCOP [16–18], suggest that these are accounted for by ≤1400 different *protein folds*, usually defined as the arrangement of secondary structural elements in space. The rate of discovery of new folds appears to be low: no new folds from CATH have been deposited in the PDB since 2012 [15], and an extended version of

SCOP [18] has reported just 15 new folds since 2009. There are problems with such analyses, however, for instance: how should protein folds and domains be defined [19]? Is the PDB subject to skewed sampling effects? And even that new folds are no longer being registered. As a result, there are disagreements on how many natural protein folds there might be, with estimates of up to 10,000 [1–4]. Nonetheless, the message is clear: natural proteins employ a limited set of 3D protein folds over again.

Enumerating and organising the protein-space that is possible

The above realisation leads to our first question: *what protein structures are possible beyond those presented to us by nature?* There are different ways to frame and consider this question [5]. One straightforward approach is to enumerate how many ways multiple SSEs — that is, α-helices and β-strands — can be combined in a linear chain, Figure 1a–d. For n SSEs, there are $(n-1)! \times 2\text{^}(2n-1)$ such permutations. We recognise that this is somewhat naïve, not least because it ignores 3D arrangements of secondary structures — that is, the overall protein fold and domain organisation — and also isolated β-strands tend not to be stable. Nonetheless, the calculation serves a purpose.

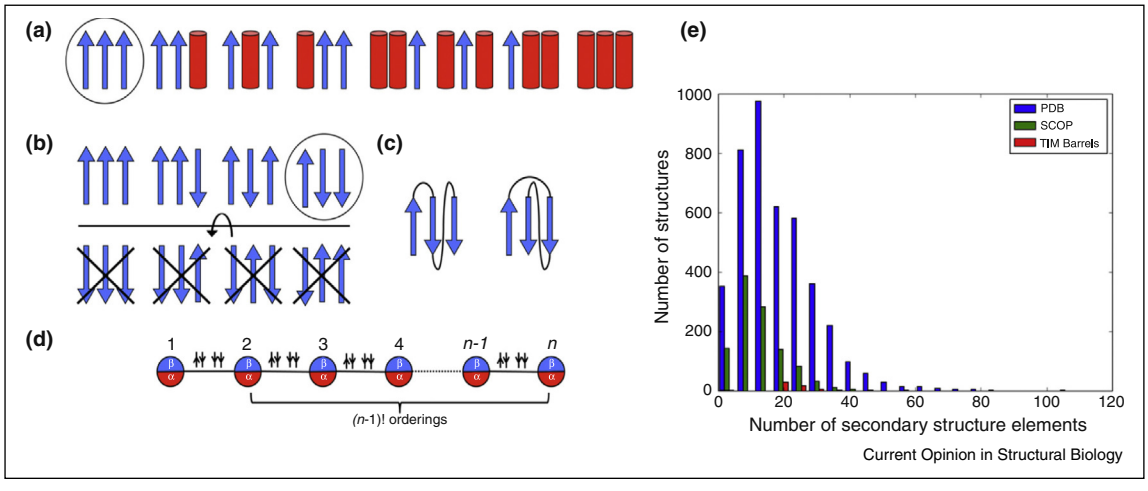
The resulting numbers of permutations would be experimentally manageable for chains with very few SSEs ($n \leq 3$). However, even for modest polypeptide chains with six SSEs, of which the 76-residue ubiquitin is an example, there are 245,760 permutations, with ubiquitin as just one of these. The same exercise on a chain with 16 SSEs, as in the $(\alpha/\beta)_8$-barrels or TIM-barrels, gives $>2.8 \times 10^{21}$ possible permutations. Again, the TIM-barrel is just one of these. True, this is a beautiful structure of alternating α-helices and β-strands. It is likely that its meandering topology and consolidated β-barrel aid its folding, stabilization and functionalization. In turn, this may help explain the predominance of the TIM barrel in nature, where it accounts for 10% of all of the known enzyme structures [20,21]. In these respects of folding kinetics and adaptability for function, it may be a *privileged fold*, or at least one very good solution to the protein folding/function problem. However, it is probable that within in the galaxy of the 2.8×10^{21} 16-element chains there will be other stable folds. The distribution of numbers of SSEs in known protein structures and domains is shown in Figure 1e.

Attempts have been made to rationalise and organise this space for specific protein folds such as four-helix bundles [22,23]; β-sandwich folds [24–26]; and structures that have similar arrangements of secondary structures neglecting the path or topology of the protein chain through these [27]. The most ambitious of these comes from Taylor, who has organised many of the possible 3D arrangements of secondary structures, referred to as *basic Forms*, into a *periodic table of protein structures* [28].

Unexplored protein space

The possibilities available to protein sequences and structures are often referred to synonymously as the

Figure 1



Possible permutations and observed numbers of the two main secondary structure elements (SSEs) in protein chains. **(a–d)** A simple calculation of the number of possible permutations, illustrated in **(a–c)** for three SSEs. In these models, the left-most SSE is the N-terminal one. **(a)** In a sequence of n SSEs, each one can be either α-helix (red cylinder) or β-strand (blue arrow), and there are $2\text{^}n$ different *orderings* of these. **(b)** For a given ordering (e.g., circled in **a**), there are $2\text{^}n$ *orientations*. Half of these are symmetry related, giving $2\text{^}(n-1)$ unique orientations. **(c)** For a specified ordering and orientation (e.g., circled in **b**), there are $(n-1)!$ possible *paths* through the remaining $(n-1)$ SSEs. **(d)** Schematic for the general case of n SSEs. At each position there are two choices of SSE, and each adjacent pair of SSEs can be arranged in a parallel or antiparallel orientation. With the left-most SSE defined as the N-terminal element, there are $(n-1)!$ ways that the remaining $(n-1)$ elements can be arranged in the primary structure of the protein. This calculation gives a total of $(n-1)! \times 2\text{^}(2n-1)$ permutations. **(e)** Observed numbers of SSEs in non-redundant (<40% sequence identity) protein chains of the PDB (blue), protein domains in SCOP (green), and those domains classified in SCOP as TIM barrels (red). α-Helices and β-strands were identified by Promotif [29].

protein universe [3,30]. We focus here on 3D protein folds rather than sequences. The difference between what is possible in protein-fold space, and what has been explored by nature — or at least what we have glimpsed of this so far — has been referred to by different names. Luisi calls these *never born proteins*. However, he is mainly concerned with generating random *de novo* sequences, and then finding structure and function within these [31]. Thus, this is more related to the studies of Szostak and others who select functional sequences from random libraries of proteins, than the structural resolution that we seek herein [32]. With this structural focus, Taylor calls the difference the *dark matter of protein space* [33]; Baker considers it part of *post-evolution biology*, which might be considered analogous to synthetic biology [34]; and we prefer *fully de novo proteins*.

Several groups are delving into this new space computationally. Taylor *et al.* have generated a large number of three-layer α–β–α structures *in silico* [33]. The starting points are known structures with this overall architecture, and multiple-sequence alignments made from these. Ambiguities in the latter, along with introduced variations, allow new *protein topologies* to be generated. Herein, protein topology refers to the string of secondary structures and the connectivities between these. In this way, protein-fold space is explored through the loss or gain of SSEs relative to the parent structure. The new topologies are then used to generate 3D models, which are compared in various ways to the known protein structures. Although this searches protein structures locally, the vast majority of the generated structures are new and unrelated to any known protein structures. Moreover, they are protein-like, and potential candidates for experimental designs.

In a different approach, Cossio *et al.* explore compact protein-like conformations accessible to a polypeptide chain of 60 valine residues through molecular-dynamics simulations [35]. This produces a large number (~7000) of tangible, independent protein topologies. Although all of the known topologies for natural proteins of a similar size appear, they represent only a small fraction (~300) of the full set. The simplicity of these models precludes immediate experimental validation, but, again, it illustrates that considerable structural complexity is possible beyond natural structures observed so far.

Moving from *in silico* design to experimental testing: parameterisation of protein structures

A key question for this new area of fully *de novo* protein design is: *how do we move from the theoretically anticipated dark matter of protein-fold space to its experimental exploration?* To put this into perspective, there are two clear advantages of designing of *de novo* proteins that mimic, or closely resemble natural proteins: First, natural backbone structures can be used as templates to start the design process. Second, sequence-to-structure relationships and/or statistical forcefields can be gleaned from natural proteins to guide the design of sequences. There are no such templates or relationships for complete *de novo* design; although, the statistical forcefields should still be useful for assessing any designs *in silico*.

Therefore, how do we kick-start the design process to move into the dark matter of protein space? There is hope: protein structures are modular, which opens possibilities for design approaches that employ secondary or supersecondary structures as building blocks. It helps that the two main secondary structures have regular backbone geometries and are scalable; that is, they are readily parameterised. Thus, in principle, larger protein structures and assemblies can be built through non-covalent association, or single-chain concatenation of standardised designs for regular smaller elements. Indeed, considerable progress is being made here, in what might be considered a combination of supramolecular chemistry and protein design/engineering. Several groups have achieved impressive, beautiful, and in some cases functional peptide and protein-based suprastructures based on *de novo* peptides [36•,37,38•], and engineered proteins [39–41,42•, 43].

Returning to generating completely *de novo* protein folds — where, arguably there is only one example to date, namely TOP7 from the Baker lab [44] — there are three main challenges: first, how do we position elements of secondary structure in space to produce new structures? That is, is this space reducible and parameterizable in any way? Second, how do we cement specific and stable interactions between elements of secondary structure? Third, how do we link the building blocks up with turns and loops to make single-chain polypeptides that direct folding as desired? The latter, loop-design problem often thwarts both protein engineering and design projects and solutions are actively being sought [45,46], but we will not address it directly herein.

Turning to the first challenge, the parametric description of protein folds has been a goal of protein scientists for decades [28,47–49]. These offer tangible routes to limit the structural space that must be searched to achieve novel protein folds. Similarly, the second challenge has been addressed in terms of understanding helix–helix and strand–strand interactions in natural protein structures for some time [50–52]. However, in both cases, there has been less effort and success in translating this understanding into *de novo* contexts. One way forward initially is to consider structures of high symmetry. Several protein structural types present as clear candidates here, for instance all-β (porin) and αβ-type (TIM) barrels, solenoid structures such as the leucine-rich repeats and so on.

Arguably, the α-helical coiled coil is the best-understood repeat structure [53,54], and, as such offers an ideal starting point for parametric design. In α-helical coiled coils two or more α-helices associate into rope-like bundles. This is programmed at the sequence level by variations on the conspicuous signature of coiled coils; namely, the heptad repeat of hydrophobic (*H*) and polar (*P*) residues, *HPPHPPP*. Coiled coils have been the subjects of many protein design studies [54,55]. Much of this work has focused on sequence-based designs, which use and embellish the heptad repeat [53]. Therefore, it is easy to forget that Crick's original postulate was as much about the geometry of helices and helical packing, as it was about the sequence patterns that might direct helix–helix interactions in coiled coils. Moreover, Crick's parametric equations, which elegantly describe classical coiled coils using just four parameters, have been demonstrated to reproduce the vast majority of subsequently experimentally determined coiled-coil structures very faithfully indeed [56,57•]. As such, Crick's parameters provide an excellent basis for the modelling and design of *de novo* coiled coils *in silico* [58–60]. Most recently, two web-accessible and user-friendly tools have become available to make coiled-coil modelling and design accessible to all. These are CCCP and CCBuilder [56,57•]. They widen the possibilities for modelling and designing new coiled-coil structures to the complete *periodic table of natural coiled coils* [61] and beyond.

Before moving onto more-recent protein design studies that are pertinent to our line of reasoning in the remainder of this review, we should like to make one thing clear: we do not regard most of the following examples as fully *de novo* protein designs as defined above. Some of them directly mimic natural protein folds; others are clear variations on natural themes; and only one is, as far as we can tell, unprecedented in the natural protein-fold space. That said, the studies are the state of the art, and they encourage us that computational routes into the dark matter of protein-fold space will follow.

Parametric protein designs achieved so far

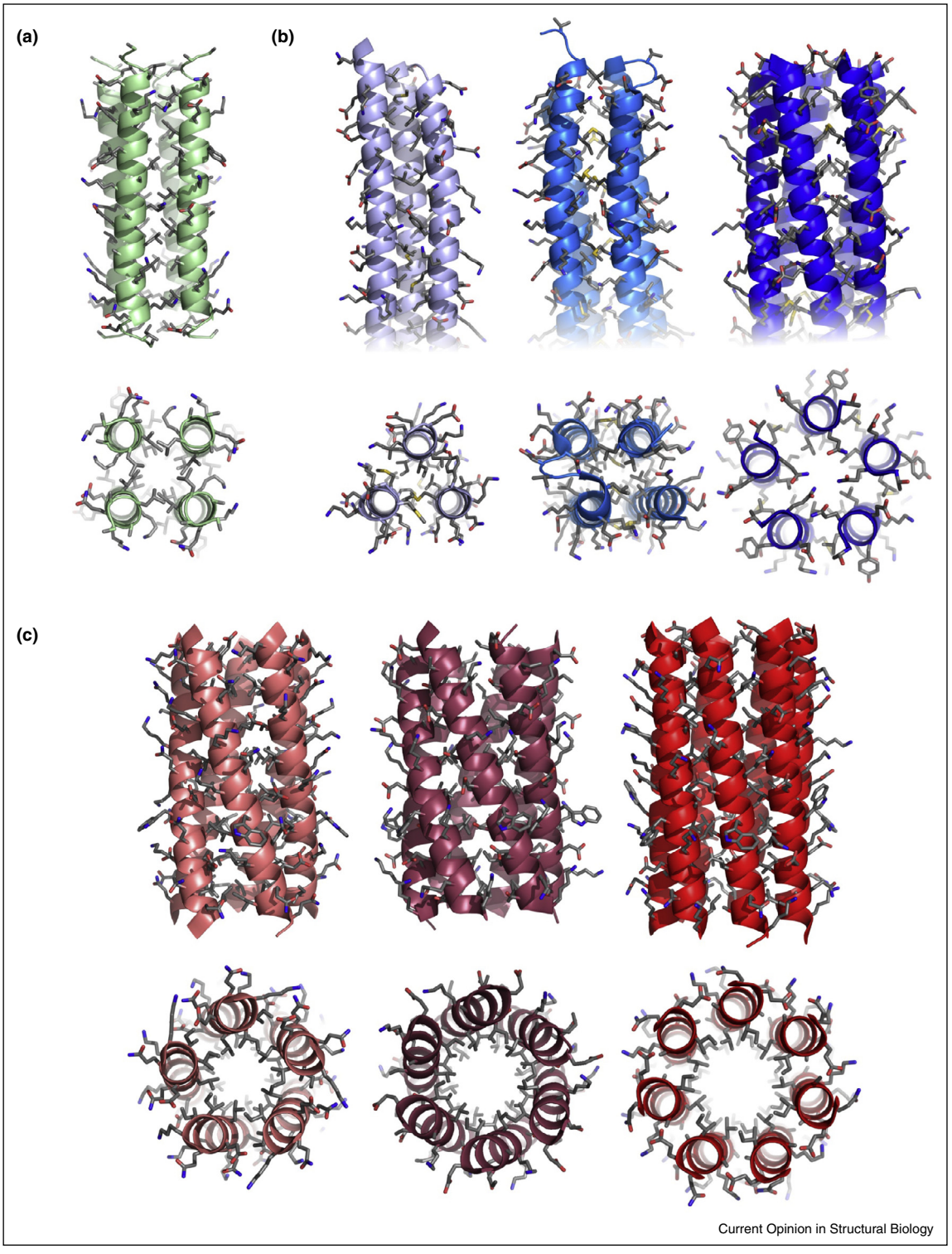
Parametric structure-based designs have been realised before now for coiled-coil proteins, including a right-handed structure that incorporates non-natural amino acids to satisfy unusual packing in the interior of the helical assembly [62], Figure 2a. However, CCCP, CCBuilder and methods developed in parallel in the Baker lab [56,57•,63••] allow structure-based designs of coiled coils and helical bundles to be tackled by more groups. A key feature of all of these methods is that they implicitly incorporate backbone variations, which has always been a challenge in computational protein design [56,57•,62,63••].

Grigoryan, DeGrado and co-workers have used CCCP to generate helices predicted to make barrel-like assemblies around carbon nanotubes [37]. Though still waiting for full structural characterisation, this is an exciting achievement with potential applications in solubilising carbon nanotubes for applications in bionanotechnology. More recently, the same group applied the parametric approach impressively to design functional membrane-spanning helical bundles [65••]. 'Rocker' is a four-helix bundle with two zinc-binding sites. The rationale is that zinc ions presented to one side of the membrane (the *cis* side) are transported to the other side (*trans*) via the two sites facilitated by the rocking of the structure such that it opens on the *cis* side, binds zinc at the first site passes it onto the second, and then exits the trans side of the membrane, Figure 3a. The group have characterised the designs to a high level of detail with a combination of liposome ion-flux assays, X-ray crystallography (of an apo, dimeric form) and NMR spectroscopy; although a high-resolution structure of the complete, functional, four-helix target remains elusive. DeGrado's group has also attempted to design peptides that switch between water-soluble and membrane-spanning peptides on lowering pH from 7.4 to 5.5 [66], Figure 3b. In assays with red blood cells, the low-pH states facilitate the release of ATP and miRNAs, but not haemoglobin. Further structural details are needed to validate the designs, and it is curious why the oligomer states of these peptides collapse in membranes, but still conduct small and macromolecules.

Baker's group has achieved parametric designs for hyperstable coiled coils [63••]. These are water-soluble assemblies including single-chain, three-helix and four-helix bundles that include both parallel and antiparallel helix–helix contacts, and a non-covalent, parallel five-helix bundle, Figure 2b. A key design feature is the structural focus on layers of hydrophobic residues that define the hydrophobic core, rather than more-traditional sequence-based repeats. By considering two-layer, three-layer and five-layer structures, which correspond to traditional seven-residue, 11-residue and 18-residue sequence repeats, the designs are for a non-covalent pentamer with a left-handed supercoil, and of single-chain constructs for a right-handed four-helix bundle, and a three-helix assembly with straight helices, respectively. These are confirmed by X-ray crystallography, Figure 2b. Interestingly, although designed using computationally using Rosetta, some of the packing solutions mimic those for natural and canonical designed coiled coils [54]; though not all have complete knobs-into-holes packing between helices as judged by SOCKET [67].

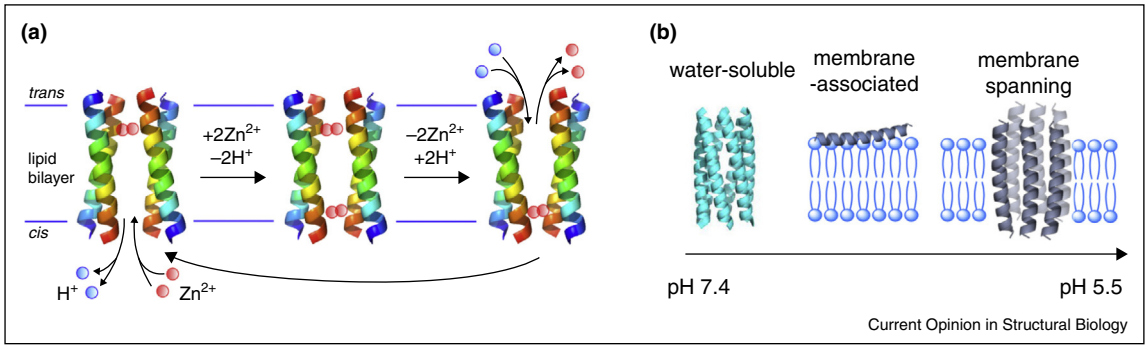
The third team is our own. Following the serendipitous discovery of a non-covalent coiled-coil hexamer (CC-Hex [68]), we have built parallel pentamers and a heptamer, CC-Pent and CC-Hept, and other hexamers by computational design as follows [64••]. The new designs are based on the realisation that successive helical interfaces

Figure 2



Structurally resolved, computational designs of coiled-coil assemblies and helical bundles. **(a)** A left-handed tetrameric assembly that incorporates non-natural side chains to achieve unusual packing in the hydrophobic core (PDB identifier 1RH4 [62]). **(b)** Hyperstable structures for three-helix and four-helix single-chain designs and a pentamer from the Baker laboratory (4TQL, 4UOS, and 4UOT, respectively [63••]). **(c)** Three designed α-helical barrels, CC-Pent, CC-Hex and CC-Hept (4PN8, 4PN9, and 4PNA, respectively [64••]), which have clear central channels.

Figure 3



Design of membrane-spanning active bundles. **(a)** The zinc-ion and proton transportation model of Rocker [65••]. This design mediates outward flux of Zn ions and inward flux of protons driven by pH gradients. **(b)** Schematic representation of water-soluble, membrane-associated and membrane-spanning states of a pH-dependent switch of a series of designed peptides [66].

in the cyclic structures can be approximated as heterodimeric faces encoded within the same peptide chain. In short, the helical interface is extended to span two helix–helix interfaces, in a so-called Type-II bifaceted coiled coil [55,69]. Next, sequences are selected computationally from one million alternatives. Then, ahead of experimental validation, the selected sequences are modelled in CCBuilder to predict the preferred oligomer state in the range tetramer to octamer. In this way, we have extended the reach of *de novo* coiled-coil design past canonical dimers, trimers and tetramers [70], which are plentiful in nature, into fully *de novo* pentamers and above for which there are few or no natural examples, Figure 2c. The structures are also noteworthy, and potentially useful, because they all have central channels the diameters of which scale with the number of helices in the assembly — the channels of CC-Pent, CC-Hex and CC-Hept are approximately 5, 6 and 7 Å across, respectively, [64••] Figure 2c. Thus, there is potential for such structures in the rational design and redesign of functional binding proteins, enzyme-like catalysts, membrane-spanning ion-channels, and peptide-based materials [71,72].

Future challenges for this aspect of the field include: the design of more structures, that is, populating empty elements of the periodic table of coiled coils [61], and beyond into the dark space of coiled-coil structures; the incorporation of dynamics, which will be key to delivering functional designs, but are poorly understood; and the further introduction of binding and catalytic functions [71,72].

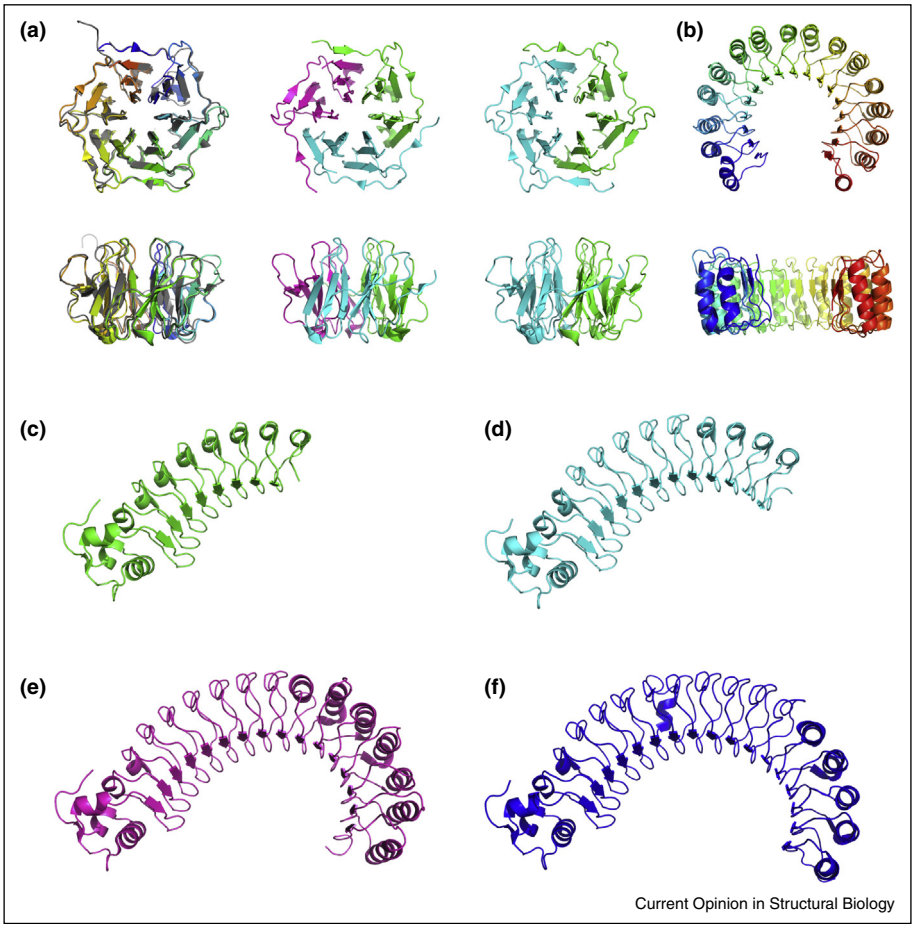
β-Propellers, solenoids and other potentially parameterizable protein structures

Moving into protein-fold space more widely will present more challenges. The coiled coil has advantages of symmetry, inherent stability, and clear parameter sets for defining and creating the possible backbone scaffolds. For other protein structures — both natural and dark — symmetry is reduced, stability is less guaranteed, and there are few or no parameter sets. Nevertheless, good progress is being made.

Tame and colleagues explore the design of new β-propeller folds, dubbed *Pizza proteins* [73•]. They start with a single blade from non-symmetric six-bladed propeller, from which they generate a fully symmetric six-bladed structure by blade duplication (the pizza slices), concatenation and optimisation in RosettaDock. The high-resolution X-ray crystal structure of the resulting single-chain protein, Pizza6, is closely similar to the designed model, Figure 4a. Shorter polypeptides, Pizza2 and Pizza3, trimerise and dimerise, respectively, to give six-bladed structures again. This demonstrates the robustness of both the designed blade and the approach. It also tallies with how multiply bladed structures may have arisen during evolution through gene duplication. Interestingly, larger sequences encoding >6 blades, form higher-order assemblies, which are mostly reconciled as discrete aggregates of the six-bladed design target; although Pizza7 behaves in more-complex ways.

Different curves and twists on another repeat protein are described by the groups of André and Baker [74••,75••]. They have both targeted the leucine-rich repeats LRRs, which form an array of horseshoe-like proteins that bind macromolecules via their inner surfaces formed by parallel β-strands connected on the outer faces by α-helices, Figure 4b. Rather than focus on similarities between repeats, and develop consensus sequences for a repeat, which has been successful in generating both structured and functional solenoid designs [76–78], both groups employ structure-based design to control the curvature (projected angle between successive repeats in the structure) and twist (the angle out of the plane between each segment). In this way, the groups aim to produce structures with all manner of shapes, combining set curves and twists

Figure 4



X-ray crystal structures for natural and designed β-propeller and α/β-solenoid proteins. **(a)** Orthogonal views of six-stranded β-propellers. Left: Superposed natural (grey; PDB identifier 1RWL [79]) and designed (rainbow; 3WW9) single-chain structures. Middle: A designed trimer 'Pizza2' (3WWF). Right: A designed dimer 'Pizza3' (3WW8). The designs are by Tame and colleagues [73•]. **(b)** Orthogonal views of the natural ribonuclease inhibitor (2BNH [80]). **(c–f)** Designed leucine-rich repeats of different sizes and subtly different shapes (4R58 **(c)**, 4R5C **(d)**, 4R5D **(e)**, and 4R6G **(f)**). These designs are from the Baker group [75••].

between repeats, to deliver high-affinity and high-specificity *de novo* binding proteins for any macromolecular target.

André's group use a repeat of the ribonuclease inhibitor (RI) fold as a starting point, Figure 4b, maintaining many of the hydrophobic-core residues that specify the local, repeat units [74••]. The other residues are mutated *in silico* to select combinations that should give a specified curvature between repeats, and zero twist. (In these respects, this approach combines elements of redesign and *de novo* design.) Concatenation of 10 self-complementary units, with additional capping units at the termini, into a single polypeptide gives stably folded and discrete proteins, although high-resolution structures will be required to verify the designs completely. Encouragingly, leaving out the capping units produces dimers, which are of the right dimensions in solution and by electron microscopy to fit the design hypothesis that the protein forms a flat semi-circular structure, two copies of which can interact head to tail to complete a circular, donut-like structure.

The approach of Baker's group is subtly different [75••]. They target a series of *de novo* self-complementary repeat modules, which when homo-oligomerized give defined curvatures. In addition, they develop junction, or wedge units to allow different repeat modules to be linked together to alter curvature within a single protein structure. Thus, in principle, these building blocks can be mixed and matched to create an array of polypeptides and define different shapes on demand. These could be regular, as with the design from the André group, or irregular. In this way, different binding surfaces could be built up for specific targets. The versatility of the approach is illustrated with 12 new proteins complete

with X-ray protein crystal structures [75••], Figure 4c–f. This approach is reminiscent of the drive in peptide self-assembly to generate toolkits of peptide-based building blocks that can be characterised once and then used in different contexts [36•,70].

What will we learn about natural proteins through this emerging area?

The topic of the evolution of protein folds is hotly debated [81]; as is the notion of what a fold is precisely, and how to classify them [82]. However, in evolutionary terms *natural protein structure space* is often considered discrete, whereas in terms of *possible protein structures* the space is by definition continuous. Grishin does not see these at odds, but simply as part of a duality in our understanding and theories of protein structure [81]. One emerging view is that this arises because evolved stable protein structures represent islands of stability within a sea of instability [83]; although, the sea does appear to be forded in places [84]. The question is: can protein designers build more islands and links *de novo*?

In *Arrival of the Fittest* [85], Wagner describes how evolution has bridged large spaces through connected *genotype networks*. This is as true of metabolic pathways, as it is for nucleic acid and protein-sequence spaces. These networks allow innovation — that is, the exploration of new genotypes and generation of new phenotypes — but, at the same time, the evolving systems can survive. Another way to put this is that for a given natural sequence or function there are many similar solutions nearby in sequence space, but it may be hard to escape from such regions.

In short, natural systems are robust. This presents a problem for protein redesign and design: if natural proteins are over-determined, are they necessarily good starting points for new designs, structural or functional? In other words, if point mutations or small numbers of changes keep you where you are, how can we move into completely new territory? If this is correct, and if we are approaching the limit of innovation with natural protein sequences and structures as scaffolds, perhaps it is time to look to the dark side? So, is there hope? Wagner also makes the point that there are multiple solutions to a given phenotype, and that sequence space is vast. Thus, solutions to the phenotype problem are not special or privileged *per se*; they are inevitable. If this extends to fully *de novo* protein folds, we should not have any problems in discovering them.

What will we do with fully *de novo proteins* once made?

There are clear uses for some of the structures described in this review. For example, a reliable set of α-helical barrels could be used as the basis for introducing binding or even catalytic properties into protein lumens of defined size and chemistry [55,71]; and, if polymerised, to produced peptide and protein-based nanotubes [72]. For the solenoid-type folds macromolecular binding presents the mostly likely function to be targeted [76,78]. Whereas access to β-propeller folds with different numbers of blades may open routes to many different functions [86]. More generally, access to a wide variety of robust, *de novo* protein folds — for which we have a clear understanding of their sequence-to-structure relationships, and that are free from sequence constraints or idiosyncrasies from millions of years of evolution — will provide future protein engineers with an extremely versatile toolkit of scaffolds to graft functions onto.

That said, with the exceptions of TOP7 and CC-Hept [44,64••], all of the *de novo* structures described so far are mimics of, or variations on natural protein folds — what is needed now is a push towards fully *de novo* folds. Aside from exploring new functions, this search for and study of such entirely new proteins is legitimate in itself for a number of reasons. Primarily, as with the entire ethos and origins of protein design, we pursue such goals as the ultimate test of our understanding of protein structure, folding, assembly and function; that is, after Feynman, '*What I cannot make, I do not understand*'. Extending this objective, the pursuit of entirely new protein structures, pushes the boundaries of what may be possible with seemingly straightforward polypeptide chains; the restrictions that are imposed on chain folding by Ramachandran; and, if we stick with proteinogenic amino acids, what can be encoded with just 20 amino acids. Furthermore, the exploration of such *de novo* proteins and assemblies fits with the emerging field of synthetic biology, for which there are high hopes in terms of new innovation and applications in biotechnology, medicine, materials science and beyond.

Acknowledgements

We thank Andrei Lupas for his insight and critical reading of an earlier version of this manuscript. DNW holds a Royal Society Wolfson Research Merit Award. The authors were supported by grants from the BBSRC (BB/J008990/1) and the ERC (340764) to DNW. AJB and CWW are supported by the EPSRC Bristol Chemical Synthesis Centre for Doctoral Training and the BBSRC South West Doctoral Training Partnership, respectively.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Chothia C: **Proteins – 1000 families for the molecular biologist**. *Nature* 1992, **357**:543-544.
2. Govindarajan S, Recabarren R, Goldstein RK: **Estimating the total number of protein folds**. *Proteins* 1999, **35**:408-414.
3. Kolodny R, Pereyaslavets L, Samson AO, Levitt M: **On the universe of protein folds**. *Annu Rev Biophys* 2013, **42**:559-582.
4. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds**. *Nature* 1994, **372**:631-634.

5. Marsh GE: **The problem of the 'prebiotic and never born proteins'**. *Int J Astrobiol* 2013, **12**:94-98.

6. Remmert M, Biegert A, Linke D, Lupas AN, Soding J: **Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin**. *Mol Biol Evol* 2010, **27**:1348-1358.

7. Kopec KO, Lupas AN: **Beta-propeller blades as ancestral peptides in protein evolution**. *PLoS ONE* 2013, **8**:e77074.

8. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG: **Theoretical and computational protein design**. *Annu Rev Phys Chem* 2011, **62**:129-149.

9. Goldsmith M, Tawfik DS: **Directed enzyme evolution: beyond the low-hanging fruit**. *Curr Opin Struct Biol* 2012, **22**:406-412.

10. Feldmeier K, Hocker B: **Computational protein design of ligand binding and catalysis**. *Curr Opin Chem Biol* 2013, **17**:929-933.

11. Kiss G, Celebi-Olcum N, Moretti R, Baker D, Houk KN: **Computational enzyme design**. *Angew Chem Int Ed* 2013, **52**:5700-5725.

12. Khoury GA, Smadbeck J, Kieslich CA, Floudas CA: **Protein folding and de novo protein design for biotechnological applications**. *Trends Biotechnol* 2014, **32**:99-109.

13. World Wide Web URL. http://en.wikipedia.org/wiki/Logan%27s_Run_(film).

14. World Wide Web URL. http://en.wikipedia.org/wiki/The_Truman_Show.

15. Rose PW, Bi CX, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M *et al.*: **The RCSB Protein Data Bank: new resources for research and education**. *Nucleic Acids Res* 2013, **41**:D475-D482.

16. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 2008, **36**:D419-D425.

17. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA: **Extending CATH: increasing coverage of the protein structure universe and linking structure with function**. *Nucleic Acids Res* 2011, **39**:D420-D426.

18. Fox NK, Brenner SE, Chandonia JM: **SCOPe: structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures**. *Nucleic Acids Res* 2014, **42**:D304-D309.

19. Schaeffer RD, Daggett V: **Protein folds and protein folding**. *Protein Eng Des Sel* 2011, **24**:11-19.

20. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions**. *J Mol Biol* 2002, **321**:741-765.

21. Richard JP, Zhai X, Malabanan MM: **Reflections on the catalytic power of a TIM-barrel**. *Bioorganic Chem* 2014, **57**:206-212.

22. Presnell SR, Cohen FE: **Topological distribution of 4-alpha-helix bundles**. *Proc Natl Acad Sci U S A* 1989, **86**:6592-6596.

23. Harris NL, Presnell SR, Cohen FE: **4 helix bundle diversity in globular-proteins**. *J Mol Biol* 1994, **236**:1356-1368.

24. Chirgadze YN: **Deduction and systematic classification of spatial motifs of the antiparallel-beta-structure in globular-proteins**. *Acta Crystallogr Sect A* 1987, **43**:405-417.

25. Woolfson DN, Evans PA, Hutchinson EG, Thornton JM: **Topological and stereochemical restrictions in beta-sandwich protein structures**. *Protein Eng* 1993, **6**:461-470.

26. Fokas AS, Gelfand IM, Kister AE: **Prediction of the structural motifs of sandwich proteins**. *Proc Natl Acad Sci U S A* 2004, **101**:16780-16783.

27. Minami S, Sawada K, Chikenji G: **How a spatial arrangement of secondary structure elements is dispersed in the universe of protein folds**. *PLOS ONE* 2014:9.

28. Taylor WR: **A 'periodic table' for protein structures**. *Nature* 2002, **416**:657-660.

29. Hutchinson EG, Thornton JM: **PROMOTIF – a program to identify and analyze structural motifs in proteins**. *Protein Sci* 1996, **5**:212-220.

30. Rekapalli B, Wuichet K, Peterson GD, Zhulin IB: **Dynamics of domain coverage of the protein sequence universe**. *BMC Genomics* 2012:13.

31. Chiarabelli C, De Lucrezia D, Stano P, Luisi PL: **The world of the "never born proteins"**. *Orig Life Evol Biosph* 2009, **39**:308-309.

32. Lane MD, Seelig B: **Advances in the directed evolution of proteins**. *Curr Opin Chem Biol* 2014, **22**:129-136.

33. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I: **Probing the "dark matter" of protein fold space**. *Structure* 2009, **17**:1244-1252.

34. Bromley EHC, Channon K, Moutevelis E, Woolfson DN: **Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems**. *ACS Chem Biol* 2008, **3**:38-50.

35. Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A, Laio A: **Exploring the universe of protein structures beyond the Protein Data Bank**. *PLoS Comput Biol* 2010:6.

36. • Fletcher JM, Harniman RL, Barnes FRH, Boyle AL, Collins A, Mantell J, Sharp TH, Antognozzi M, Booth PJ, Linden N *et al.*: **Self-assembling cages from coiled-coil peptide modules**. *Science* 2013, **340**:595-599.
This paper describes how standardised *de novo* peptide building blocks can be combined rationally to produce discrete water-soluble hubs, which then co-assemble into novel cage-like, supramolecular protein assemblies.

37. Grigoryan G, Kim YH, Acharya R, Axelrod K, Jain RM, Willis L, Drndic M, Kikkawa JM, DeGrado WF: **Computational design of virus-like protein assemblies on carbon nanotube surfaces**. *Science* 2011, **332**:1071-1076.

38. • Gradisar H, Bozic S, Doles T, Vengust D, Hafner-Bratkovic I, Mertelj A, Webb B, Sali A, Klavzar S, Jerala R: **Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments**. *Nat Chem Biol* 2013, **9**:362-366.
In this paper *de novo* and natural dimeric coiled-coil blocks are combined in a single-chain construct to direct the folding of a novel protein tetrahedron.

39. Sinclair JC, Davies KM, Venien-Bryan C, Noble MEM: **Generation of protein lattices by fusing proteins with matching rotational symmetry**. *Nat Nanotechnol* 2011, **6**:558-562.

40. Lai YT, Cascio D, Yeates TO: **Structure of a 16-nm cage designed by using protein oligomers**. *Science* 2012, **336**:1129.

41. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, Andre I, Gonen T, Yeates TO, Baker D: **Computational design of self-assembling protein nanomaterials with atomic level accuracy**. *Science* 2012, **336**:1171-1174.

42. • King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, Yeates TO, Baker D: **Accurate design of co-assembling multi-component protein nanomaterials**. *Nature* 2014, **510**:103-108.
This paper follows on from Ref. [41]. It describes the rational design of self-assembling polyhedra built from natural proteins the surfaces of which are engineered rationally to drive new protein–protein interactions to specify the polyhedra.

43. Song WJ, Tezcan A: **A designed supramolecular protein assembly with in vivo enzymatic activity**. *Science* 2014, **346**:1525-1528.

44. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy**. *Science* 2003, **302**:1364-1368.

45. Hu XZ, Wang HC, Ke HM, Kuhlman B: **High-resolution design of a protein loop**. *Proc Natl Acad Sci U S A* 2007, **104**:17668-17673.

46. Boyle AL, Bromley EHC, Bartlett GJ, Sessions RB, Sharp TH, Williams CL, Curmi PMG, Forde NR, Linke H, Woolfson DN: **Squaring the circle in peptide assembly: from fibers to**

discrete nanostructures by de novo design. *J Am Chem Soc* 2012, **134**:15457-15467.

47. Crick FHC: **The packing of alpha-helices – simple coiled-coils**. *Acta Crystallogr* 1953, **6**:689-697.

48. Murzin AG, Finkelstein AV: **General architecture of the alpha-helical globule**. *J Mol Biol* 1988, **204**:749-769.

49. Murzin AG, Lesk AM, Chothia C: **Principles determining the structure of beta-sheet barrels in proteins: 1. A theoretical-analysis**. *J Mol Biol* 1994, **236**:1369-1381.

50. Chothia C, Levitt M, Richardson D: **Helix to helix packing in proteins**. *J Mol Biol* 1981, **145**:215-250.

51. Walther D, Eisenhaber F, Argos P: **Principles of helix–helix packing in proteins: the helical lattice superposition model**. *J Mol Biol* 1996, **255**:536-553.

52. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN: **Determinants of strand register in antiparallel beta-sheets of proteins**. *Protein Sci* 1998, **7**:2287-2300.

53. Lupas AN, Gruber M: **The structure of alpha-helical coiled coils**. *Adv Protein Chem* 2005, **70**:37-78.

54. Woolfson DN: **The design of coiled-coil structures and assemblies**. *Adv Protein Chem* 2005, **70**:79-112.

55. Woolfson DN, Bartlett GJ, Bruning M, Thomson AR: **New currency for old rope: from coiled-coil assemblies to alpha-helical barrels**. *Curr Opin Struct Biol* 2012, **22**:432-441.

56. Grigoryan G, DeGrado WF: **Probing designability via a generalized model of helical bundle geometry**. *J Mol Biol* 2011, **405**:1079-1100.

57. • Wood CW, Bruning M, Ibarra AA, Bartlett GJ, Thomson AR, Sessions RB, Brady RL, Woolfson DN: **CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies**. *Bioinformatics* 2014, **30**:3029-3035.
The authors describe an open-access, user-friendly, web-based tool for generating and scoring fully atomistic models of coiled coils for any given sequence and across a swathe of parameter space.

58. Harbury PB, Tidor B, Kim PS: **Repacking protein cores with backbone freedom – structure prediction for coiled coils**. *Proc Natl Acad Sci U S A* 1995, **92**:8408-8412.

59. Offer G, Hicks MR, Woolfson DN: **Generalized crick equations for modeling noncanonical coiled coils**. *J Struct Biol* 2002, **137**:41-53.

60. Offer G, Sessions R: **Computer modeling of the alpha-helical coiled-coil – packing of side-chains in the inner-core**. *J Mol Biol* 1995, **249**:967-987.

61. Moutevelis E, Woolfson DN: **A periodic table of coiled-coil protein structures**. *J Mol Biol* 2009, **385**:726-732.

62. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS: **High-resolution protein design with backbone freedom**. *Science* 1998, **282**:1462-1467.

63. •• Huang PS, Oberdorfer G, Xu CF, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D: **High thermodynamic stability of parametrically designed helical bundles**. *Science* 2014, **346**:481-485.
This paper describes the parametric design, recombinant production, characterisation and X-ray protein crystal structures of hyperstable, single-chain, *de novo* coiled coils. The targetted structures are three-, four- and five-helix bundles based on non-canonical sequence repeats.

64. •• Thomson AR, Wood CW, Burton AJ, Bartlett GJ, Sessions RB, Brady RL, Woolfson DN: **Computational design of water-soluble alpha-helical barrels**. *Science* 2014, **346**:485-488.
In this paper computational designs are described for completely *de novo* coiled-coil-based alpha-helical barrels, including pentamers, hexamers and a heptamer. The approach combines sequence sifting and parametric modelling of the different oligomer states. The designs are made by peptide synthesis and characterized through to X-ray protein crystal structures.

65. •• Joh NH, Wang T, Bhate MP, Acharya R, Wu YB, Grabe M, Hong M, Grigoryan G, DeGrado WF: **De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle**. *Science* 2014, **346**:1520-1524.
Here rational computational design is applied to the generation of a membrane-spanning four-helix bundle that exchanges zinc ions and protons across biological membranes. The designs are characterised biophysically, using membrane-transport assays, and partly by structural biology.

66. Zhang Y, Bartz R, Grigoryan G, Bryant M, Aaronson J, Beck S, Innocent N, Klein L, Procopio W, Tucker T *et al.*: **Computational design and experimental characterization of peptides intended for pH-dependent membrane insertion and pore formation**. *ACS Chem Biol* 2015 http://dx.doi.org/10.1021/cb500759p.

67. Walshaw J, Woolfson DN: **SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures**. *J Mol Biol* 2001, **307**:1427-1450.

68. Zaccai NR, Chi B, Thomson AR, Boyle AL, Bartlett GJ, Bruning M, Linden N, Sessions RB, Booth PJ, Brady RL *et al.*: **A de novo peptide hexamer with a mutable channel**. *Nat Chem Biol* 2011, **7**:935-941.

69. Walshaw J, Woolfson DN: **Extended knobs-into-holes packing in classical and complex coiled-coil assemblies**. *J Struct Biol* 2003, **144**:349-361.

70. Fletcher JM, Boyle AL, Bruning M, Bartlett GJ, Vincent TL, Zaccai NR, Armstrong CT, Bromley EHC, Booth PJ, Brady RL *et al.*: **A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology**. *ACS Synth Biol* 2012, **1**:240-250.

71. Burton AJ, Thomas F, Agnew C, Hudson KL, Halford SE, Brady RL, Woolfson DN: **Accessibility, reactivity, and selectivity of side chains within a channel of de novo peptide assembly**. *J Am Chem Soc* 2013, **135**:12524-12527.

72. Xu CF, Liu R, Mehta AK, Guerrero-Ferreira RC, Wright ER, Dunin-Horkawicz S, Morris K, Serpell LC, Zuo XB, Wall JS *et al.*: **Rational design of helical nanotubes from self-assembly of coiled-coil lock washers**. *J Am Chem Soc* 2013, **135**:15565-15578.

73. • Voet ARD, Noguchi H, Addy C, Simoncini D, Terada D, Unzai S, Park SY, Zhang KYJ, Tame JRH: **Computational design of a self-assembling symmetrical beta-propeller protein**. *Proc Natl Acad Sci U S A* 2014, **111**:15102-15107.
Tame and colleagues describe how six-stranded beta-propeller structures can be constructed by concatenation of single, designed blade.

74. •• Ramisch S, Weininger U, Martinsson J, Akke M, Andre I: **Computational design of a leucine-rich repeat protein with a predefined geometry**. *Proc Natl Acad Sci U S A* 2014, **111**:17875-17880.
This paper describes a combined redesign and *de novo* design approach that delivers an idealised leucine-rich repeat, which when concatonated should form a highly regular semi-circular structure. The solution-phase characterization is strong. High-resolution structures are required to completely validate the designs.

75. •• Park K, Shen BW, Parmeggiani F, Huang PS, Stoddard BL, Baker D: **Control of repeat-protein curvature by computational protein design**. *Nat Struct Mol Biol* 2015, **22**:167-174.
This detailed paper describes the rational computational design of a series of leucine-rich-repeat modules. These are then combined to give a number of repeat solenoid proteins with different curvatures, many of which are characterized thoroughly through to high-resolution X-ray protein crystal structures.

76. Grove TZ, Cortajarena AL, Regan L: **Ligand binding by repeat proteins: natural and designed**. *Curr Opin Struct Biol* 2008, **18**:507-515.

77. Main ERG, Lowe AR, Mochrie SGJ, Jackson SE, Regan L: **A recurring theme in protein engineering: the design, stability and folding of repeat proteins**. *Curr Opin Struct Biol* 2005, **15**:464-471.

78. Boersma YL, Pluckthun A: **DARPins and other repeat protein scaffolds: advances in engineering and applications**. *Curr Opin Biotechnol* 2011, **22**:849-857.

79. Good MC, Greenstein AE, Young TA, Ng HL, Alber T: **Sensor domain of the *Mycobacterium tuberculosis* receptor Ser/Thr**

protein kinase, PknD, forms a highly symmetric beta propeller. *J Mol Biol* 2004, **339**:459-469.

80. Kobe B, Deisenhofer J: **Mechanism of ribonuclease inhibition by ribonuclease inhibitor protein based on the crystal structure of its complex with ribonuclease A**. *J Mol Biol* 1996, **264**:1028-1043.

81. Sadreyev RI, Kim BH, Grishin NV: **Discrete-continuous duality of protein structure space**. *Curr Opin Struct Biol* 2009, **19**:321-328.

82. Alva V, Koretke KK, Coles M, Lupas AN: **Cradle-loop barrels and the concept of metafolds in protein classification by natural descent**. *Curr Opin Struct Biol* 2008, **18**:358-365.

83. Lupas AN, Koretke KK: **Evolution of protein folds**. In *Computational Structural Biology: Methods and Applications.* Edited by Pitsch M, Schewede T. Hackensack, NJ: World Scientific; 2008:131-152.

84. Shah PK, Aloy P, Bork P, Russell RB: **Structural similarity to bridge sequence space: finding new families on the bridges**. *Protein Sci* 2005, **14**:1305-1314.

85. Wagner A: *Arrival of the Fittest*. Great Britain: Oneworld Publications; 2014.

86. Chen CKM, Chan NL, Wang AHJ: **The many blades of the beta-propeller proteins: conserved but versatile**. *Trends Biochem Sci* 2011, **36**:553-561.