Review

Special Issue – Synthetic Cell Biology

Principles for designing ordered protein assemblies

Yen-Ting Lai[1], Neil P. King[2], and Todd O. Yeates[3,4]

[1] Biomedical Engineering Interdepartmental Degree Program, University of California, Los Angeles, CA, USA
[2] Department of Biochemistry, University of Washington, Seattle, WA, USA
[3] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA
[4] University of California Los Angeles (UCLA)–Department of Energy Institute for Genomics and Proteomics, University of California, Los Angeles, CA, USA

In nature, many proteins have evolved to have self-complementary shapes. This drives them to assemble into supramolecular structures, sometimes of great complexity, and often carrying out sophisticated cellular functions. Designing novel proteins that can self-assemble into similarly complex structures is a longstanding goal in bioengineering. New ideas, combined with continually improving computer algorithms, are making it possible to advance on that goal, bringing wide-ranging applications in synthetic biology within reach. Prospective applications range from vaccine design to molecular delivery to bioactive materials. Recent strategies and examples of successfully designed protein cages, layers, and crystals are reviewed.

Protein assemblies as a synthetic biology goal

The emerging research area of synthetic biology seeks to recreate various complex phenomena exhibited by biological systems, especially at the molecular level [1]. The phenomena of interest are often characterized by a high degree of order – either in time or space. The emergent behavior or 'output' of synthetic systems can be considerably more complex than the behavior of the individual components [2]. Often, this arises from the introduction of non-trivial or self-referring interactions between components. For example, if rather than simply A affecting B, instead B also affects C, and C affects A, then an output behavior may arise where molecular concentrations oscillate in time [3]. Likewise, if rather than simply A binding to B, instead A binds to itself in multiple ways, surprisingly large and complex molecular assemblies can arise, leading to spatial organization of various types, such as compartmentalization or long-range propagation of forces by rigid structures. We concern ourselves here with strategies for designing protein molecules that self-associate to produce large, complex assemblies with potential synthetic biology applications.

Diverse efforts in the area of designing protein-based assemblies and materials can be divided into two groups: stochastic and deterministic [4,5]. The stochastic group encompasses several design strategies where the self-assembling protein molecule is highly flexible [4,6]. When a highly flexible molecule self-associates to form a higher-order assembly, the result is typically an extended and geometrically irregular material. Such network or mesh-like materials have interesting bulk properties, which can sometimes be modulated in useful ways [7]. The second group, where the assembly behavior is intended to be deterministic, encompasses those strategies aimed at producing specific 3D structures [5]. These structures, which may be finite in size (e.g., clusters or cages) or indefinite in extent (e.g., arrays or crystals), can be built with atomic level features in mind.

Early work on designing geometrically specific protein assemblies focused on filamentous structures as design targets. That early focus reflects the relatively simple design requirements for filamentous assemblies: a single self-associating interface can produce end-to-end polymerization. Specific well-studied self-associating protein motifs have been a rich source of building blocks for designing filamentous structures. Helical coiled-coils have been especially useful [8,9]. Cyclic polypeptides composed of β-strand-preferring amino acids have provided another self-associating motif, in this case leading to rigid tubular assemblies [10]. Variations on filamentous designs have sought to produce more complex patterns, such as branching [11], but the end-to-end polymerization strategies central to filament design do not extend easily to the problem of creating highly specific 3D architectures.

In this review we focus on strategies for designing proteins that self-assemble to give defined structures with complex architectures, including cages and extended 2D and 3D crystalline arrays.

Underlying principles

In nature, wherever supramolecular structures are built up by the assembly of multiple copies of the same subunit (or similar subunits), the subunits are nearly always assembled in a symmetrical fashion [12,13]. The reason for this was anticipated as early as 1956 by Crick and Watson [14]: symmetric assemblies require fewer distinct kinds of specific interaction interfaces compared to asymmetric assemblies. It is natural then that efforts to design ordered protein assemblies should rely on principles of symmetry. We articulate three connected ideas that we believe are important to permit full exploitation of symmetry-based

Corresponding author: Yeates, T.O. (yeates@mbi.ucla.edu).
Keywords: protein design; symmetry; cages; biomaterials; nanomaterials; assembly.

0962-8924/$ – see front matter http://dx.doi.org/10.1016/j.tcb.2012.08.004 Trends in Cell Biology, December 2012, Vol. 22, No. 12

approaches for designing self-assembling molecular systems:

(i) A symmetric molecular assembly, whether it is a finite structure or an extended array, is characterized by its symmetry group. A symmetry group is an exact mathematical idea that expresses the complete set of spatial operations that interrelate a set of individual components, in this case a set of structurally identical protein subunits in a 3D assembly. For example, a 24-subunit assembly built with the symmetry of a cube (referred to as octahedral symmetry) is described by a symmetry group with 24 spatial operations. Every pair of subunits is related by some operation in the group. For chiral biological molecules, such as proteins, these operations must be rotations, potentially combined with translations, but excluding reflections.

(ii) Any interface formed between two subunits in a symmetric assembly corresponds to one of the operations of the symmetry group (*vide supra*).

(iii) For a hypothetical, symmetric constellation of subunits (whether finite or indefinite) to constitute a plausible, physically connected assembly – in other words, to not be two or more disjoint sub-complexes – the following condition is both necessary and sufficient. Taking any individual subunit as a reference, the distinct interface types it uses to contact its neighbors comprise a subset of the operations of the complete symmetry group, and this subset of operations must be capable of generating the full symmetry group. That is, repeated combinations of this typically small subset of operations must be able to reform the whole symmetry group. In physical terms, this is equivalent to stating that it must be possible to trace a path from any one molecule to any other molecule in the assembly through the contacts between molecules in the assembly (Box 1). Otherwise, the collection of molecules would be disjoint.

The first two ideas are relatively well-known to those familiar with crystallography or molecular symmetry, but the third is less obvious. It was first articulated in the context of crystal symmetries [15], and then in the context of designed protein assemblies [16]. Particularly surprising, and important from a design perspective, was the realization that large, complex symmetries (including some 3D crystal symmetries) could be generated using only two distinct symmetry elements [16]. This translates to the idea that if one can design two distinct, geometrically specific, self-associating interfaces into a single molecule, then a wide range of assembly architectures can be realized. Although more than the minimum number of distinct contact types can be introduced in a design – and large natural assemblies such as viral capsids almost always exhibit more than the minimum number of distinct contact types [17] – the minimum contact number (which is just two for many cases) establishes an important design principle.

Design strategies and successes

Owing largely to the complexity of protein molecules, and our incomplete understanding of the rules by which they fold and recognize each other, designing complex protein assemblies has been a difficult challenge. However, efforts along multiple lines are beginning to bear fruit. Recalling the discussions above on the design requirement of introducing two (or more) modes of self-association into a single protein molecule, varied approaches to the problem of designing self-assembling proteins can be grouped according to the strategy used to satisfy this central requirement. Different strategies rely to different degrees on natural (or native-like) protein–protein interfaces versus novel interfaces created by amino acid sequence design (Figure 1).

Fusion of natural oligomers

An early idea for introducing two self-associating interfaces into a single protein molecule emerged at a time when the prospects for designing novel protein–protein interfaces by computational methods were limited. In 2001, we proposed that genetically fusing two different, naturally oligomeric domains into a single protein chain could satisfy the design requirement of combining two self-associating motifs [16]. The problem was how to dictate, or at least predict in advance, what the relative orientation would be between two genetically fused domains; free backbone rotations occur at a point of fusion, and this would make the final geometry unpredictable. The solution for how to predict the relative orientation in advance was to use oligomeric domains that began or ended in an α-helix, such that the protein backbone (and any α-helix-preferring amino acids introduced as a linker) might adopt an unbroken α-helix running from within one oligomeric domain into the next. In this way, pairs of oligomeric domains of known structure would be combined in hypothetical, predicted arrangements, and a search could be made for pairs that would satisfy specific geometric rules for constructing different architectures such as cages or crystals. The same paper laid out geometric rules for how the symmetry axes of the component domains would have to be oriented relative to each other to obtain various architectures ([16], Box 1). That initial set of geometric rules only covered combinations of dimers and trimers because the protein structure database at that time contained relatively few proteins with both higher oligomeric symmetry and subunits ending in helices [16]. A tremendous range of assembly architectures can be achieved using higher symmetry building blocks; those possibilities have not been completely enumerated yet.

The oligomeric fusion method was first used in the design of protein filaments – a relatively easy design target – and a 12-subunit molecular cage with a tetrahedral shape, representing the first protein construction of its kind. However, that initially designed protein sequence formed cage-like assemblies whose sizes were too heterogeneous to characterize in atomic detail; crystals could not be obtained [16]. This barrier was surmounted in recent work. When the original design was revisited, and two amino acid changes were made based on a visual identification of potential steric conflicts, a homogenous 12-subunit cage was obtained and crystallized [18]. The designed cage is roughly 16 nm in diameter and contains a central opening about 5 nm in diameter (Figure 2a,b). Interestingly, despite an overall match to the design, the observed assembly exhibited significant deviations (about 8 Å root-mean-square deviation, RMSD) from perfect symmetry [18].