Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Editorial overview: Sequences and topology: 'paths from sequence to structure'

Nir Ben-Tal and Andrei N Lupas

Current Opinion in Structural Biology 2021, **68**:vi–viii

For a complete overview see the Issue

https://doi.org/10.1016/j.sbi.2021.05.005

Nir Ben-Tal

George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel
e-mail: bental@tauex.tau.ac.il; Web: https://www.bentalab.com/.

Nir Ben-Tal studied Biology, Chemistry and Physics at the Hebrew University of Jerusalem and obtained his DSc in Chemistry from the Technion, Israel Institute of Technology. After postdoctoral training at Columbia University, Ben-Tal joined Tel Aviv University as group leader. In collaboration with his colleagues, Ben-Tal develops and uses computational tools to study protein structure, function, dynamics, and evolution, and also contributes to drug discovery. Recently, their studies focus mostly on a fundamental question in molecular evolution: How do novel protein architectures emerge.

Andrei N Lupas

Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany
e-mail: andrei.lupas@tuebingen.mpg.de

Andrei Lupas studied Biology at the Technical University Munich and obtained his PhD in Molecular Biology from Princeton University. After postdoctoral training with Andreas Plückthun and Wolfgang Baumeister at the MPI for Biochemistry in Martinsried, Lupas joined SmithKline Beecham, later GlaxoSmithKline, as a Senior Computational Scientist and Assistant Director of Bioinformatics. In 2001 Lupas returned to the Max Planck Society as Director of the Department of Protein Evolution at the MPI for Developmental Biology. He studies the evolution of proteins by means of bioinformatics, biochemistry, and structural biology, particularly the emergence of folded proteins at the origins of Life.

Sixty years ago the atomic structure of myoglobin and its surprising irregularity triggered the realization that protein structures could not be inferred from the properties of polypeptide sequences as simply as the structure of DNA had been from those of nucleic acids. Expectations of structural regularity in proteins, nourished by the secondary structures described by Linus Pauling and the first three-dimensional models for protein fibers by Pauling, Francis Crick, and G. N. Ramachandran gave way to the realization that for proteins the relationship between sequence and structure, although deterministic, was complicated. This 'protein folding problem' became one of a handful of grand challenges in molecular biology, as more and more elaborate and determined efforts failed to solve it.

Since the mid 1990s, progress on this problem has been tracked by the biennial CASP experiments (Critical Assessment of Structure Prediction). These showed that, once the pervasive use of 'postdiction' — predictions for which the result was already known — was eliminated, the methods used for structure prediction were not very performant. The doublyblind setup of CASP experiments and their rigorous assessment helped to focus researchers on the most promising approaches and the field showed clear progress from CASP1 (1994) to CASP5 (2002). After that, progress slowed down considerably through to CASP12 (2016), possibly because the most important source of structural information, proteins of known structure homologous to the target, had been mined comprehensively by increasingly powerful sequence comparison methods.

During this time a related question came to the fore, prompted by the difficulties in making progress on the protein folding problem. If this problem was so complex, how had nature solved it, given that all processes of life are substantially dependent on folded proteins? Work by bioinformaticians and evolutionary biologists had produced comprehensive databases of protein domains, which showed that since the time of the Last Universal Common Ancestor (LUCA) some 3.5 billion years ago, new proteins had mostly arisen by the amplification, recombination and differentiation of more ancient domain prototypes. Since domains are defined biochemically as autonomously folding parts of a polypeptide chain, nature had thus apparently bypassed the folding problem since LUCA. However, searches for folded proteins in libraries of random polypeptide chains had shown that at chain lengths typical for domains, the likelihood of encountering a folded exemplar was extremely low, begging the question of how domain prototypes had arisen in the first place, before LUCA split into Bacteria and Archaea.

One scenario, based on the observation that similar subdomain-sized fragments appear to have been reused in various contexts, proposed that

primordial peptides evolved as cofactors of RNA-based replication and catalysis (the 'RNA world') and assumed structure on scaffolds, be they RNA, the first membranes, or abiotic surfaces. They gradually achieved the ability to fold independently through an increase in complexity, for which several mechanisms have been discussed. These include repetition of the same peptide in an oligomer or within the same polypeptide chain, extension by relocations of the start and stop codons, recombination of non-identical peptides into longer chains, or evolutionary optimization of peptides that were originally only structured in complex with a cofactor (such as a nucleotide, heme, or an iron-sulfur cluster).

In this issue of Current Opinion in Structural Biology, we have assembled articles that illuminate three of the paths by which protein sequences may become folded structures, in nature and by design. The first, repetition, was already proposed in 1967 by Richard Eck and Margaret Dayhoff and, indeed, many proteins feature repetitions of various sizes, from short linear motifs all the way through whole domains. Mylemans, Voet & Tame survey β-propellers, which are among the clearest manifestations of repetitions. They suggest evolutionary pathways that may have led from a single blade to the contemporary propeller repertoire by amplification, and point out the usefulness of the pathways for protein engineering. Outer membrane β-barrels are also excellent candidates for an origin by repetition, and Dhar & Slusky present our current view on this. They argue that most barrels share a common ancestor — with efflux pumps illustrating an exception to this rule — and suggest the underlying evolutionary pathways. They further describe how the pathways can be used within the context of rational protein engineering.

Whereas β-propellers and β-barrels are both toroidal folds, in which the repeats form a closed circular structure, most repeat proteins in nature are either fibers (coiled coils, collagens) or solenoids, that is, open-ended stacks of repeating units forming a helical array. Gidley & Parmeggiani review recent advances in the engineering of new solenoid proteins, both by customization of natural repeat units and by *de novo* design. They conclude that the modular nature of solenoids and their open-ended nature makes them ideally suited to build large protein structures without the need to redesign more than limited parts of the repeating unit. Natural repeats are frequently α-helical in nature and show local packing geometries familiar from coiled coils. In their survey of coiled coils and helix bundles, both naturally occurring and engineered, ElGamacy & Hernandez Alvarez consolidate the understanding that inter-helix packing via known periodic motifs facilitates formation of stable symmetrical structures. They also highlight the fact that deviations from these motifs, which are occasionally tolerated, result in local asymmetry and destabilization. The local impairments offer various functional opportunities such as ligand binding, and conformation changes. All in all, this is yet another demonstration of how stability is compromised for the emergence of function, one of the general principles in evolution and engineering.

Homo-oligomeric proteins can be viewed as yet another manifestation of repetition. Ragonis-Bachar & Landau survey recent advances in studies of amyloids, a unique form of homo-oligomerization. The ability of amyloids to readily form highly ordered and extraordinarily stable structures makes them particularly attractive within the context of the emergence of proteins. Recent studies reveal that some of them are polymorphic, that is, stable in several conformations, and are functional. Both qualities land further support for the idea that amyloids, which are currently notorious for their involvement in neurodegenerative diseases, may have contributed to the origins of life.

The second path to the emergence of structured proteins is the recombination of previously evolved fragments. Given the frequency of illegitimate (or nonhomologous) recombination in genomic DNA, it seems reasonable to expect that this process should occasionally lead to the formation of new domains by the merger of unrelated but mutually compatible protein fragments. Recombination leaves traces in contemporary proteins. However, these might not always be readily detected because of natural mutational drift. Kolodny surveys various search strategies that have been used to this end, highlighting their pros and cons, focusing on similarity between protein segments that are smaller than domains. She then argues that protein segments that share the same evolutionary origin may nevertheless be found in different conformations in current day proteins. Taking a somewhat different angle on the same subject, Romero-Romero, Kordes, Michel, & Höcker focus mostly on reuse where structure is retained. They survey how the detection of such sequence and structure conserving segments is useful for grafting within the context of protein engineering, nicely demonstrating it with TIM barrels and related folds.

The third path to autonomous protein structure is *de novo*, by evolution under functional constraints from previously non-coding genetic sequence. Special cases of this are presented by alternate reading frames on the same DNA strand as coding sequences (overprinting) or by in-frame messages on complementary strands (bidirectional coding), both constrained by the nature of the genetic code. Carter provides a careful discussion of these two cases and concludes that, whereas overprinting is still common as a birthing ground for new proteins in viruses, bidirectional coding was an important but transient stage in the early evolution of proteins. In cellular organisms today, the emergence of new proteins appears to proceed largely from DNA sequence previously non-coding in any frame,

and Bornberg-Bauer, Hlouchova, & Lange describe possible underlying mechanisms for this process, arguing that it suggests the existence of many virtual proteins of biological relevance in sequence space, not yet explored by natural evolution. Tong, Lee & Seelig provide a complementary view, exploring how functional proteins can be isolated from randomized sequence libraries by combing evolution with rational design, while Skolnick & Gao discuss the physicochemical limits of the process, concluding that the native structure of a polypeptide chain is determined mostly by compactness, and that therefore fold space is likely saturated.

Skolnick & Gao and Pearce & Zhang provide the concluding overview of paths leading to folded protein structures by discussing the impact of deep learning on their prediction and design. The Zhang group in particular has been one of the leaders in protein structure prediction for over a decade and their discussion provides conclusive insight into the developments that led to the astonishing breakthrough in structure prediction by DeepMind at the recent CASP14 (2020). Combining superb software engineering with raw number-crunching power, the DeepMind group, AlphaFold2, was able for the first time to submit models that rivaled experimental structures for most targets, leading the CASP organizers to conclude that the structure prediction problem for single protein chains was solved. This breakthrough promises to energize all life sciences, and protein science in particular, by making protein structure information available at the speed of computation and relieving the need for time-consuming and resource-intensive experimental determination. There is much to look forward to in the coming years.