Published in final edited form as:

Curr Opin Struct Biol. 2016 June ; 38: vii–ix. doi:10.1016/j.sbi.2016.06.016.

Editorial overview: Sequences and topology: Protein function - synthesizing information from sequence, structure and disorder

Elizabeth M Meiering and

Department of Chemistry, University of Waterloo, Waterloo, ON N2L 3G1, Canada

L Aravind

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

The expansive advances in genome sequencing, large-scale structure determination and parallel methodological developments have left us with enormous protein sequence and structure data. The information from these resources is helping us answer long-standing questions in biology and biochemistry but at the same time is uncovering new puzzles. A great variety of approaches are being used to mine these data. Consequently, connecting several areas of investigation into a common framework is rather challenging. With this in mind we have sought to bring together a series of reviews on diverse current topics in sequence, structure and methodological studies which are interconnected and feature uniting themes.

The wealth of data demands concomitant methodological developments and the need for new methods is acutely felt in regard to leveraging the huge accumulations of protein sequences to extract new biochemical information. In this context Neuwald reviews recent developments in the use of residue correlations between positions in protein multiple sequence alignments to obtain structural and functional information. There appears to be considerable promise in the Direct Coupling Analysis (DCA) and the complementary Bayesian Partitioning with Pattern Selection (BPPS) methods to identify residues relevant to folding of a domain as also biochemical function (e.g. active sites). As proteins in extant organisms are the products of historical processes, methods for ancestral sequence reconstruction have emerged. The state of this field is examined in a succinct review by Wheeler *et al.* A flurry of recent experiments have characterized the properties of reconstructed sequences. The authors define two key questions pertaining to these results: (1) are ancestral sequences more thermostable, and (2) are they more promiscuous in function? The answers to these questions have profound implications for protein evolution and design. Based on the available data the answer to the first question is 'yes', noting also that some existing organisms have proteins with comparable properties. The answer to the second is uncertain and defines a key area for future research. Going from analysis of protein sequences to function also needs the integration of other types of information. This aspect is treated by Lobb and Doxey in their review, where they discuss the role of integrating sequence and structure analysis with contextual information from gene-neighborhoods, domain architectures, phyletic patterns and structural modeling. They provide examples of recent successes of inferring biological function by combining different methods and performing such analysis on the genome-scale. One such case of concrete

functional inference regards the rich repertoire of secreted carbohydrate metabolism proteins seen in members of the *Bacteroidetes* clade which are an important component of the human microbiome.

One of the key successes of applying sequence and structure analysis methods to genomic and protein 3D-structure data has been the delineation of the evolutionary, tertiary structural and functional building blocks of proteins, that is, domains. Accordingly, we have additional reviews that delve deep into the natural history of domains. Orengo and coworkers provide a primer in this regard by surveying the evolution of function in protein domains, building on sequence-based and structure-based classification of domains into families and superfamilies. Recent progress includes insights into the mechanisms of evolution and functional innovation both for individual domains and for combinations of domains. Different domains clearly vary in their ability to support the evolution of new biochemistry. Moreover, function may be maintained despite extreme structural changes, for example, of fold topology, or a structured domain may become intrinsically disordered and regulated in particular domain architectural contexts. This also brings home the important point that not just structured domains but also disordered regions of proteins have much potential in terms of biochemical function.

The theme of domains and function is further explored in depth by a series of reviews focusing on particular classes of domains of functional and structural interest. Starting with the cell-membrane, Saier reviews the structure and evolution of transmembrane proteins. It is proposed that the majority of such proteins mediating traffic across the membrane with membrane-spanning α -helices appear to have emerged from simple precursors with just 1–3 helices. It is also suggested that such precursors spawned all major classes of transmembrane transport proteins via an evolutionary pathway which passes through secondary carriers all the way to group translocators that modify their substrates. The secondary carriers are also proposed to have given rise to transmembrane enzymatic domains such as the rhomboid peptidases. Moving outside the cell, Vadnais and Smider briefly review the interesting case of dramatic domain evolution on a relatively short timescale as exhibited by the bovine antibodies. Cows produce a special type of antibody in which the third complementarity determining region of the heavy chain has a long β -hairpin stalk where an unprecedented disulfide-bonded domain termed the knob is inserted at the tip. These knob-domains diversify via an as-yet-unclear hypermutation process to assume considerable variety of structure stabilized by distinct patterns of disulfide-bonds. Thus, these antibodies may provide a model for understanding how structures of disulfide-bonded secreted proteins can rapidly evolve.

Moving into the cellular interior, Matthews, MacKay and coworkers have contributed reviews on how proteins bind nucleic acid sequences. Focusing on the rich structural diversity among eukaryotic transcription factors, Smith and Matthews survey established structural models for protein–DNA interactions, and then tackle the elaborate and diverse mechanisms that determine specificity. All binding partners generally participate in multiple multi-component complexes, creating varying intricate networks of interactions in the cell. Different regions of a given protein, including intrinsically disordered and ordered domains, contribute in complex ways to binding, which may be regulated by a range of covalent

modifications of proteins and DNA. Computational modeling is playing an increasing role in understanding and informing experiments, including the design of functional transcription factors *in vivo*. Nevertheless, the intricacies of eukaryotic transcription regulation mean that, despite impressive progress, there is challenging work ahead to attain full understanding and control of transcription factor function. Helder *et al.* survey intense recent research suggesting that RNA recognition by proteins is even more complex than that of DNA. We are only beginning to understand the enormous diversity in modes of RNA recognition exhibited by RNA-binding proteins: the interactions of many enigmatic RNA-binding domains are yet undefined and the status of long non-coding RNA–protein interactions remains unclear. An emerging new paradigm involves fluctuating, multivalent interactions between many binding sites and partners, which underpin the formation of RNA–protein droplets and gel phases (membraneless organelles). Such binding is central to diverse natural and disease processes, which relates to aggregate states formed in cells by disordered proteins.

Halfmann develops this theme further by exploring low complexity protein sequences (LCS) or compositionally biased sequences that are enriched in some amino acids. These sequences, which are particularly common in eukaryotic proteins, can make varying multivalent inter-molecular interactions (with self or other molecules) resulting in phase transitions to dynamic metastable liquid droplets and subsequent solidification to one dimensional crystals (i.e. various fibers) or disordered glasses. On one hand there is increasing evidence that such phases adopted by LCS in proteins have biological roles in regulation of transcription and translation, cellular antiviral responses, inflammation and cell fate determination. On the other hand, when these processes ‘run-away’ or arise from misfolding of normally folded soluble proteins they can prove toxic. The resulting protein and protein–nucleic acid aggregation is associated with several neurodegenerative diseases, such as Alzheimer’s, Huntington’s and Amyotrophic Lateral Sclerosis. Recent work is unpacking the differences in mechanisms and rates of interconversion between different conformations, the variation in which can be important in determining the biological ramifications of the transitions, in space, time and upon mutation. There is increasing recognition and definition of the distinctions between classical prion, prion-like, amyloid, amyloid-like and the liquid phases; this is a critical and propitious area for future research.

One aspect of protein folding, misfolding and aggregation relates to the rates of translation, which is reviewed by Jacobson and Clark. They summarize the recent progress in understanding how synonymous codons affect translation rate and co-translational protein-folding mechanisms. Codons that are more common tend to be translated faster than rare synonyms, and mutations that alter translation rate have been demonstrated to significantly alter co-translational folding mechanism, resulting in alternatively folded domains or aggregates, and consequent altered cell function. Variations in the speed of translation along the polypeptide chain can thus impact natural and disease-related folding and misfolding; also, the tuning of codon usage during heterologous protein expression is of considerable practical importance.

A key corollary to gleaning information from natural sequences, as has been covered in the above reviews, is using the principles learned therefrom to design new proteins in the

laboratory. We are offered a survey of the challenges and potential solutions in achieving this by Broom and Meiering. A notable point is that successful design of new proteins requires multiple iterations before converging to a successful solution and that the current metrics used to distinguish failures from successes are not strongly discriminative. Nevertheless, the use of symmetries, which are common in natural proteins due the evolutionary process of tandem duplications (or repetition) of starting units, also prove to be effective in engineering artificial proteins. Starting with such symmetric structures, designs may be further refined via coarse-grained simulations to analyze the energy landscape of the engineered protein domain. This can also be combined with atomistic simulations to ‘work out’ more difficult elements of the design like unintended aggregation.

In selecting this set of reviews we have attempted to cover the gamut of major issues in the study of protein structure and function. One consistent theme that emerges is the power of bioinformatics and computational methods in retrieving functional information regarding proteins from the large stores of data that are being accumulated. Thus, laboratory investigations of protein functions as well as design of new proteins are increasingly guided by this information. In this way we continue to test our understanding of what we are learning from the experiments that have already been done over billions of years of protein evolution in nature.

Biographies

Liz Meiering completed her PhD in Biological Chemistry at the University of Cambridge, UK, followed by postdoctoral research in Biological Chemistry and Molecular Pharmacology at Harvard Medical School, USA. She is a Professor in Chemistry and Biology at the University of Waterloo. Her research team uses biophysical and biochemical experimental methods in combination with computational modeling to elucidate relationships between protein sequence, folding, and function. Recent studies have focused on the molecular mechanisms of allostery, misfolding and aggregation in protein expression and disease, as well as protein engineering and design.

L Aravind obtained his PhD in Biology from the Texas A&M University while doing most of his research at the National Center for Biotechnology information, NIH, USA. Ever since, he has been principal investigator at the NCBI. His expertise is in comparative genomics, protein sequence and structure analysis. His research has helped solve several problems in biochemistry and evolution relating to nucleic acid repair, protein and nucleic acid modifications, secondary metabolite synthesis and biological conflicts.