Towards a Better Understanding of Biological Structures with Unknown Functions

Herman, M. A.[1]*, Trakselis, M.A.[2]*, and Hwang, W.[3]*

[1]School of Biological Sciences, University of Nebraska-Lincoln;
[2]Department of Chemistry and Biochemistry, Baylor University;
[3]Departments of Biomedical Engineering,Materials Science & Engineering, and Physics & Astronomy, Texas A&M University

*For correspondence
Michael A. Herman (mherman5@unl.edu)
Michael A. Trakselis (michael_trakselis@baylor.edu)
Wonmuk Hwang (hwm@tamu.edu)

Summary

The relationship between structure and function is a major constituent of the rules of life. These relationships occur at multiple levels of biological organization. With reams of data from molecules, genes, genomes, cells, tissues, whole organisms, populations, communitites and even ecosystems, we still encounter structures whose functions are unknown and/or cannot be predicted based on similar structures with known functions. In order to arrive at a comprehensive understanding of the functions of biological structures, their dynamics (*i.e.*, time-dependent motion and interaction), as well as their role within the context of bigger assemblies need to be extensively incorporated. Recent developments in experimental modalities with increasing spatial and temporal resolution, as well as advances in computational capability, will allow us to explore these questions in unprecedented ways.

Defining and measuring biological structure. Inherent to structure is shape, organization, and patterning. We refer to a basic structure as a biological entity that has form and substance and as such, structures inhabit all levels of biological organization. We need to have an inclusive definition of structure in order to map onto one or more functional outputs. Structure can take on multiple forms or states from the very static to increasingly dynamic. At the innermost level, commonly there is a central core that serves as a foundation. The core may provide the most visual representation of shape but may not be the best predictor of function. Flexibility exists at the next level where global dynamics and conformational change can affect shape or organization. Finally, the most dynamic state includes a level of disorder or stochastic action. The continuum of structural dynamics that emanate from these elements will be the most likely predictor of function.

The concepts of core, flexibility, and disorder are meant to translate across and through biological levels from the molecular up to and through ecological systems. Of course, there can be many alternative intermediaries along a spectrum of structures, and these concepts are meant to serve as benchmarks or guidelines. With this foundation, models to predict function from structure developed at one level may be used to understand at another level. We also recognize that many biological structures may have multiple functions within or across given levels but also at different time scales or conditions. What principles direct these multidimensional structure-function relationships is an extremely interesting question to pursue.

Our current understanding of structure relies on a variety of experimental tools from the nano to the macroscale. For example, once we had large datasets of genome sequences from divergent

species, comparison of those genome structures led to the development of Structural Genomics that can describe the organization of genomes. Inherent in these comparisons is the use of sequence conservation to understand the evolutionary relationships of genome sequences. Similarity of sequences that is arrived at through descent from a common ancestor is referred to as homology, which then informs structure-function relationships of genes and genomes (Goldsmith-Fischman and Honig, 2003). More recently, DNA sequence data and associations of sequences across the genome have been utilized to capture three dimensional architecture of larger chromosome (or genome) interactions involved in the regulation of gene expression (Kempfer and Pombo, 2019). Advances in microscopy techniques have allowed visualization of structure down to the nanometer scale, more recently using superresolution imaging (Huang et al., 2009). Techniques including resonance imaging and ultrasound have been adapted in many specific ways to provide information on tissue and organismal structures (Goth et al., 2016). Sub-nanometer resolution structures have been resolved primarily from tools such as nuclear magnetic resonance (NMR), X-ray crystallography, and electron microscopy that can provide atomic and molecular level details for increasingly larger complexes (Moore, 2017). Finally, increased use of remote sensing data and macro level imaging of ecological systems are being increasingly used to help understand ecosystem dynamics. (Kennedy et al., 2014).

To enhance all of the above techniques, a time dimension has been added to better reveal structural dynamics and change representing the flexibility and disordered components of a unified model. Compared to static snapshots, structural dynamics is the more likely predictor of function. If we are to better predict function of unknown structures, it is essential to include dynamic information in our measurements and represented in predictive homology models.

What is a biological function? We refer to function as the role that structure plays within the biological entity. This could be a molecular interaction, product production, or building blocks for larger structures. Depending on the context, many biological components may have more than one functional output. In most cases, we would only consider those functions that evolved through natural selection and contribute to organismal fitness. However, the development of techniques to accurately predict evolutionarily new (or old) or previously uncharacterized functions would fundamentally change our understanding of biological systems and would be a great platform to reintegrate biology.

Why is this viewpoint poignant? We have reached a state across many scientific fields where more than enough data have been generated to start synergizing predictive output models. Biologists have been working for over two centuries to develop such models and have produced many examples that have come about through advances in microscopy that revealed cellular structures, to the establishment of the GenBank or the Protein Data Bank to deposit genome sequences or protein structures, respectively, to the visualization of cellular dynamics, to the development of tissue models and engineering of organ growth. The progression has ultimately moved from visualization of a static state to recognition of the importance of dynamics for function. At the cellular level, many organelle structures and correlating functions have been defined and can be predictive across cell types. For recently emerging substructures including microgranules and biomolecular condensates, predicting function would be highly valuable. We have reached a tipping point, where we can now begin to define commonalities in structure across various biological levels that can be predictive of function. However, we need a common language and definitions that can transcend biological levels of structure that can be used to better predict function.

Reliance on homology to predict function. One concept that might be limiting our ability to characterize structures of unknown function is our primary reliance on similarity and homology.

Similarity can be in the form or shape, or pattern of a structure that for macromolecules is usually assessed by the primary sequence of components. Homology is arrival of structural and functional similarity through descent from a common ancestor. When we encounter a structure of interest, a molecule, a gene, or an organ, for which the function is unknown, we attempt to understand it through comparison of similarity to known structures at the given level. For example, when we identify genes in the genome that are conserved at least among close relatives, but for which a function has not been discovered, we compare its sequence to a database of sequences that have been annotated with a function. The problem is that available databases of known functions is limited to structures that have been previously identified and we ascribe that known function to similar structures. Thus, new functions are not identified, and the reasoning becomes circular in nature. While this example refers to genes, our ways of knowing functions of structures at higher and lower levels of organization are also limited by the reliance on similarity or homology.

One way forward could be to improve our ways to detect similarities by improving databases of structures with ascribed function. The problem is that these functions are usually discovered in a limited number of organisms in a limited number of conditions; usually a model species in a laboratory setting. Because of our reliance on laboratory experiments, our knowledge of known functions is limited. We can only characterize those structures that are similar to what is known. The result is that we gather more things that are similar to what we know and we tend to exclude uncharacterized things that are not. We need to expand the ways we characterize functions of known structure. In our gene example, this might be to include more different conditions, preferably more natural conditions, and to capture metadata on those conditions more explicitly in available databases. In addition, better characterizations of functions in a more diverse set of organisms might also expand the databases of structures of known function.

In terms of core, flexibility, and disorder in genomes, highly conserved genes could be thought of as the core set and retain stable and homologous protein structures. Genome flexibility is introduced through mutations and domain swaps that impart new functions based on one or more central core structures. Introduction of flexible loops and insertions are more variable and contribute more to specificity of function. Finally, we include aspects of disorder that can be introduced into protein coding genes that are generated by a variety of mechanisms, including exon swapping, accumulation, or skipping or lateral gene transfer that can provide increased disorder and result in the dynamic accumulation of new functions. Similar arguments can be made for noncoding genomic regions which may also introduce flexibility through the acquisition of mutations that affect promoter, regulatory, or other unknown functions. It is also known that heterochromatic regions of the genome are more prone to base substitution mutations (*i.e.* core ⊏ flexible), while open euchromatin more rapidly accumulates larger genomic changes including indels and substitutions (Makova and Hardison, 2015) (*i.e.* more disorder).

We could also expand our understanding of function by better integrating knowledge of complex interactions among structures or molecules with changes in the biochemical and physiological states of cells. Currently, we gather increasing amounts of data about the structures or molecules that function within biological systems using various “omic” methods. We characterize genomes, transcriptomes, and metabolomes of various organisms under a variety of conditions. While individually, these data are extremely useful for understanding organismal functions, we need to integrate the various types to better understand the functions of the individual structures. For example, how do changes in metabolites affect gene expression, and how are these dependencies affected by polymorphisms? This is an active area of investigation (Wu et al., 2019) and there is promise that these methods will become more robust, however current databases do not adequately capture those dependencies. Improvement in this area

would accelerate progress on multiple fronts, including a more complete understanding of the structure-function relationship. Structural knowledge of other molecules is similarly limited.

Dynamics. A critical aspect of structure-function relationship is dynamics. This relates directly to time-dependent motion of biomolecular structures such as domain motions and loop dynamics, or at a higher scale, assembly and disassembly of macromolecular structures, all of which are essential for their functions. Many proteins have modular design where a conserved core motif serves as a chassis on which more mobile peripheral loops and subdomains play more determinate roles for their functions (Branden and Tooze, 1999). Likewise, multidomain proteins connected by flexible hinge or linker carry out domain motion. For example, matrix metalloproteinase (MMP) cleaves the collagen triple helix; and in plants, cellulase cleaves cellulose. Both MMP and cellulase have two domains connected by a linker (Henrissat 1994; Murphy and Knäuper, 1997). And the dynamic interaction between the two domains and their substrates is necessary for cleavage. Also, certain multidomain viral proteins assemble into different structures to carry out different functions during the infection cycle, which is a strategy to overcome the limited genomic information that viruses can carry to provide more flexibility of function (Carrillo, et al., 2014).

A prime example of dynamics at the meso-scale is the mitotic spindle, which is the subcellular machine that drives cell division. The mitotic spindle is dynamically organized and maintained by constant polymerization and depolymerization of microtubule filaments, a phenomenon named as `dynamic instability' (Burbank and Mitchison, 2006). There are anti-cancer drugs targeting microtubule polymerization dynamics (Vicente and Wordeman, 2019), highlighting the functional importance of dynamics.

In a broader sense, structural dynamics also refers to time-dependent genomic organization and pattern formation in higher-order structures. In these cases, dynamics is described by time-dependent equation of smaller-scale constituents such as chemical concentrations, interaction, or behavior of individual cells as variables, rather than directly using measures of shape (Murray, 2003). In these cases, `structures' are emergent properties of the smaller scale constituents that are important for their collective function.

Recent developments in computational methods are enabling us to reconcile a plethora of structural and functional data into mechanistic understanding. For example, a typical molecular dynamics simulation involves a system consisting on the order of 10^5 atoms (including water molecules), with simulation time reaching microseconds. While this is far limited compared to the experimental length and time scale, the computational capability is constantly improving. Likewise, different experimental methods have their own limitations. It is thus crucial to recognize the limitations of different modalities and combine them to advance our understanding of the role of dynamics for mapping between structure and function.

At a higher scale where macroscopic shapes and patterns are measured, resolution of images becomes less of an issue. But these structures are emergent properties through underlying processes at the molecular and cellular level. Mathematical theories linking the properties and interactions of individual agents to the emergent behaviors of the ensemble of the agents exist, such as nonlinear dynamics (Murray, 2003) and theory of stochastic processes (van Kampen, 2007). We thus have ample computational, theoretical, and experimental tools and data to better understand structure-function-dynamics relationships. Again, the challenge is to how to suitably integrate disparate methodologies and concepts. Among many, one example area that need further development in order to tackle mesoscale phenomena is image analysis. Computational capability to extract quantitative structural information is lagging far behind the

ability to produce high-resolution images. Examples include reconstructing the microtubule network based on cryo-electron microscopy images (Lučić, et al., 2013) and neuronal networks (Hildebrand, et al., 2017), which are still done largely by manual mouse-clicking by human operators. In relatively simple and routine cases, machine learning methods are showing promises (Ronneberger, et al., 2012), but they are difficult to apply to quantitative segmentation and measurements of fine structures such as the mitotic spindle and neuronal network, or for amorphous structures where landmarks needed for machine learning are poorly defined, for which different analysis approach would be needed.

Broader impacts. One intention is to create a universal way to think about structure that can be better predictive of function across levels of biological organization (Engler et al., 2009). Are specific aspects of structure more predictive of function than others? For example, can methods for predicting protein structure be applied at other scales to guide predictions for currently unknown or undiscovered tissue, organismal, or ecological functions; as suggested in (Kim et al., 2019)? If underlying principles are discovered, this may allow us insight into the functions of structures preserved in the fossil record, remnants of genes remaining in the genome, or images obtained from other galaxies. At the same time, we should also consider the inverse problem of whether function can be predictive of structure.

Interdisciplinary knowledge will be required to strengthen our understanding of function, which can then feedback onto the understanding of important principles of structure. Our current databases contain information from only a few organisms or systems under a limited set of conditions, and integration of information from divergent species and environments may lead to significant insights into our understanding of the structure-function relationship. To achieve this, developing multidisciplinary quantitative biology curricula are also needed. Such programs encompassing multiple departments are emerging in some universities and institutes; a trend which should be expanded so that we can better and more rapidly understand structures at all levels of biology that provide function to our world.

Acknowledgments: This collaborative effort was a result of the Reintegrating Biology Jumpstart Meeting supported by the National Science Foundation (NSF) #1940791, held in Austin, TX in 2019. Individual research support is from the NSF (1613534 to M.A.T.) and the National Institutes of Health (R15GM135791 to M.A.T.; R01AI136301 and R01GM127723 to W.H.)

REFERENCES

Branden, C.I. and Tooze, J. (1999). Introduction to protein structure, 2nd ed. (Garland Science).
Burbank, K.S., and Mitchison, T.J. (2006). Microtubule dynamic instability. Curr Biol *16*, R516-517.

Carrillo, B., Choi, J.M., Bornholdt, Z.A.,Sankaran, B., Rice, A.P., and Prasad, B.V.V. (2014). The influenza A virus protein NS1 displays structural polymorphism. J Virol *88*, 4113-4122.

Engler, A.J., Humbert, P.O., Wehrle-Haller, B., and Weaver, V.M. (2009). Multiscale modeling of form and function. Science *324*, 208-212.
Goldsmith-Fischman, S., and Honig, B. (2003). Structural genomics: computational methods for structure analysis. Protein Sci *12*, 1813-1821.

Goth, W., Lesicko, J., Sacks, M.S., and Tunnell, J.W. (2016). Optical-Based Analysis of Soft Tissue Structures. Annu Rev Biomed Eng *18*, 357-385.
Henrissat, B. (1994). Cellulases and their interaction with cellulose. Cellulose 1, 169-196.

Hildebrand, D.G.C., Cicconet, M., Torres, R.M., Choi, W., Quan, T.M., Moon, J., Wetzel, A.W., Champion, A.S., Graham, B.J., Randlett, O., Plummer, G.S., Portugues, R., Bianco, I.H., Saalfeld, S., Baden, A.D., Lillaney, K., Burns, R., Vogelstein, J.T., Schier, A.F., Lee, W.C.A., Jeong, W.K., Lichtman, J.W., and Engert F. (2017). Nature *545*, 345-349.

Huang, B., Bates, M., and Zhuang, X. (2009). Super-resolution fluorescence microscopy. Annu Rev Biochem *78*, 993-1016.
Kempfer, R., and Pombo, A. (2019). Methods for mapping 3D chromosome architecture. Nature Reviews Genetics.
Kennedy, R.E., Andrefouet, S., Cohen, W.B., Gomez, C., Griffiths, P., Hais, M., Healey, S.P., Helmer, E.H., Hostert, P., Lyons, M.B.*, et al.* (2014). Bringing an ecological view of change to Landsat-based remote sensing. Front Ecol Environ *12*, 339-346.
Kim, H., Smith, H.B., Mathis, C., Raymond, J., and Walker, S.I. (2019). Universal scaling across biochemical networks on Earth. Sci Adv *5*, eaau0149.
Lučić, V., Rigort, A., and Baumeister, W. (2013). Cryo-electron tomography: the challenge of doing structural biology *in situ*. J Cell Biol, *202*, 407-419.
Makova, K.D., and Hardison, R.C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. Nat Rev Genet *16*, 213-223.
Moore, P.B. (2017). Structural biology: Past, present, and future. New biotechnology *38*, 29-35.
Murphy, G. and Knäuper, V. (1997). Relating matrix metalloproteinase structure to function: why the “hemopexin” domain? Matrix Biol *15,* 511-518.
Murray, J.D. (2003). Mathematical biology II: spatial models and biomedical applications (3rd ed., Springer).
Ronneberger, O., Liu, K., Rath, M., Rueβ, D., Mueller, T., Skibbe, H., Drayer, B., Schmidt, T., Filippi, A., Nitschke, R., Brox, T., Burkhardt, H., Driever, W. (2012). ViBE-Z: a framework for 3D virtual colocalization analysis in zebrafish larval brains. Nat Methods, *9*, 735-742.

van Kampen, N.G. (2007). Stochastic processes in physics and chemistry (3rd ed., Elsevier).

Vicente, J.J., and Wordeman, L. (2019). The quantification and regulation of microtubule dynamics in the mitotic spindle. Curr Opin Cell Biol *60*, 36-43.
Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019). A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High Throughput *8*.