Can sequence specific and dynamics-based metrics representing the conformational ensemble allow us to decipher the encoded function in IDP sequences?

S. Banu Ozkan[1]

[1]. Department of Physics Center for Biological Physics Arizona University, Tempe AZ

Among the biomolecules common to all living organisms on Earth, proteins conduct a diverse array of functions that their nucleotide-based counterparts do not. What makes proteins Nature's miracle is their specific amino-acid sequence, which encodes all their biophysical properties. Traditionally sequence-function relation is decoded by structural characterization of proteins in their native states However, not all protein sequences require a structure in their native (functional) state. Such structure-independent, a.k.a. intrinsically disordered proteins (IDPs), are critical in many biological functions and constitute a considerable fraction of the proteome (1–3). Despite two decades of extensive studies, it remains challenging to decipher sequence-function code for due to failures in classical sequence-structure-function mapping. (3). Lacking an ordered compact form in the unbound state along with exhibiting structural promiscuity in bound forms brings challenges in structural alignments (2). Furthermore, IDPs appear to evolve faster than structured proteins, as insertion and deletion of amino acids are more common in IDPs, thereby complicating sequence alignments (3).

What unifies IDPs and structured proteins is the conformational ensemble in the functional state (i.e., native state for structural proteins) and associated equilibrium dynamics (Figure 1A). This ensemble, dictated by the 1-D sequence, underlies the function. For structural proteins, sampling a native state ensemble through all-atom simulations or coarse-grain models of experimentally determined 3-D interactions allows us to decipher the sequence- function paradigm (4). However, this process is much more challenging for IDPs due to large conformational space. Despite the success of development of coarse-grain models and modifications in all-atom forcefields to sample IDP conformations (5), accurate representation of the conformational ensemble is still under debate (6,7). Thus, sequence-based first principle theoretical models which describe ensemble features of IDPs are desperately needed to decipher their sequence-encoded functions.

In this issue Huihui and Ghosh(14) address this daunting challenge by expanding their earlier theoretical work (8), a first principle approach of constructing a Hamiltonian based on the IDP sequence; particularly, positions of the charged amino acids. The Hamiltonian also includes two body excluded volume interaction and repulsive three-body interaction that allow the two-body interaction to be attractive (9). This fundamental analytical formalism models one of the critical aspects of sequence, the charge distribution with long-range electrostatic interactions. This charge distribution captures the conformational heterogeneity of IDPs as well as changes in their conformational ensemble features at different temperature, and salt concentration as well as phosphorylation, a major post-translational process of IDPs (9,10). Particularly, the success of this heteropolymer theory is that it analytically computes the ensemble average distance profiles $<R_{ij}^2>$ between any two amino acids for IDPs as a function of the sequence. Remarkably, it also produces similar distance profiles obtained computationally expensive all atom simulations (8). Thus, these higher resolution distance profiles provide a detailed description of the IDP in contrast to typical metrics such as scaling exponents, radius of gyration, or end-to-end distance.

Since this theoretical framework provides the much needed high-dimensional conformational ensemble features (i.e., any pairwise distance profile analytically derived directly from sequence), Huihui and Ghosh have tested whether the model can distinguish functional versus non-functional IDP sequences of various IDP systems (8). They constructed the sequence charge decoration matrix (SDCM), an interaction matrix that captures the specific electrostatic interactions for each IDP sequence which dictate conformational heterogeneity and subsequently clustered these matrices using principle component analysis. Clustering reveals that SDCM successfully groups functional IDP sequences together and can distinguish them from non-functional groups in two distinct IDP families: Ste50 (11) and PSC (12). Finally, they also applied the same SDCM to the synthetic variants of RAM sequences sharing the same net charge and showed that their approach is in agreement with experimental classification of measured binding data (13). Previously, the number of contiguous negative charges were used to discriminate between repressive and non-repressive PSC systems, whereas the basal net charge was used as a marker for functional Ste50s. While this charge patterning that was previously used to classify RAM sequences yielded some correlation with experimental functional characterization, including binding affinities and transcriptional activation values, Huihui and Ghosh present SCDM as an universal algorithm which can successfully discriminates the functional IDPs based on sequences in all three protein systems . This strongly suggests that SCMD is a powerful approach to decipher sequence-function paradigm of IDPs. The general applicability of SCDM could be a result of two possible factors. First, as discussed above, SCDM is a high-dimensional set of metrics that can tease out cryptic features encoded in the sequence. Second, SCDM captures critical ensemble average conformational features by taking into account every pairwise interaction dictating the conformational ensemble properties. A recent work of Cohan et al. (14) also highlights importance of the accurate representation of the conformational ensemble to map IDP sequence-function and addresses this by constructing an information entropy-based matrix. Using joint and single probability distributions of ensemble features which include radius of gyration, end to end distance and asphericity obtained from all-atom simulations this matrix can functionally cluster different RAM sequences. This type of computational all-atom simulation-based approach is complimentary to the easily obtainable analytical form developed by Huihui and Ghosh.

Efforts toward accurate representation of the conformational ensemble of IDPs through combined computational and experimental methods (15) is indeed critical ,as it paves a way to provide the blue prints of IDPs. A vast majority of studies on resurrected ancestral globular proteins (16) and intrinsically disorder proteins (17) also present that evolution proceeds through the modulation of conformational dynamics, particularly the fine-tuning of protein ensembles facilitated by mutations. Moreover, evolutionary selection proceeds using the principle of “minimum perturbation, maximum response” through allosteric control of functional site dynamics (i.e. binding /catalytic sites) by mutating distal positions, rather than introducing mutations at functional sites themselves (4). Taken together, these analyses suggest that Nature’s main tool in utilizing the evolutionary landscape to alter function through variations in sequence is allosteric regulation.

Compared to structural proteins, the broader native state ensembles from which IDPs can sample also allows for a much larger mutational landscape to modulate protein dynamics towards a desired function. This also makes them key player in allosteric regulation(18,19) Therefore, it is no surprise to observe that the fraction of IDPs increases in evolution from Bacteria and Archaea to Eukaryota proteomes/ Indeed, controlling protein ensembles of IDPs through different perturbations including mutations, posttranslational modification and/or different ionic strength in cells makes IDPs the most valuable players of a cell in their ability to orchestrate any complex processes that Nature may frequently exploit. (Figure 1B). The rigorous analytical form developed by Huihui and Ghosh (8-10) is therefore a first step towards providing not only a complete understanding of the relationship between an amino acid sequence of IDPs and the function it encodes, but also may allow us to explain how allostery spatiotemporally regulates global signal propagation within a cell. Deciphering the sequence-function relationship of IDPs would find immediate application to many long-standing problems including the understanding and treatment of disease, design of novel drugs, the development of green syntheses of commodity chemicals, and new energy sources.

Figure 1: (A) Specific IDP sequence underlies the conformational ensemble, and dynamics within this ensemble dictates function. Unlike structured proteins, ensemble is much more broader. (B) This broader ensemble allows to exploit dynamic allostery where different perturbations in the sequence such as change in ionic strength, post-translational modification, mutations lead to a shift in conformational ensemble. This change in conformational ensemble modulates functions.

Acknowledgement

S.B.O. acknowledges support from the Gordon and Betty Moore Foundations and National Science Foundation (Award: 1715591 and 1901709).

References:

1. Csizmok, V., A.V. Follis, R.W. Kriwacki, and J.D. Forman-Kay. 2016. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem. Rev.* 116:6424–6462.

2. Schuler, B., A. Borgia, M.B. Borgia, P.O. Heidarsson, E.D. Holmstrom, D. Nettels, and A. Sottini. 2020. Binding without folding – the biomolecular function of disordered polyelectrolyte complexes. *Current Opinion in Structural Biology*. 60:66–76.

3. Light, S., R. Sagit, O. Sachenkova, D. Ekman, and A. Elofsson. 2013. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol*. 30:2645–2653.

4. Campitelli, P., T. Modi, S. Kumar, and S.B. Ozkan. 2020. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. *Annual Review of Biophysics*. 49:267–288.

5. Das, R.K., K.M. Ruff, and R.V. Pappu. 2015. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol*. 32:102–112.

6. Nerenberg, P.S., and T. Head-Gordon. 2018. New developments in force fields for biomolecular simulations. *Curr Opin Struct Biol*. 49:129–138.

7. Rahman, M.U., A.U. Rehman, H. Liu, and H.-F. Chen. 2020. Comparison and Evaluation of Force Fields for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* 60:4912–4923.

8. Huihui, J., and K. Ghosh. 2021. Intra-chain interaction topology can identify functionally similar Intrinsically Disordered Proteins. *Biophysical Journal*.

9. Huihui, J., and K. Ghosh. 2020. An analytical theory to describe sequence-specific inter-residue distance profiles for polyampholytes and intrinsically disordered proteins. *J. Chem. Phys.* 152:161102.

10 Huihui, J., T. Firman, and K. Ghosh. 2018. Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. *J. Chem. Phys.* 149:085101.

11. Zarin, T., B. Strome, A.N. Nguyen Ba, S. Alberti, J.D. Forman-Kay, and A.M. Moses. 2019. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife*. 8:e46883.

12. Beh, L.Y., L.J. Colwell, and N.J. Francis. 2012. A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *PNAS*. 109:E1063–E1071.

13 Sherry, K.P., R.K. Das, R.V. Pappu, and D. Barrick. 2017. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *PNAS*. 114:E9243–E9252.

14 Cohan, M.C., K.M. Ruff, and R.V. Pappu. 2019. Information theoretic measures for quantifying sequence–ensemble relationships of intrinsically disordered proteins. *Protein Engineering, Design and Selection*. 32:191–202.

15 Lazar, T., E. Martínez-Pérez, F. Quaglia, A. Hatos, L.B. Chemes, J.A. Iserte, N.A. Méndez, N.A. Garrone, T.E. Saldaño, J. Marchetti, A.J.V. Rueda, P. Bernadó, M. Blackledge, T.N. Cordeiro, E. Fagerberg, J.D. Forman-Kay, M.S. Fornasari, T.J. Gibson, G.-N.W. Gomes, C.C. Gradinaru, T. Head-Gordon, M.R. Jensen, E.A. Lemke, S. Longhi, C. Marino-Buslje, G. Minervini, T. Mittag, A.M. Monzon, R.V. Pappu, G. Parisi, S. Ricard-Blum, K.M. Ruff, E. Salladini, M. Skepö, D. Svergun, S.D. Vallet, M. Varadi, P. Tompa, S.C.E. Tosatto, and D. Piovesan. 2021. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Research*. 49:D404–D411.

16. Modi, T., P. Campitelli, I.C. Kazan, and S.B. Ozkan. 2021. Protein folding stability and binding interactions through the lens of evolution: a dynamical perspective. *Current Opinion in Structural Biology*. 66:207–215.

17. Hultqvist, G., E. Åberg, C. Camilloni, G.N. Sundell, E. Andersson, J. Dogan, C.N. Chi, M. Vendruscolo, and P. Jemth. 2017. Emergence and evolution of an interaction between intrinsically disordered proteins. *eLife*. 6:e16059.

18. Li, J., J.T. White, H. Saavedra, J.O. Wrabl, H.N. Motlagh, K. Liu, J. Sowers, T.A. Schroer, E.B. Thompson, and V.J. Hilser. 2017. Genetically tunable frustration controls allostery in an intrinsically disordered transcription factor. *eLife*. 6:e30688.

19. Wright, P.E., and H.J. Dyson. 2015. Intrinsically Disordered Proteins in Cellular Signaling and Regulation. *Nat Rev Mol Cell Biol*. 16:18–29.