CellPress

Molecular Cell

Commentary

# Intrinsically disordered regions (IDRs): A vague and confusing concept for protein function

Kevin Struhl[1,*]
[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: kevin@hms.harvard.edu
https://doi.org/10.1016/j.molcel.2024.02.023

**The term "intrinsically disordered region" (IDR) in proteins has been used in numerous publications. However, most proteins contain IDRs, the term refers to very different types of structures and functions, and many IDRs become structured upon interaction with other biomolecules. Thus, IDR is an unnecessary, vague, and ultimately confusing concept.**

Two fundamental principles of proteins were established in the 1950s, neither of which were obvious beforehand. Fred Sanger discovered that proteins are precisely ordered sequences of amino acids.[1,2] Max Perutz[3] and John Kendrew[4] demonstrated that proteins can have the intrinsic ability to fold into three-dimensional structural domains that can be solved in atomic detail by X-ray crystallography. Structured domains have multiple high-specificity interactions, indicating evolutionary selection and biological function. Their atomic details are the basis of enzymatic catalysis, interactions with small molecules resulting in structural changes, and other biochemical functions, leading to the initial premise that protein functions are mediated by structured domains.

Structured domains are relatively resistant to proteases, and protease cleavage has often been used to generate isolated domains amenable to high-resolution structural analysis. However, early estimates[5] and AlphaFold-based predictions[6,7] estimate that 30%–50% of the proteome is not part of structured domains; such regions—termed unstructured, disordered, or unfolded—are easily digested by proteases. Work over the past two decades has indicated that protein regions exist in a continuum of states that range across rigid structured domains, dynamic structured domains, partially folded or molten globules, disordered regions with transient structures, and disordered statistical coils.[8] The term "intrinsically disordered region" (IDR) was originally defined as a physicochemical property of a protein sequence,[9] often identified by its hydrophobicity and charge content.[10] As this definition of IDR solely depends on the sequence, it does not incorporate any notion of biological, biochemical, or genetic function.

Over the past few years, the term IDR has been used indiscriminately in thousands of publications concerned with protein function. It lumps all unstructured regions into a single negative idea—i.e., the absence of a folded domain—even though some disordered regions are biologically important whereas others are not. Furthermore, proteins with a structured domain typically have unstructured regions, multiple regions within an individual protein can be unstructured, and the physicochemical properties of sequences exist in a continuum; hence, the identity, number, and locations of IDRs are arbitrary. Lastly, the word "intrinsic" (defined by the Merriam-Webster dictionary as "belonging to the essential nature or constitution of a thing") is unnecessary, meaningless, and misleading—unnecessary because structured domains are defined by their intrinsic behavior, so unstructured regions are also intrinsically determined; meaningless because there is no difference between intrinsically or non-intrinsically disordered regions; and misleading because, as discussed below, some so-called IDRs become structured upon specific interactions with other proteins, nucleic acids, or other biomolecules. In a biological context, a region of a protein that is conditionally structured cannot be intrinsically disordered.

Although it was originally believed that proteins carry out their biological function via precise interactions within structured domains, it became clear in the 1980s that this is frequently not the case. Consequently, the term "domain" is now generally used to denote a protein region that carries out a biological function. Two early examples of functional, but not intrinsically structured, domains are those mediating mitochondrial import[11] and transcriptional activation.[12] In both cases, the functional regions of the proteins are too short to form a structured domain and are proteolytically sensitive. Strikingly, many different sequences suffice to mediate the same function, indicating that these protein regions have some common sequence preferences with a general structural property.[11–13]

As initially proposed, these functional domains, though unstructured on their own, become structured upon interaction with the appropriate target protein(s). For transcriptional activation domains, the key interactions involve hydrophobic and aromatic amino acids that lead to an amphipathic $\alpha$-helical structure when bound to cognate surfaces of target proteins, but the precise sequence does not matter.[14] In a slightly different example, the DNA-binding surface of bZIP proteins is not part of a structured domain, but rather undergoes a folding transition to an $\alpha$ helix upon high-affinity binding to its target DNA sites.[15,16] More generally, numerous protein-protein interactions critical for biological functions are mediated by relatively short "adhesive surfaces"[17] that are unfolded on their own but form a defined and relatively stable structure upon interaction.

More recently, some so-called IDRs are critical for the formation of biomolecular

condensates, which are large macromolecular entities that typically contain proteins involved in a common biological process.[18,19] Condensates are often large enough to be microscopically visualized in cells as discrete non-membranous bodies, but they are distinct from other large entities such as non-specific protein aggregates and virus particles that have specific structures. The structural basis for condensates involves a very large number of weak and multivalent interactions between protein components. Unlike protein-protein interactions involving short adhesive surfaces, condensates are highly dynamic entities that adopt multiple conformations due to different constellations of multivalent interactions.

Thus, in the biological context, it is extremely confusing that IDR can refer to regions that either (1) are conditionally ordered, (2) mediate very different types of protein interactions, (3) behave as linkers between domains, or (4) have no known interactions and may have no biological function at all. As such, I suggest that the term IDR be restricted to studies on protein structure *per se* and otherwise be replaced by more useful and descriptive terms linked to biological, biochemical, or genetic functions. The domain(s) of protein X that mediate adhesive interactions with a domain(s) of protein Y should simply be called "Y interaction domain" and vice versa. Domains can also be defined by their biological function, such as transcriptional activation or mitochondrial import domains. Regions that are sufficient for condensate formation should simply be called "condensate forming domains," perhaps with an additional descriptor that indicates the specific type of condensate. In cases where domains need to be separated by sequences between them, linker region is the current and preferred term. Lastly, protein regions of no apparent function should simply be ignored until such time as a function is identified.

Lemke et al. (this issue of *Molecular Cell*)[20] correctly note that some protein regions can have multiple biological or biochemical functions. For example, a Y interaction domain could also be an A or B interaction domain, and such a domain could be regulated by post-translational modifications and perform distinct biological functions (e.g., transcriptional activation or repression depending on the context). This situation is analogous to a single gene having multiple biological functions, often associated with different names depending on the history of its identification and/or the functions. This is a well-known and inherent nomenclature issue, and it is not a reason to avoid using functional terms to describe protein regions.

This suggested terminology, some of which is already used, is simple to understand and much more informative than IDR. Semantics matter for scientific clarity.

## ACKNOWLEDGMENTS

I thank Sun Hur and Steve Harrison for thoughtful discussion and comments on the manuscript. This work was supported by a grant to K.S. from the National Institutes of Health (GM131801).

## DECLARATION OF INTERESTS

The author declares no competing interests.

## REFERENCES

1. Sanger, F., and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J. *49*, 481–490.

2. Sanger, F., and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J. *49*, 463–481.

3. Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., and North, A.C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature *185*, 416–422.

4. Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., and Shore, V.C. (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. Nature *185*, 422–427.

5. Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztányi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., et al. (2013). D(2)P(2): database of disordered protein predictions. Nucl. Acids Res. *41*, D508–D516.

6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

7. Ruff, K.M., and Pappu, R.V. (2021). AlphaFold and implications for intrinsically disordered proteins. J. Mol. Biol. *433*, 167208.

8. Babu, M.M., Kriwacki, R.W., and Pappu, R.V. (2012). Structural biology. Versatility from protein disorder. Science *337*, 1460–1461.

9. Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., et al. (2013). What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. Intrinsically Disord. Proteins *1*, e24157.

10. Uversky, V.N. (2002). What does it mean to be natively unfolded? Eur. J. Biochem. *269*, 2–12.

11. von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. EMBO J. *5*, 1335–1342.

12. Hope, I.A., and Struhl, K. (1986). Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell *46*, 885–894.

13. Ma, J., and Ptashne, M. (1987). A new class of yeast transcriptional activators. Cell *51*, 113–119.

14. Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2020). A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. Mol. Cell *78*, 890–902.e6.

15. Patel, L., Abate, C., and Curran, T. (1990). Altered protein conformation on DNA binding by Fos and Jun. Nature *347*, 572–575.

16. Weiss, M.A., Ellenberger, T., Wobbe, C.R., Lee, J.P., Harrison, S.C., and Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature *347*, 575–578.

17. Ptashne, M., and Gann, A. (2002). Genes & Signals (Cold Spring Harbor Press).

18. Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

19. Sabari, B.R., Dall'Agnese, A., and Young, R.A. (2020). Biomolecular condensates in the nucleus. Trends Biochem. Sci. *45*, 961–977.

20. Lemke, E.A., Babu, M.M., Kriwacki, R.W., Mittag, T., Pappu, R.V., Wright, P.E., and Forman-Kay, J.D. (2024). Intrinsic disorder: A term to define the specific physicochemical characteristic of protein conformational heterogeneity. Mol. Cell *84*. https://doi.org/10.1016/j.molcel.2024.02.024.