Identifying key residues in intrinsically disordered regions of proteins using machine learning

Wen-Lin Ho[1], Hsuan-Cheng Huang[2], and Jie-rong Huang[1,2,3,*]

[1]Institute of Biochemistry and Molecular Biology, National Yang Ming Chiao Tung University, No. 155 Section 2, Li-nong Street, Taipei, Taiwan
[2]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, No. 155 Section 2, Li-nong Street, Taipei, Taiwan
[3]Department of Life Sciences and Institute of Genome Sciences, National Yang Ming Chiao Tung University, No. 155 Section 2, Li-nong Street, Taipei, Taiwan

To whom correspondence should be addressed: jierongh@nycu.edu.tw

Abstract

Conserved residues in protein homolog sequence alignments are structurally or functionally important. For intrinsically disordered proteins (IDPs) or proteins with intrinsically disordered regions (IDRs), however, alignment often fails because they lack a steric structure to constrain evolution. Although sequences vary, the physicochemical features of IDRs may be preserved in maintaining function. Therefore, a method to retrieve common IDR features may help identify functionally important residues. We applied un-supervised contrastive learning to train a model with self-attention neuronal networks on human IDR orthologs. During training, parameters were optimized to match sequences in ortholog pairs but not in other IDRs. The trained model successfully identifies previously reported critical residues from experimental studies, especially those with an overall pattern (e.g. multiple aromatic residues or charged blocks) rather than short motifs. This predictive model can therefore be used to identify potentially important residues in other proteins.

Availability and implementation

The training scripts are available on GitHub (https://github.com/allmwh/IFF). The training datasets have been deposited in an Open Science Framework repository (https://osf.io/jk29b). The trained model can be run from the Jupyter Notebook in the GitHub repository using Binder (mybinder.org). The only required input is the primary sequence.

Introduction

DNA/RNA sequences and the proteins they encode carry their evolutionary history, and multiple sequence alignment methods can reveal phylogenetic relationships. For instance, our extinct Neanderthal and Denisovan cousins were identified from the DNA extracted in ancient bones [1, 2]; prokaryotic ribosomal 16S RNA sequences contributed to the discovery of Archaea domain [3]; the tracing of myoglobin and hemoglobin protein sequences back to their globin origin is another textbook example [4, 5]. Protein structures also provide insights into how proteins have evolved, being conserved in some cases despite changes in the primary sequence. For example, the structural similarity between the motor domains of kinesin and myosin hints that they have a common ancestor, despite low sequence identity [6]. The shape of a protein also constrains how it evolves and functionally important residues are conserved. Indeed, when sequence conservation levels are mapped onto 3D structures, the most conserved residues are typically found in key locations, such as the folding core [7] or catalytic sites [8].

However, these structural constraints on evolution do not apply to intrinsically disordered proteins (IDPs) or proteins with intrinsically disordered regions (IDRs), and as a result, the sequences of these proteins, which represent more than half of the proteome [9], vary more widely than do those of their folded counterparts (see example in Supplementary Figure S1). Although some structural evolutionary restraints still apply to some IDRs, especially those that undergo folding-upon-binding [10, 11], the evolution of IDRs is mainly constrained by function. One recently recognized function of IDRs is their ability to undergo liquid-liquid phase separation (LLPS) [12, 13]. This mechanism contributes to the formation of membraneless organelles and explains the spatiotemporal control of many biochemical reactions within a cell [14, 15]. The proteins within these condensates do not adopt specific conformations (i.e. they still behave like random coils) [16, 17] and thus evolve without structural restraints. Although multiple sequence alignment may work in some instances (for example, the aromatic residues in the IDRs of TDP-43 and FUS are conserved, highlighting their potential importance for LLPS [18]), most IDRs cannot be aligned, especially when there are sequence gaps between orthologs [19].

The functionally important physicochemical properties of IDPs/IDRs encoded in their primary sequence may be retained during evolution. Aromatic residue patterns [20], prion-like amino-acids [21], charged-residues blocks [22], and coiled-coil content [23] all contribute to LLPS, but these features cannot be revealed by sequence alignment. Multiple sequence alignment methods are, therefore, of limited use in identifying critical residues in IDRs. To overcome this challenge, we propose an unsupervised contrastive machine learning model trained using self-attention neuronal networks on human IDR orthologs. Our results show that the trained model “pays attention” to crucial residues or features within IDRs. We also provide online access to our model that uses primary sequences as input.

Methods

Training dataset preprocessing

Human protein sequences were retrieved from UniProt [24] and the corresponding orthologs were obtained from the Orthologous Matrix (OMA) database [25]. Chordate orthologs were aligned using Clustal Omega [26]. The PONDR [27] VSL2 algorithm was used to predict the IDR of the human proteins and to define the boundaries of the aligned sequences (Figure 1A). Aligned regions were defined as subgroups. N-terminal methionines were removed to assist learning (methionine is coded by the start codon in protein synthesis). After removing gaps within the aligned sequences, all sequences were padded to a length of 512 amino acids (repeating from the N-terminus; Figure 1A). The few sequences longer than 512 amino acids (56,086 out of 2,402,346, 2.3 %) were truncated from the C-terminus. The training dataset thus consisted of 28,955 ortholog subgroups from 13,476 human protein families with IDRs longer than 40 amino acids.

Each training batch consisted of fifty randomly selected subgroups (Figure 1B). The human sequence from each subgroup was paired with one of its orthologs (one of the non-human sequences in the same subgroup, Figure 1C). The selection probability was weighted by the Levenshtein distance [28] from the human sequence to favor low similarity pairings. Supplementary Figure S2 shows how different the sequences typically were in these ortholog subgroups, along with the corresponding selection probabilities. The most dissimilar sequences (high probability of being selected for training) in each ortholog group were also deposited in Open Science Framework. A classifier token (CLS) was added to the start of the selected sequences, and these were mapped to a matrix with an embedding dimension of 128 (*embed_dim*; Figure 1C).

Training architecture

The training architecture was a self-supervised contrastive learning model, Momentum Contrast version 3 (MoCo v3) [29]. The base encoder in MoCo v3 was replaced with a classical self-attention network [30]. We used 8-head attention and tested six attention layers. Fifty human sequences from the same batch and their corresponding orthologs (the ones with the lowest similarity to each human sequence, as mentioned above) were sent to the momentum encoders (f_q, f_k respectively, following the original nomenclature [29]), and calculated in parallel (Figure 1D). The outputs from each human sequence and its ortholog were a query (q) and key (k+; the positive sample for each query). The output of the other 49 orthologs were the negative samples (k–). All 50 combinations of q, k+, and k– were formulated to minimize a contrastive loss using the adopted InfoNCE [31]:

$$\mathcal{L}_q = -\log \frac{\exp\,(q{\cdot}k^{+}/\tau)}{\exp\,(q{\cdot}k^{+}/\tau)+\sum_{k^-}\exp\,(q{\cdot}k^{-}/\tau)} \qquad (1)$$

where τ is a temperature hyper-parameter (set to 0.02). The loss was computed in a symmetrized manner [29], i.e. the human sequences (q) were also sent to f_k, and the orthologs (k) were sent to the f_q with correspondent outputs for calculating the InfoNCE loss. The parameters between the attention layers of f_q (light purple blocks in Figure 1D) were updated according to a gradient to minimize the cross-entropy loss (Equation (1)). The parameters in f_k (dark purple blocks) were updated by the momentum encoder: (1–m) • query_encoder + m • key_encoder, with m set to 0.999 by default [29]. This scheme (Figure 1B–D) was repeated ~580 times to include all 28,955 subgroups in each training epoch. The training consists of 400 epochs, and the InfoNCE loss is sufficiently converged (Supplementary Figure S3). The model was built on PyTorch and the training was performed on a Nvidia Telsa P100 16G GPU.

Results

The trained model attributes a high attention score to experimentally confirmed critical residues.

Studies have shown that the aromatic residues (phenylalanine, tyrosine, and tryptophan) in the IDRs of TDP-43 [32], FUS [33], and hnRNP-A1[34] are critical for LLPS-related functions. These residues obtain a high attention score in our model (Figure 2A). The aromatic residues (two tryptophans and ten tyrosines) in galectin-3 [35] also score highly (Figure 2B, left panel). Interestingly, although zebrafish galectin-3 differs substantially in primary sequence from human galectin-3 (Supplementary Figure S4), the aromatic residues (mostly tryptophan instead of tyrosine) also have high attention scores (Figure 2B, right panel). Note that zebrafish galectin-3 was not in the OMA ortholog database used for training (OMA number: 854142). Charged residues (purple arrows in Figure 2C) reported to be associated with condensation in NPM1 [36], FMRP [37], and Caprin1 [38] also obtain high attention scores (Figure 2C). Our model also assigns high attention scores to the methionines in Pbp-1 (labeled in Figure 2D; Pbp-1 is the yeast ortholog of human

Ataxin-2), which have been shown to be critical for redox-sensitive regulation [39]. Altogether, these results indicate that the trained model correctly identifies known key IDR residues.

Most amino acids have broadly distributed attention scores except tryptophan and cysteine, whose presence in IDRs hints at potential importance.

Figure 2E compares the attention score distributions of the amino acids in human IDRs. The differences are striking, but the attention scores are not correlated with other physical properties, such as disorder/order propensity [40, 41], prion-likeness [42], or prevalence in human IDRs (Supplementary Figure S5). The attention scores of alanine are always low. Although alanine promotes α-helix formation, which is known to contribute to IDR functions such as LLPS [43-45], our model ignores these residues. This is probably because α-helices are also promoted by other amino-acid types, e.g. leucine or methionine [46, 47], in different combinations not involving alanine. Also, since the training process did not include structure information, structure-related sequence motifs were ignored. At the other end of the distribution, tryptophan and cysteine systematically obtain high attention scores. These structure-promoting amino acids rarely appear in unstructured regions [40, 41, 48]; therefore, their appearance in IDRs hints at their potential importance. Although little is known about the role of cysteine in IDRs, its involvement in tuning structural flexibility and stability has been recently discussed [49]. Tryptophan, in contrast, is well-known to act as LLPS-driving “stickers” in IDRs [32, 50, 51], and bioinformatic analysis shows that they may have evolved in the IDRs of specific proteins to assist LLPS [18].

Finally, the fact that most amino acids, including those highlighted in Figure 2A–D, have broad attention score distributions (Figure 2E), excludes the possibility that our model is biased toward particular amino-acid types rather than sequence content as a whole. Moreover, in the machine learning procedure, the protein sequences were embedded into higher dimension matrices (as sequences of digits; Figure 1C), and amino-acid type information was lost when the matrices were transformed into tensors along with the self-attention layers (Figure 1D). These results support the predictive ability of the trained model.

Discussion

Genetic information, in the form of a linear combination of nucleic or amino acids, becomes more diverse over time. Comparing levels of diversity between different species reveals how closely related they are. In terms of amino acids, multiple sequence alignment not only highlights phylogenetic relationships between proteins but also facilitates homology modeling for structure prediction [52-54]. Machine learning approaches have recently been used to incorporate information from evolution to train structure prediction models [55, 56], and the highly accurate predictions from AlphaFold [57] and RoseTTAFold [58] have revolutionized structural biology. In contrast, the structural conformations of IDRs do not have a one-to-one correspondence with the primary sequence, and multiple sequence alignment often fails [18, 59]. These limitations make IDR structural ensembles challenging to predict. A few attempts have been reported, such as using generative autoencoders to learn from short molecular dynamics simulations [50]. The potential and challenges of machine learning in IDR ensemble prediction are also discussed [59].

Sequence pattern prediction faces similar challenges, including the lack of a sufficient stock of "ground-truth" training data, such as image databases or the Protein Data Bank. However, unsupervised learning architectures have been developed to train models without labeled datasets [60], and this type of approach is especially well-suited for IDRs. For instance, Saar et al. used a language-model-based classifier to predict whether IDRs undergo LLPS [61]. Moses and coworkers pioneeringly applied unsupervised contrastive learning, using protein orthologs as augmentation [62], to train their model to identify IDR characteristics [63]. Although we also used ortholog sequences as training data, our approach differs in many aspects. Instead of convolution neural networks, we used self-attention networks to capture the distal features in the entire protein sequence. Additionally, we trained our model using the latest contrastive learning architecture (MoCo v3), which greatly reduces memory usage for larger batches and enhances efficiency. In contrast to other masked language models [64-66], our approach is the first, to the best of our knowledge, to combine contrastive learning and self-attention in extracting features using natural language processing for protein sequence analysis. Our trained model directly "pays attention" to potentially critical residues in the entire sequence instead of mapping the primary sequence to learned motifs [63]. In other words, our model identifies overall features in an IDR sequence, for example, a predominance of aromatic residues or blocks of charged residues (Figure. 2). Moreover, our model provides intuitive results that point out potentially important residues for researchers to target for example in mutagenesis or truncation experiments.

Conclusion

Although the model could be improved by training on larger datasets (e.g., including more orthologs other than human's) or with larger batch sizes (requiring a supercomputer), these results show that self-supervised contrastive learning with self-attention networks can be used to identify key residues in IDRs, something that cannot be achieved by conventional multiple sequence alignment. The model, IFF – for *I*DP *F*eature *F*inder, can be accessed online using a primary sequence as the only input. We expect our model to be useful in various research fields, notably cell biology, to efficiently identify critical residues in proteins with IDRs, such as those that undergo LLPS.

Acknowledgments

The authors thank the IT Service Center at NYCU for GPU access. This work was supported by the National Science and Technology Council of Taiwan (110-2113-M-A49A-504-MY3). The authors are also grateful to Prof. Wen-Shyong Tzou (National Taiwan Ocean University) for his help.

Figure legends

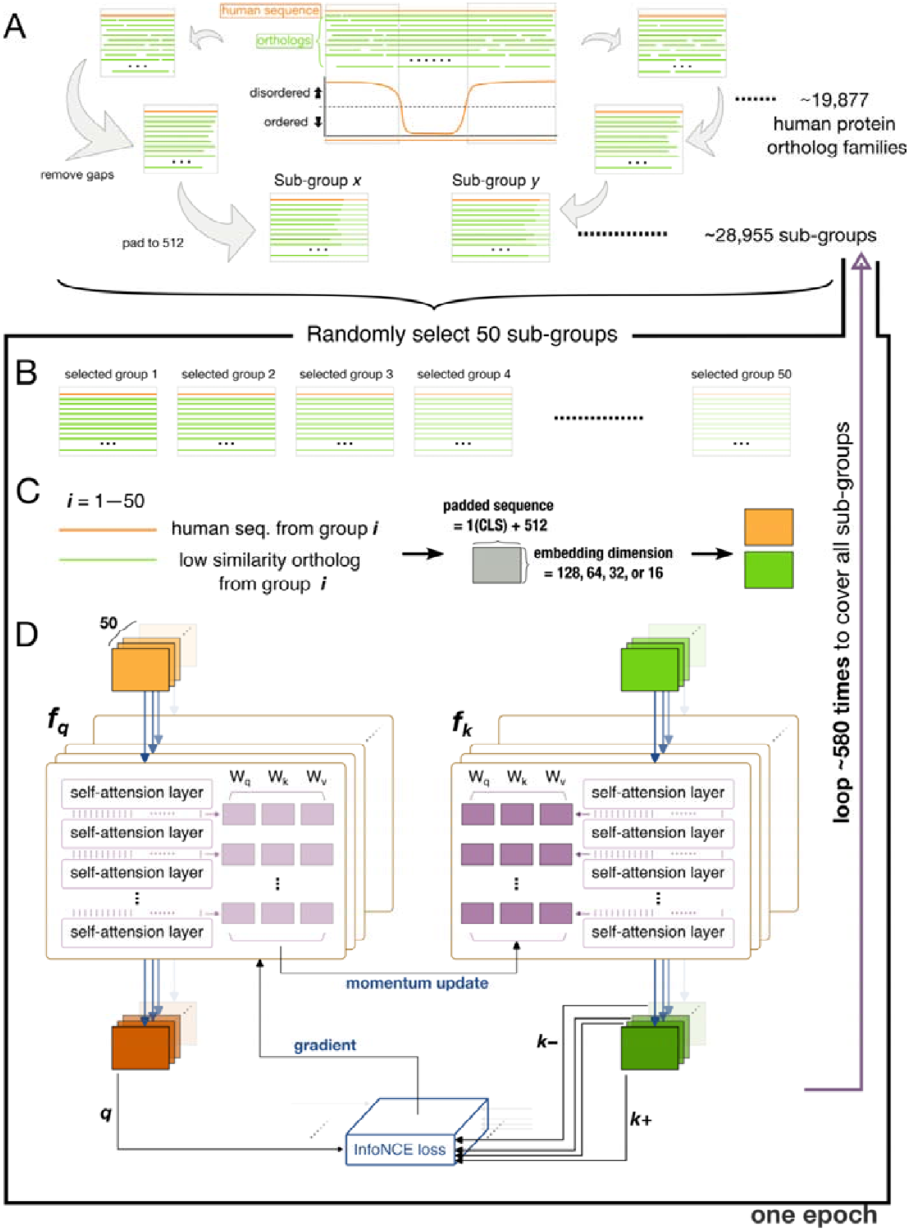

Figure. 1. Flowchart of the training scheme. (A) Schematic representation of how the training datasets were constructed from human sequences (orange lines) and orthologs (green lines). (B) A training batch made up of 50 randomly selected subgroups. (C) Embedding of the human sequence and one of its orthologs from the same subgroup (selection probability weighted by dissimilarity) to different dimensions (as a tensor for each sequence). (D) The architecture of the training model. The steps in panels B–D were repeated 580 times to cover all subgroups in the training set, and the whole process (a training epoch) was repeated 400 times.

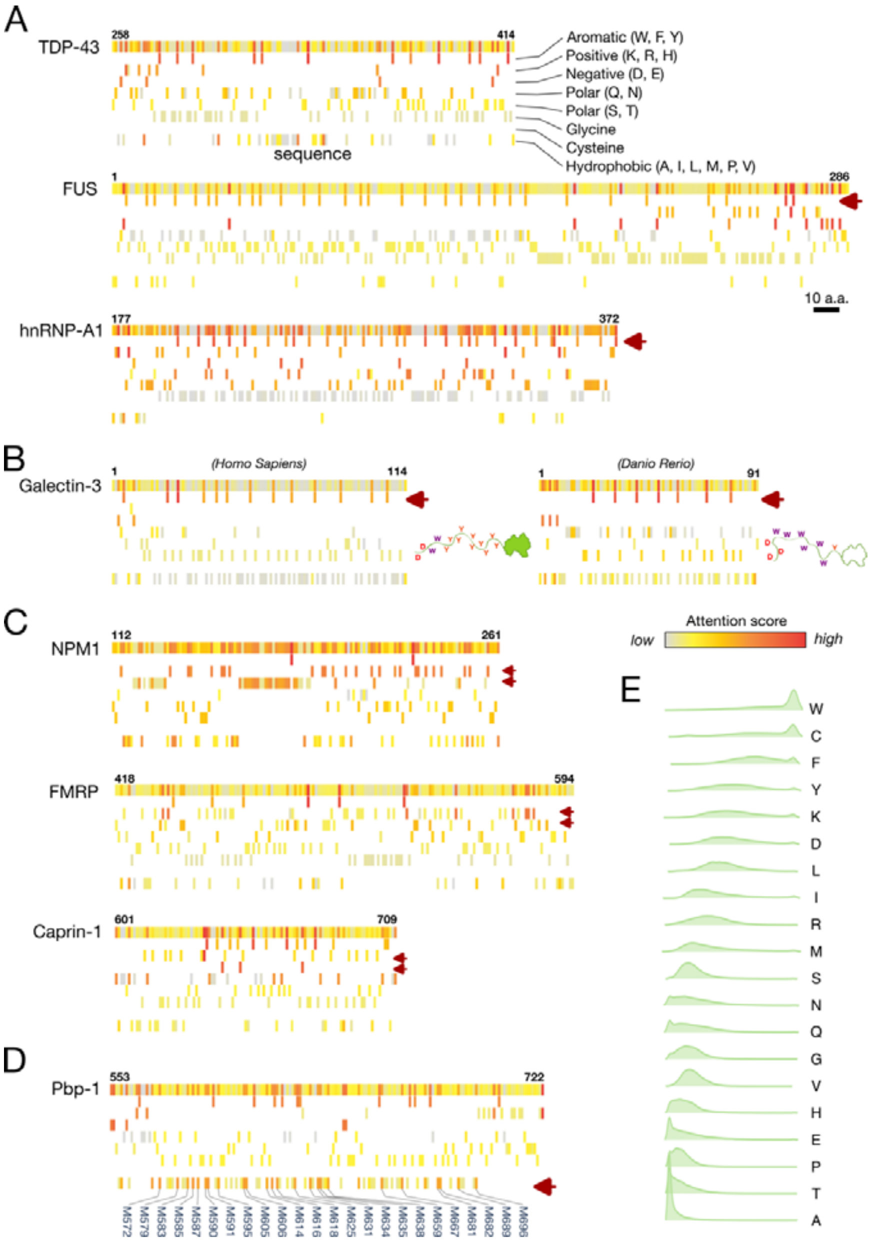

Figure 2. Results of the trained model for reference proteins and attention score distributions for individual amino acids. (A–D) Sequences and attention scores for the intrinsically disordered regions of (A) the RNA-binding proteins TDP-43, FUS, and hnRNP-A1, (B) human and zebrafish galectin-3, (C) NPMA, FMRP, and Caprin-1, and (D) Pbp-1. The attention scores appear as heatmaps from high (red) to low (grey) in the top row of each protein along with residue numbers. Amino acids with different physical properties are shown on separate rows as indicated in panel (A). Purple arrows indicate amino acids of known functional importance. (E) Half-violin plots of the distribution of attention scores in human IDRs for each amino acid, sorted by median value from high (tryptophan, W) to low (alanine, A).

References

1. Meyer, M., et al., *A high-coverage genome sequence from an archaic Denisovan individual.* Science, 2012. **338**(6104): p. 222-6.
2. Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. **328**(5979): p. 710-722.
3. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5088-90.
4. Suzuki, T. and K. Imai, *Evolution of myoglobin.* Cell Mol Life Sci, 1998. **54**(9): p. 979-1004.
5. Hardison, R.C., *Evolution of hemoglobin and its genes.* Cold Spring Harb Perspect Med, 2012. **2**(12): p. a011627.
6. Kull, F.J., et al., *Crystal structure of the kinesin motor domain reveals a structural similarity to myosin.* Nature, 1996. **380**(6574): p. 550-5.
7. Echave, J., S.J. Spielman, and C.O. Wilke, *Causes of evolutionary rate variation among protein sites.* Nat Rev Genet, 2016. **17**(2): p. 109-21.
8. Craik, C.S., et al., *The catalytic role of the active site aspartic acid in serine proteases.* Science, 1987. **237**(4817): p. 909-13.
9. Dunker, A.K., et al., *Intrinsic protein disorder in complete genomes.* Genome Inform Ser Workshop Genome Inform, 2000. **11**: p. 161-71.
10. Jemth, P., et al., *Structure and dynamics conspire in the evolution of affinity between intrinsically disordered proteins.* Sci Adv, 2018. **4**(10): p. eaau4130.
11. Karlsson, E., et al., *The dynamic properties of a nuclear coactivator binding domain are evolutionarily conserved.* Commun Biol, 2022. **5**(1): p. 286.
12. Alberti, S. and A.A. Hyman, *Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing.* Nat Rev Mol Cell Biol, 2021. **22**(3): p. 196-213.
13. Alberti, S., A. Gladfelter, and T. Mittag, *Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates.* Cell, 2019. **176**(3): p. 419-434.
14. Shin, Y. and C.P. Brangwynne, *Liquid phase condensation in cell physiology and disease.* Science, 2017. **357**(6357).
15. Banani, S.F., et al., *Biomolecular condensates: organizers of cellular biochemistry.* Nat Rev Mol Cell Biol, 2017. **18**(5): p. 285-298.
16. Burke, K.A., et al., *Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II.* Mol Cell, 2015. **60**(2): p. 231-41.
17. Brady, J.P., et al., *Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation.* Proc Natl Acad Sci U S A, 2017. **114**(39): p. E8194-E8203.

18. Ho, W.L. and J.R. Huang, *The return of the rings: Evolutionary convergence of aromatic residues in the intrinsically disordered regions of RNA-binding proteins for liquid-liquid phase separation.* Protein Sci, 2022. **31**(5): p. e4317.

19. Light, S., et al., *Protein expansion is primarily due to indels in intrinsically disordered regions.* Mol Biol Evol, 2013. **30**(12): p. 2645-53.

20. Martin, E.W., et al., *Valence and patterning of aromatic residues determine the phase behavior of prion-like domains.* Science, 2020. **367**(6478): p. 694-699.

21. Patel, A., et al., *A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation.* Cell, 2015. **162**(5): p. 1066-77.

22. Greig, J.A., et al., *Arginine-Enriched Mixed-Charge Domains Provide Cohesion for Nuclear Speckle Condensation.* Mol Cell, 2020. **77**(6): p. 1237-1250 e4.

23. Fang, X., et al., *Arabidopsis FLL2 promotes liquid-liquid phase separation of polyadenylation complexes.* Nature, 2019. **569**(7755): p. 265-269.

24. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.

25. Altenhoff, A.M., et al., *OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more.* Nucleic Acids Res, 2021. **49**(D1): p. D373-D379.

26. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**: p. 539.

27. Romero, Obradovic, and K. Dunker, *Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family.* Genome Inform Ser Workshop Genome Inform, 1997. **8**: p. 110-124.

28. Levenshtein, V.I. *Binary codes capable of correcting deletions, insertions, and reversals*. in *Soviet physics doklady*. 1966. Soviet Union.

29. Chen, X.L., S.N. Xie, and K.M. He, *An Empirical Study of Training Self-Supervised Vision Transformers.* 2021 Ieee/Cvf International Conference on Computer Vision (Iccv 2021), 2021: p. 9620-9629.

30. Vaswani, A., et al., *Attention Is All You Need.* Advances in Neural Information Processing Systems 30 (Nips 2017), 2017. **30**.

31. van den Oord, A., Y. Li, and O. Vinyals, *Representation Learning with Contrastive Predictive Coding.* CoRR, 2018. **abs/1807.03748**.

32. Li, H.R., et al., *TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues.* J Biol Chem, 2018. **293**(16): p. 6090-6098.

33. Lin, Y., S.L. Currie, and M.K. Rosen, *Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs.* J

Biol Chem, 2017.

34. Molliex, A., et al., *Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization.* Cell, 2015. **163**(1): p. 123-33.

35. Lin, Y.H., et al., *The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions.* J Biol Chem, 2017. **292**(43): p. 17845-17856.

36. Mitrea, D.M., et al., *Self-interaction of NPM1 modulates multiple mechanisms of liquid-liquid phase separation.* Nat Commun, 2018. **9**(1): p. 842.

37. Tsang, B., et al., *Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation.* Proc Natl Acad Sci U S A, 2019. **116**(10): p. 4218-4227.

38. Wong, L.E., et al., *NMR Experiments for Studies of Dilute and Condensed Protein Phases: Application to the Phase-Separating Protein CAPRIN1.* J Am Chem Soc, 2020. **142**(5): p. 2471-2489.

39. Kato, M., et al., *Redox State Controls Phase Separation of the Yeast Ataxin-2 Protein via Reversible Oxidation of Its Methionine-Rich Low-Complexity Domain.* Cell, 2019. **177**(3): p. 711-721 e8.

40. Vihinen, M., E. Torkkila, and P. Riikonen, *Accuracy of protein flexibility predictions.* Proteins, 1994. **19**(2): p. 141-9.

41. Radivojac, P., et al., *Intrinsic disorder and functional proteomics.* Biophys J, 2007. **92**(5): p. 1439-56.

42. Lancaster, A.K., et al., *PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition.* Bioinformatics, 2014. **30**(17): p. 2501-2.

43. Chiu, S.H., et al., *Phase separation driven by interchangeable properties in the intrinsically disordered regions of protein paralogs.* Commun Biol, 2022. **5**(1): p. 400.

44. Li, H.R., et al., *The physical forces mediating self-association and phase-separation in the C-terminal domain of TDP-43.* Biochim Biophys Acta, 2018. **1866**(2): p. 214-223.

45. Conicella, A.E., et al., *TDP-43 alpha-helical structure tunes liquid-liquid phase separation and function.* Proc Natl Acad Sci U S A, 2020. **117**(11): p. 5883-5894.

46. Pace, C.N. and J.M. Scholtz, *A helix propensity scale based on experimental studies of peptides and proteins.* Biophys J, 1998. **75**(1): p. 422-7.

47. Levitt, M., *Conformational preferences of amino acids in globular proteins.* Biochemistry, 1978. **17**(20): p. 4277-85.

48. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding.* Biochim Biophys Acta, 2010. **1804**(6): p. 1231-64.

49. Bhopatkar, A.A., V.N. Uversky, and V. Rangachari, *Disorder and cysteines in proteins: A design for orchestration of conformational see-saw and modulatory functions.* Prog Mol Biol Transl Sci, 2020. **174**: p. 331-373.

50. Wang, J., et al., *A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins.* Cell, 2018. **174**(3): p. 688-699 e16.

51. Sheu-Gruttadauria, J. and I.J. MacRae, *Phase Transitions in the Assembly and Function of Human miRISC.* Cell, 2018. **173**(4): p. 946-957 e16.

52. Balakrishnan, S., et al., *Learning generative models for protein fold families.* Proteins, 2011. **79**(4): p. 1061-78.

53. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.

54. Weigt, M., et al., *Identification of direct residue contacts in protein-protein interaction by message passing.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 67-72.

55. Xu, J., *Distance-based protein folding powered by deep learning.* Proc Natl Acad Sci U S A, 2019. **116**(34): p. 16856-16865.

56. AlQuraishi, M., *End-to-End Differentiable Learning of Protein Structure.* Cell Syst, 2019. **8**(4): p. 292-301 e3.

57. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

58. Baek, M., et al., *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.

59. Lindorff-Larsen, K. and B.B. Kragelund, *On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins.* J Mol Biol, 2021. **433**(20): p. 167196.

60. *Unsupervised Learning: Foundations of Neural Computation*, ed. G.S. Hinton, T. J. 1999: MIT Press.

61. Saar, K.L., et al., *Learning the molecular grammar of protein condensates from sequence determinants and embeddings.* Proc Natl Acad Sci U S A, 2021. **118**(15).

62. Lu, A.X.a.L., Alex X. and Moses, Alan, *Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning.* arXiv, 2020.

63. Lu, A.X., et al., *Discovering molecular features of intrinsically disordered*

regions by using evolution for contrastive learning. PLoS Comput Biol, 2022. **18**(6): p. e1010238.

64. Brandes, N., et al., *ProteinBERT: A universal deep-learning model of protein sequence and function*. Bioinformatics, 2022. **38**(8): p. 2102-10.

65. Elnaggar, A., et al., *ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning*. IEEE Trans Pattern Anal Mach Intell, 2022. **44**(10): p. 7112-7127.

66. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. Proc Natl Acad Sci U S A, 2021. **118**(15).