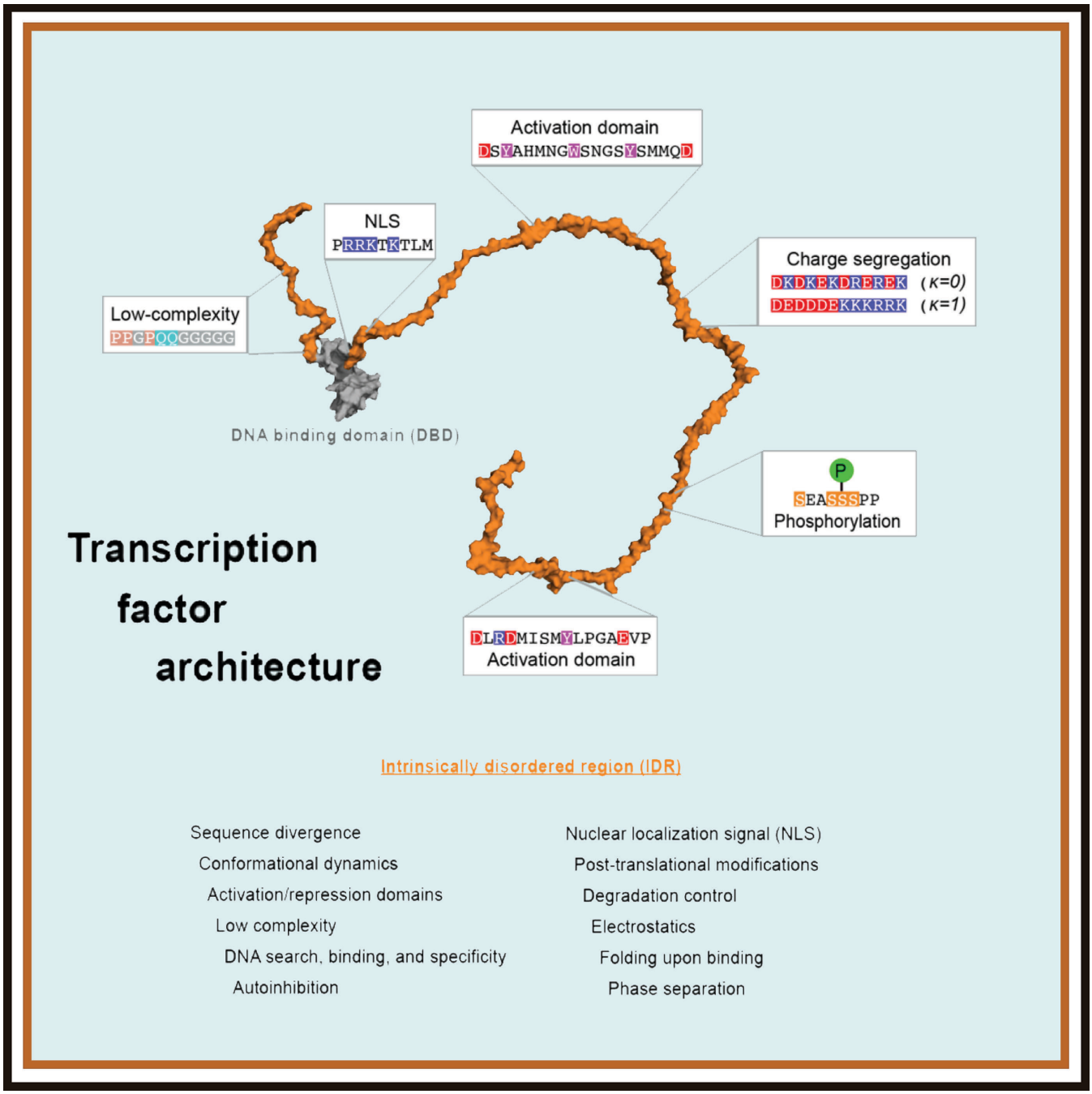
Chemistry–A European Journal
Review
doi.org/10.1002/chem.202203369
Chemistry Europe
European Chemical Societies Publishing
www.chemeurj.org

Multifunctional Intrinsically Disordered Regions in Transcription Factors

Matti Már,[a] Kateryna Nitsenko,[a] and Pétur O. Heidarsson*[a]



Abstract: Eukaryotic transcription factors (TFs) are the final integrators of a complex molecular feedback mechanism that interfaces with the genome, consolidating information for transcriptional regulation. TFs consist of both structured DNA-binding domains and long intrinsically disordered regions (IDRs) embedded with motifs linked to transcriptional control. It is now well established that the dynamic multifunctionality of IDRs is the basis for a wide spectrum of TF functions necessary to navigate and regulate the human genome. This review dissects the chemical features of TF IDRs that endow them with structural plasticity that is central to their functions in the nucleus. Sequence analysis of a set of over 1600 human TFs through AlphaFold was used to identify key features of their IDRs. Recent studies were then highlighted to illustrate IDR involvement in processes such as protein interactions, DNA binding and specificity, chromatin opening, and phase separation. To expand our understanding of TF functions, future directions are suggested for integrating experiments and simulations, from in vitro to living systems.

1. Introduction

Gene transcription, the process of converting a DNA sequence into mRNA that can be translated into a protein, is governed by an exquisitely complex and temporally coordinated network of around 1600 transcription factors (TFs). TFs recognize a specific sequence, for example within a promoter region, and subsequently recruit cofactors that activate or repress transcription.[1] Knowing how, where, and when TFs bind to the genome is crucial for understanding how they control gene expression. The majority of TFs interact with a multitude of different cofactors and can regulate many different genes in a cell-type specific manner.[1–3] Dysfunctions of TFs, are the cause of many human diseases from cancers and neurological disorders to cardiovascular diseases and obesity.[4]

Eukaryotic TFs share a global architecture which is organized into DNA-binding domains (DBDs) and effector domains (EDs),[5] which were famously described in 1988[6] by Paul Sigler as 'acid blobs and negative noodles'. Despite keen interest in understanding how TFs regulate gene expression, it remains challenging to determine how the precise genomic binding sites of TFs are specified and how TF binding ultimately relates to regulation of transcription. Sigler pointed out that the EDs were 'conformationally ill-defined' and their functions seemed to be less dependent on sequence but rather on composition such as numbers of negative charges. Since then, much has been learned about their sequence and structural features but how these relate to function has yet to be revealed in full detail, and in some sense Sigler's description of the elusive blobs still holds true. It is now widely recognized that the functional malleability of TFs is afforded by their structurally heterogenous EDs which are the primary functional units for controlling transcription.[5,7–9] Due to these flexible EDs, most TFs are classified as intrinsically disordered proteins (IDPs); they have one or more structured DBDs and usually long intrinsically disordered regions (IDRs) that can cycle dynamically between different conformations. Despite major technical challenges in studying disordered proteins, the many roles of the IDRs are increasingly coming into view, expanding beyond the classical model of orchestrating the transcriptional machinery.

In this review we focus on the rich biological output of IDRs in TFs, how they shape the conformational dynamics of the proteins and their search for correct binding sites, and even how they directly interact with DNA sequences.[1] As we will explore, the IDRs of TFs have a multifunctional role which is reflected in their dynamic conformational ensembles, allowing them to function as interaction hubs with other proteins and modulate the architecture of chromatin on a genome-wide scale. We limit ourselves to discussing TFs in eukaryotes and note that the general principles of their action are shared across many species (a recent review highlights the large differences with respect to bacterial TFs).[7] Our focus is mainly on the IDRs of TFs but it is worth pointing out that the DBDs of TFs are themselves frequently disordered in isolation.[10] For a more thorough overview of IDPs and IDRs, we refer the reader to excellent reviews on the topic.[11–15]

2. Structured and Disordered Regions in TFs

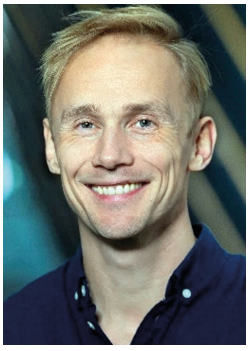
Protein structures are best described as a continuum of conformational states. One end of the spectrum contains highly structured proteins; these are rich in secondary structure and have generally fixed tertiary structure. By fixed tertiary structure we do not imply that it is necessarily static. Rather, a well-structured protein may adhere to a relatively narrow conformational space between structures whose atomic coordinates are highly similar, potentially with rare excursions to different conformations. In the other extrema of the structural spectrum are IDPs; they display little signs of fixed secondary or tertiary structures but instead dynamically explore a broad ensemble of conformational states, typically on a sub-millisecond timescale. The majority of proteins in the human proteome (~60%) exists between the two extremes, where structured domains are

[a] *M. Már, Dr. K. Nitsenko, Dr. P. O. Heidarsson*
Department of Biochemistry
Science Institute
School of Engineering and Natural Sciences
University of Iceland
Sturlugata 7, 102 Reykjavík (Iceland)
E-mail: pheidarsson@hi.is

separated or flanked by disordered regions.[16] It is important to note, however, that both structured and disordered regions can transiently adopt different conformations, either spontaneously or as a response to environmental cues, and these conformational changes can be critical for their function. TF architecture tends to cluster on the more disordered end of the structural spectrum as they usually contain long IDRs (Figure 1, Figure 2C).[17] It remains a major challenge to study large structurally heterogeneous polypeptides- these tend to be difficult to purify and they are usually not resolved in X-ray or cryo-electron microscopy (cryo-EM) studies. Nonetheless, integrative approaches have proven powerful to study dynamic protein ensembles, by combining ensemble experimental techniques (e.g., NMR and small-angle X-ray scattering) with single molecule approaches (e.g., single-molecule FRET), predictions, and computer simulations.[18–20]

Matti Már earned a BSc degree in Biochemistry and Molecular Biology from the University of Iceland in 2021 and is currently a MSc student in Biochemistry working under Associate Professor Pétur O. Heidarsson and Professor Eiríkur Steingrímsson. His MSc thesis research is focused on understanding the dynamic structure of transcription factor IDRs and their function using single-molecule fluorescence spectroscopy.

Dr. Kateryna Nitsenko obtained her BSc degree in medical radiation physics at the Kyiv National University after Taras Schevchenko, and a joint MSc degree in condensed mater physics at the Kyiv National University after Taras Schevchenko and the University of Strasbourg. She completed a PhD degree in biophysics and biochemistry from the University of Paris in 2020 as a Marie Skłodowska-Curie Actions fellow under Horizon 2020 funding program. She is currently a postdoctoral researcher at the University of Iceland in Pétur O. Heidarsson research group.



Dr. Pétur O. Heidarsson completed his BSc and MSc degrees in Biochemistry at the University of Iceland before completing a PhD degree in biophysics from the University of Copenhagen in 2013 under the supervision of Prof. Birthe B. Kragelund. He spent a short postdoctoral period in the same laboratory and as an EMBO Short-term fellow at the University of Cambridge with Prof. Jane Clarke. He then performed postdoctoral work in single-molecule biophysics in the group of Prof. Ben Schuler at the University of Zürich from 2015–2019. Pétur subsequently moved to Reykjavík to establish an independent research group at the University of Iceland. His research group uses single-molecule techniques to study transcription factor structure, function, and interactions with chromatin.

2.1. TF architecture

The DBD amino acid sequence is usually highly conserved within families and the structures of most DBDs are well characterized by classical structural biology techniques. Based on the structure of their DBDs, TFs have been broadly organized into 25 families including zinc-fingers (ZF), homeodomains, basic helix-loop-helix (bHLH), and forkhead.[1] The short cognate DNA sequence (6-12 bases) is usually recognized through one or more DBDs and even though the human genome contains millions of possible sites for each TF, they usually only occupy a subset of those[21] (discussed below). Even though disorder is much more prevalent in the EDs than in the DBDs,[5] many DBDs are disordered in isolation and then fold upon binding to DNA.[8,22] In these cases, it can be challenging to distinguish the details of the binding mechanism which can entail conformational selection, induced fit, or a mixture of both.[8,23]

The EDs, which are broadly classified into activator domains (ADs) and repressor domains (RDs), are much more divergent in sequence, which reflects both their multifunctional role and how they direct specific functions. The EDs are well documented to be enriched in structural disorder, where the majority of TFs have been predicted to contain extended disordered regions (Figure 2A,C).[17,24,25] Why are TF EDs so enriched in disorder, that is, what is the functional advantage to the TF in regulating transcription? The answer may lie in the unique physical and chemical features of such regions; i) IDRs assume a dynamic range of conformation, allowing for promiscuous interactions with several partners, enabling TFs to act as hubs to initiate transcription, ii) IDRs are frequently the sites of post-translational modifications (PTMs) which allows sensitive tuning of conformational ensembles and consequently functional output, iii) IDR-DBD interactions can enhance specificity and modulate protein-protein interactions (PPIs), iv) IDR dynamics allow multivalent interactions that can drive phase separation, and many more. These functions are not expected to be mutually exclusive in IDRs. For example, strong correlations have been found between transcriptional activation and phase separation.[26] Nonetheless, IDRs classically take on the role of ADs which function by recruiting coactivators to their binding site.[1,5] These regions are poorly conserved and often bind a wide array of coactivators.[3,5,27] We will explore some of the functional roles of TF IDRs that have been uncovered recently and show that TFs possess remarkable functional versatility that is enabled by their complex molecular architecture. We begin by looking at the amino acid sequence and use predictive tools to quantify structural disorder in human TFs.

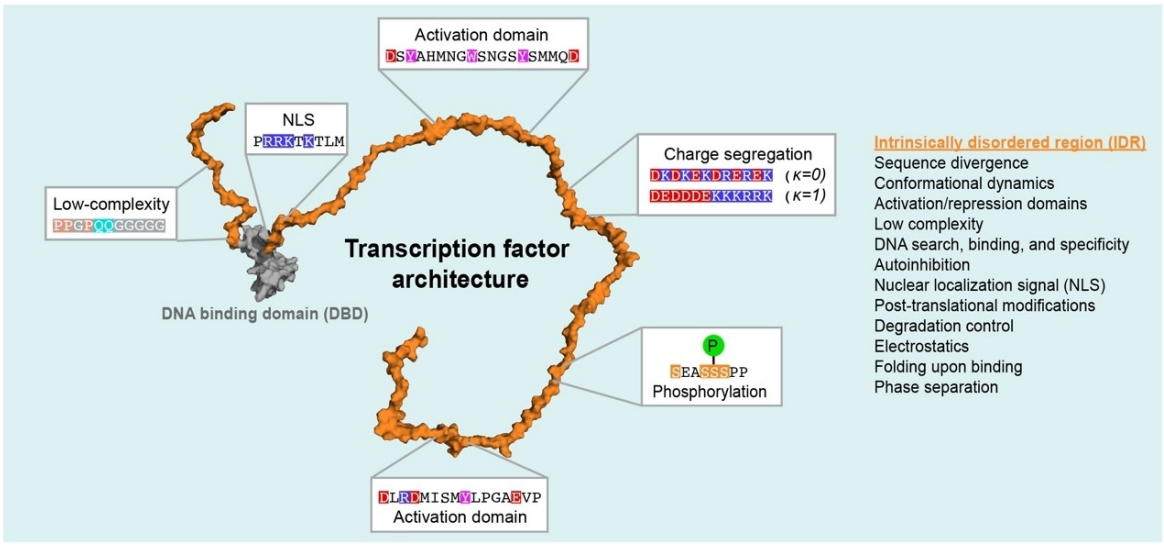


Figure 1. Schematic of general TF architecture and features of IDR sequence, structure, and function discussed in this review. The schematic model is based on the sequence and structure of the TF Sox2 (PDB 2LE4). Amino acids are coloured according to the RasMol "amino" colour scheme that groups together amino acids with similar physical properties.

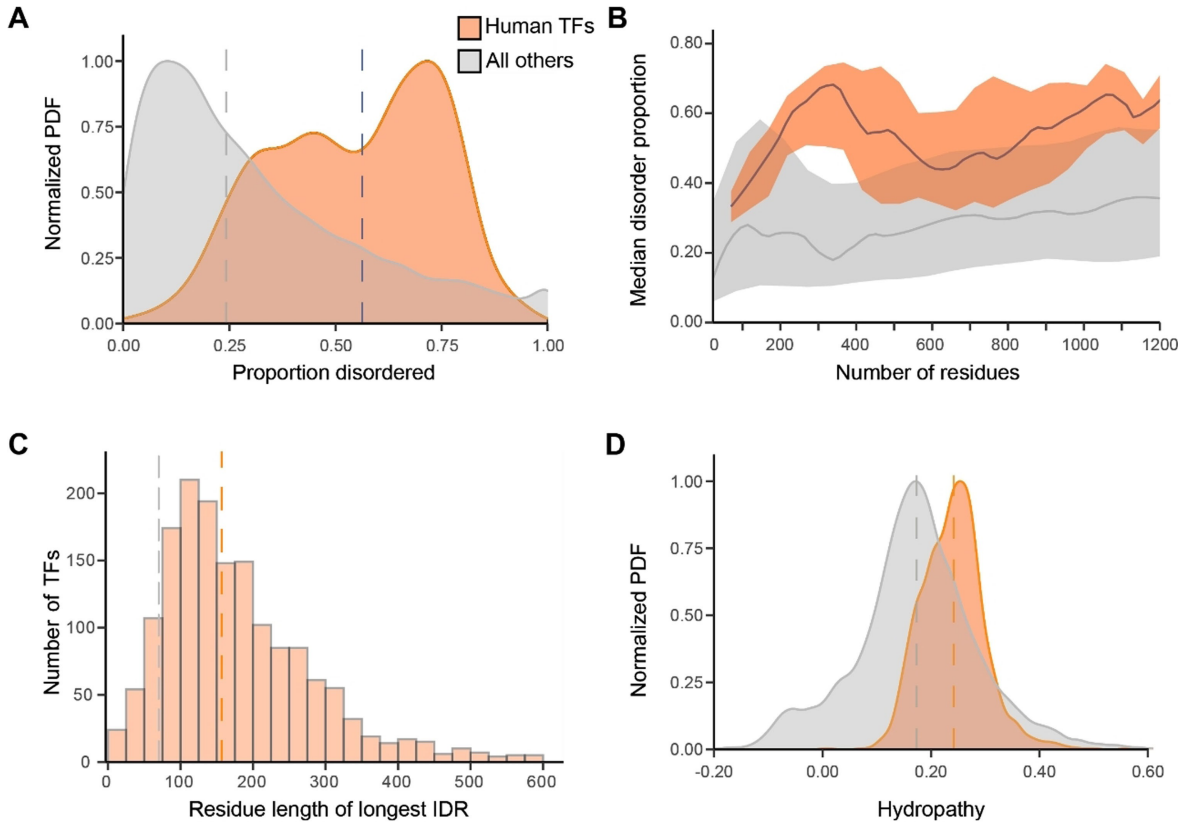


Figure 2. Properties of TF disorder. **A)** Plot comparing proportion of sequence predicted to be disordered by AF2 in human TFs (n = 1613) and the rest of the human proteome (n = 18514). The dashed lines represent the mean values. **B)** Plot comparing predicted median disorder proportion of the same datasets by residue length. Shaded areas represent the range from the 25th to 75th percentile of values. **C**) Histogram of the lengths of the longest IDRs in 1613 human TFs. Orange dashed line shows median value at 157 amino acids and grey dashed line shows 10th percentile value at 70 amino acids. **D**) Plot comparing the hydropathy (Wimley-White interface scale[35]) of human TFs to the rest of the human proteome.

2.2. Disorder prediction of TFs using AlphaFold

For sequence analysis we used a list of human TFs compiled by Lambert et al., with available reviewed entries fetched from Uniprot; for comparison we used a list of all reviewed human protein entries in Uniprot.[1,28] Disordered and ordered residues were predicted using AlphaFold2 (AF2) using the local per-residue estimate of confidence (pLDDT) values averaged over five amino acids as a proxy for disorder.[29–31] We considered pLDDT values above 67 to represent ordered residues,[31] and contiguous ordered or disordered residues were grouped together as ordered regions or IDRs. AF2 is a competitive predictor of disorder compared to existing predictors[31] and has a large available database of pre-computed pLDDT values for every human protein. However, it should be noted that a potential caveat of AF2 is that it may infer structure to be present that is only present in complex with a ligand.[31,32] AF2 also does not make any prediction on the structural ensemble of IDRs and conformationally mobile proteins, merely their lack of stable structure. We then used localCIDER to compute Das-Pappu kappa (κ) values for regions grouped by AF2, with $\kappa = 1$ corresponding to a maximally charge segregated sequence and $\kappa = 0$ a perfectly mixed one.[33,34] AF2 predicts all TFs to contain some structured regions and long IDRs, with enrichment for disorder across all length scales, which is reflected in the high hydropathy of TFs (Figure 2A–D). In what follows, we discuss some of our findings using AF2 predictions.

3. Sequence-Function Properties of TF IDRs

It is well known that IDRs are generally devoid in hydrophobic residues, rich in charged and polar residues, and often have low sequence complexity.[36] These features are general enough that whether a region is disordered in isolation can be robustly predicted by computational models such as Spot-Disorder2, AF2[29,31] and many others.[37] The improvement of such models is constrained by variations in the definition of IDRs. Some might consider a certain dynamic region an IDR, others simply a dynamic fold.[38] Adding to these human classification biases is the issue of biological context; a sequence may be disordered in the nucleus but adopt a folded state when bound to a partner. Determining whether a sequence is disordered is a starting point for determining how its sequence relates to function. Predictions of IDR functions from sequence have multiplied in recent years with predictors for regions that fold upon binding, DNA and RNA binding regions, and general function prediction.[39,40] Some of these are discussed in more detail in a later section. Though these methods may give predictive insight into the functions of TF IDRs they do not infer the dynamics and structure of IDRs that is vital to gaining a molecular understanding of their behaviour.

3.1. Charge effects on TF IDR ensembles

As IDRs populate dynamic and heterogenous structural ensembles, physical methods to understand their functions are ideally able to characterize the ensemble of conformations associated with each TF IDR or shared statistical properties (or distributions of these) of the ensemble. For these purposes, molecular dynamics approaches have proven very useful for investigating TF IDRs, giving deep insights into their dynamics and function, and synergistically improving with experimental approaches.[41,42] However, these methods are computationally expensive and are not applicable on proteome wide scales-simpler physical descriptions of IDR ensembles are also needed.

Some of the most notable models are built on observations that IDR ensemble dimensions are dominated by charges,[43] with net charge per residue (NCPR) and the fraction of charged residues (FCR) determining the position of a sequence in a "diagram of states" that predicts the expansion of the polyelectrolyte ensemble.[33,44] Plotting all TF IDRs that we identified with AF2 on such a diagram of states (Figure 3A) indicates that more than half of TF IDRs are weak polyampholyte globules and most of the remainder are somewhere between coil and globule. A comparison of TF IDR charge characteristics with all human IDRs (Figure 3B,C) shows that TF IDRs are enriched for negative net charges while being depleted in total charge content, consistent with the observed importance of acidic ADs.[45] The diagram of states is complemented by considering the charge decoration of the sequence,[46] which describes how well mixed negative and positive charges are. Molecular dynamics simulations have shown that sequences that segregate like charges together tend to adopt more collapsed states (Figure 1, charge segregation).[33,46] This collapse can be captured by the factor κ that encapsulates the charge segregation of the sequence (Figure 1). Plotting κ over FCR shows that IDRs tend to have more variable κ values than structured regions (Figure 3D). These analyses indicate that most TF IDRs lean towards a more charge poor, collapsed state than other human IDRs, while a substantial fraction are stronger polyampholytes that have more collapsed ensembles than their NCPR would imply due to charge clustering.[46] Despite these possibly collapsed ensembles, the dynamic nature of IDRs and the favourable solvation of charged side-chains will especially expose charged residues that have been found to be vital to their function.[33]

3.2. TF EDs are enriched in IDRs but IDRs are not always EDs

Though they have many varied functions, the central role of TF IDRs is transcriptional regulation, accomplished through their transcriptional EDs.[1] EDs are PPI hubs through which TFs bind cofactors.[3,27,47] EDs are traditionally divided into ADs and RDs depending on their effect on transcription when the ED is assayed.[5] Adding to this simple picture, some studies have found bifunctional domains that activate or repress transcription in a context or time dependent manner.[48,49] These results suggest that some EDs, perhaps far more than currently

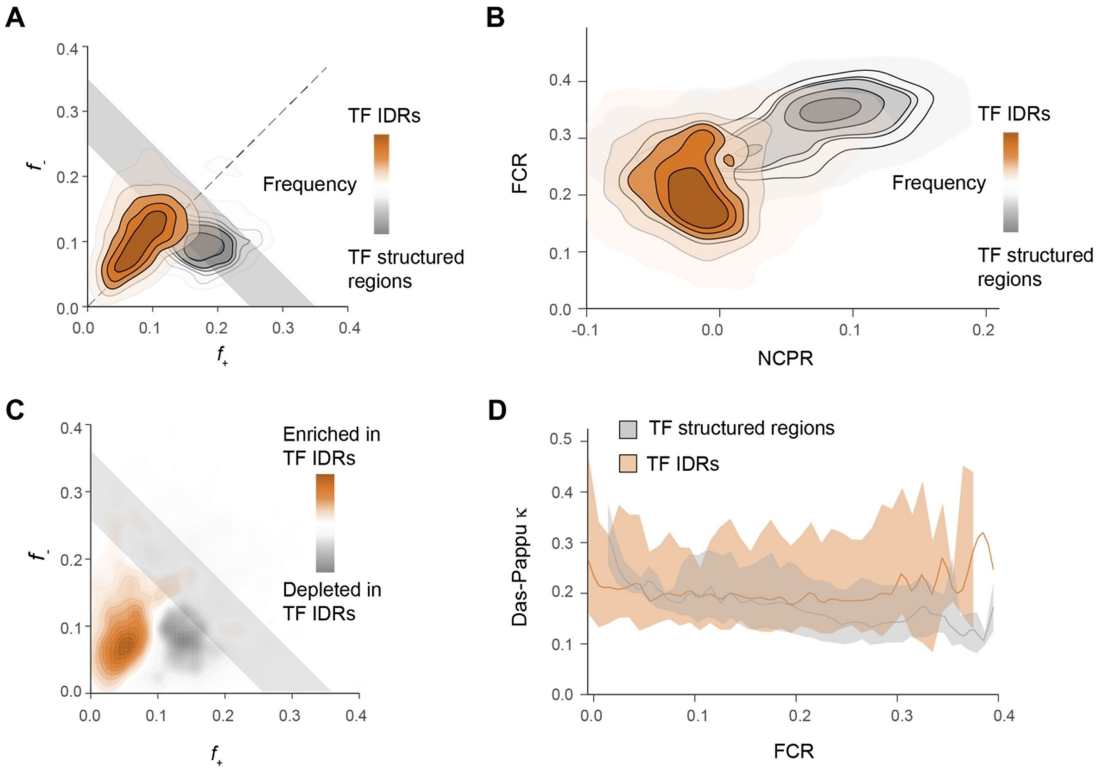


Figure 3. Charge properties of TFs. **A**) 2-dimensional (2D) density map comparing the fraction of negative (f_-) and positive (f_+) charges in human TF IDRs (n = 4177) and structured regions (n = 4109) longer than 20 residues. The regions below, above, and within the grey shaded area correspond to weak, strong, and intermediate polyampholytes, respectively. **B**) 2D density map comparing the fraction of charged residues (FCR) and net charge per residue (NCPR) of the same dataset. **C**) A 2D density difference map showing the relative enrichment of TF IDRs (n = 2792) longer than 50 residues compared to all human proteome IDRs (n = 17953) longer than 50 residues. Shaded areas same as A. **D**) Plot of Das-Pappu κ values by FCR for human TF IDRs and structured regions longer than 50 residues. Lines depict median values in the centre of a 0.01 FCR wide window and shaded regions depict the range between the 10th and 90th percentile values of the same window.

known, are sensitive to their environment and that some molecular features are likely to be common to both ADs and RDs.

From the early studies of eukaryotic transcriptional activation, it was recognized that most ADs lack any stable structure,[17] are generally acidic,[50] and hydrophobic.[51] Despite an intensive search for common motifs and grammar in the decades following Sigler's famous piece in 1988,[6] his description of ADs functioning "almost irrespective of sequence, provided only that there is a sufficient excess of acidic residues clustered or peppered about" still holds mostly true.

The frequent lack of structure in isolation among EDs has in some cases lead to erroneous conflation of IDRs with EDs. The confusion is added to by the gradual way in which activity can be reduced in heterologous assays and the context sensitivity of disordered motifs.[52] While studies using traditional experimental techniques have shed light on TF EDs, they cover only a subset of all TF EDs and lack a shared definition of what constitutes an "effector domain" other than the area must be able to activate or repress transcription.[5] This region can also contain parts that are not strictly necessary for function. To narrow down the sequence and physical chemical features of EDs, a more stringent and useful definition is the minimum amino acid sequence required for function. Such a dissection requires many overlapping regions of the sequence to be assayed for their activity and the minimum sequence required be deduced by the amino acids shared by all highly active regions.

3.3. Large-scale screens of EDs enable deep inferences of function and dynamics

Over the last few years, a spate of large-scale screens leveraging deep sequencing methods to parallelize TF ED dissections in yeast, *Drosophila melanogaster* (*dmel*) and human cells have been undertaken. These screens use a standard DBD and fuse to it a variable ED sequence (a 'tile') to investigate its function, mainly focusing on its activation properties (Figure 4A). ED constructs are varied, with some screens assaying whole TFs,[2,53] some tiling through sequences,[45,49,54,55] others using random sequences and combinations thereof,[56–58] and two screens have performed directed deep mutational scanning.[59,60] Combined, these screens have assayed millions of random and natural sequences for ED function and their results have enabled concrete steps towards prediction and classification of TF EDs.

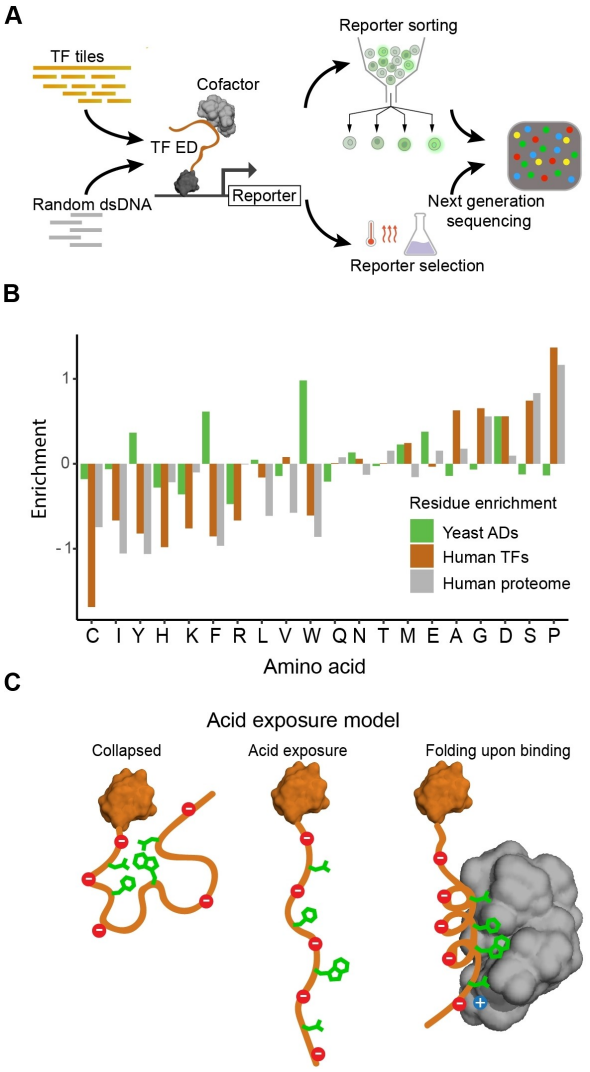


Figure 4. Screens of ED activity inform models for their function. **A)** A generalized scheme for core steps shared by most high-throughput screens of TF ED transcriptional activity. **B**) Residue enrichment in IDRs over structured regions in human TFs and all human proteins, defined as $Log_2(F_{Disordered}/F_{Ordered})$, as well as residue enrichment in active sequences in a screen performed by Ravarani et al., defined as $Log_2(F_{active}/F_{inactive})$.[56] **C**) In the acid exposure model proposed by Staller et al.,[59,60] ADs move between a buried and exposed state where the exposed state can bind cofactors and partially fold (adapted from Staller et al.[60]).

3.4. Acidic ADs in yeast contain solvent exposed aromatic residues in an acidic context

The focus of most large-scale screens has been on ADs, of which the acidic ADs have featured most prominently. In the case of Gcn4 and Gal4 ADs, these domains have been shown to fold upon binding to the Mediator subunit Med15 through fuzzy hydrophobic interactions of several regions of the AD with three activator binding domains on Med15.[47,61] Accordingly, all screens performed in yeast recover hydrophobic residues, particularly aromatic residues as key to activation with acidic residues a close second[45,56,57,59] (Figure 4B). The screens also find that the positioning of acidic and hydrophobic residues relative to each other is important, with acidic-hydrophobic residue pairs[56,57] and sequence proximity of the two being linked to activation.[59] These results generally concur with the *acid exposure model* in which acidic residues keep hydrophobic residues that interact with key coactivators exposed to solvent for binding (Figure 4C).[59] However, screens in yeast do not show enrichment for all hydrophobic residues, with the hydrophobic valine and isoleucine generally not enriched in active sequences unless adjacent to an acid and leucine is only mildly enriched.[57] Sanborn et al. extended their screen of yeast TF activation potential to an mRNA display screen for Med15 binding and found that a majority of activating tiles bound Med15 and that binding correlated with activation potential.[45] The requirement to bind Med15 has been invoked to explain why yeast ADs tend to favour aromatics that are more compatible with its shallow binding groove than other hydrophobic residues.[60]

3.5. Subtle motifs and structure in disordered ADs

Despite a preference for specific residues, de novo motifs as well as previously proposed motifs such as the 9aa TAD[62] are not commonly found in functional sequences in both random peptide screens and Sanborn et al.'s screen of yeast TFs.[45,56,57] In scramble mutants of yeast ADs, Sanborn et al. observed that most scrambles reduced activity little and only a third of ADs were sensitive to being scrambled, underscoring the importance of sequence composition over residue positioning. In addition to this lack of sequence motif specificity, random screens generally found a small correlation between helicity and activation potential, especially at the core of the AD.[56,57] Sanborn et al. found that proline insertions in some ADs strongly decreased activation, while most were unaffected. They noted that susceptibility to proline mutation was correlated with sensitivity to scrambling and the presence of basic residues, inferring that these ADs are likely sequestering their basic residues away from the binding interface by forming an α-helix (Figure 4C). They also found that when they assayed hydrophobic 9 amino acid peptides, a common amphipathic helical motif was associated with activation, and required for leucine rich peptide activation. This suggests that leucine-rich ADs rely on the folding of an α-helix, possibly explaining the depletion of isoleucine and valine in ADs, which are not favourable for helix formation.[45,63] Furthermore, acidic residues are more highly activating when located N-terminally to hydrophobic residues rather than C-terminally,[57] presumably stabilizing helical conformations by offsetting the backbone hydrogen bond dipole.[45,63] Such observations may be explained by the *acidic exposure model*, in which helical ADs are a convenient way of simultaneously exposing the maximal number of hydrophobic residues while hiding basic ones.[45,60]

3.6. Metazoan ADs share many similarities with yeast ADs alongside added diversity and complexity

Transferability of AD screening in yeast to higher eukaryotes suffers from the limited number of ADs found in yeast TFs and their almost universal acidic character. Models trained on yeast data fail to account for other classes of ADs found in higher eukaryotes.[49] Despite these differences, screens of metazoan sequences reveal an acidic and hydrophobic bias in activating sequences.[49,55,58,60] Human acidic ADs reveal an additional twist compared to yeast in that they show a stronger enrichment for leucine, possibly owing to differences in coactivators.[60] Acidic ADs in human TFs also fit within the *acid exposure model:* they display context-dependent increases in activity upon addition of hydrophobic residues that depends on the level of exposure afforded by acidic residues in the AD.[59,60] Substitutions that cannot be solvated drive hydrophobic collapse and reduce activity.[60]

Metazoan TFs are more commonly thought to harbour ADs distinct from the classical acidic ADs. These are usually classified by the residues they are enriched in, which tend to be disorder-promoting residues such as proline, glutamine and serine.[64,65] Compositionally biased regions and homopolymeric repeats are more commonly found in TFs than other classes of proteins.[2,66] These regions are found associated with activation in some screens of metazoan ADs but mechanistic detail of their function is lacking.[49,58] In their recent preprint, DelRosso et al. found that nearly half of all activating sequences contained a compositional bias.[49] Surprisingly they find little effect on activation from the deletion of homotypic repeats of hydrophilic residues but commonly find reduction in activation when leucine or alanine repeats are deleted, suggesting some repeats are not critical to AD function. Furthermore, when compositionally biased regions have their biased residue replaced with alanine, only acidic residues and leucine are found to be important for activation. This implies serine, proline and glutamines in ADs might fine-tune function, possibly through phosphorylation[67] or proline cis-trans isomerization[68] in ways that might elude detection in synthetic screens. In aggregate, DelRosso et al. find that nearly all ADs contain a hydrophobic character, combined with either acidic residues, disorder promoting S, P, Q, or a combination of the two.[49] This classification suggests that as in yeast, a modified *acid exposure model* suffices for most metazoan ADs, with added layers of PTM regulation and possibly using disorder to promote exposure to a larger extent.

3.7. Screens of RDs identify conserved structured domains and corepressor binding motifs

The general picture of disordered, ill-conserved and promiscuous ADs stands in some contrast to RDs that include the folded and well conserved KRAB and POZ/BTB domains.[5,69] Despite the common occurrence of these domains, a substantial portion of repressive sequences are predicted to be disordered.[49,54] Unlike ADs, several conserved cofactor interaction motifs are associated with RDs that are necessary for their function,[70,71] these are generally thought to be embedded in a disordered context and fold on binding.[71] RDs do not share the heavy acidic residue enrichment seen in ADs yet share their generally hydrophobic nature, some of which can be attributed to the folded domains.[5] In their activity screen of Pfam domains, Tycko et al. found far more RDs than ADs,[55] suggesting that RDs are far more commonly conserved well enough to allow for Pfam classification than ADs are and further evidence from screens suggests RDs are more structured.[49]

Alongside domain conservation, conserved corepressor binding motifs have been found in a large number of RDs,[2,49,54] with more than half of identified repressor tiles containing motifs. These motifs tend to contain conserved hydrophobic residues alongside a context of disorder with proline and charges that bind to corepressors. When mutated, these motifs were almost universally found to be important for function.[49,54] A common motif found to be important for repression was the SUMO interacting motif, albeit less so than the directly binding motifs. Finally, a strong association between repressive tiles and SUMOylation motifs was found with far more RDs containing SUMOylation motifs than any other single motif. The vital role of SUMOylation in transcription is excellently discussed in a dedicated review by Boulanger et al..[72] Despite the frequent conservation of motifs and domains found in screens of RDs, a large proportion of RDs function through unknown mechanisms, and the two most recent screens were unable to discover motifs or domains overlapping 45% and 28%, respectively, of screened tiles.[49,54] It is possible that RDs lacking identified motifs interact in much the same way as motif-bearing RDs where the constraints on the required interactions are flexible enough to allow large variations, as has been found for other IDR motifs.[52]

In summary, ADs and RDs both appear to have many commonalities, both possessing disordered, charged sequences that promote solvent exposure of hydrophobic residues which facilitate binding of cofactors.[49,54] However, how this binding interaction then progresses seems fundamentally different between the two, with large differences in conservation, sequence grammar and the relative importance of different PTMs. A recent work that quantified TFs and cofactors in human cells showed that TFs and corepressors are orders of magnitude more common than coactivators.[73] This discrepancy in the stoichiometry of cofactors may underlie the differences between ADs and RDs, since recruitment of a coactivator requires far higher affinity and competition than recruitment of corepressors. Conversely, recruitment of corepressors encounters a different problem, saturating amounts of similar corepressors may require a higher degree of specificity.

3.8. Limitations of high-throughput screens and AD predictors

It is possible that the screens of ED function fail to capture the full complexity of TF EDs by simplifying to one DBD and one or two promoter types. An earlier screen of *dmel* TFs found

different transcriptional effects depending on the type of enhancer the TF was targeted to, suggesting that some EDs require a proper cis-regulatory context to operate.[2] This is further corroborated by the finding that different enhancer and promoter types are dependent on specific cofactors.[74,75] A recent study by Jacobs and colleagues found that the presence of specific TF binding motifs determines the ability of a corepressor to repress transcription from an enhancer.[76] This suggests that the cis-regulatory effect is due to specific TFs interacting with corepressors to modulate their activity. Some TFs have been found to not depend on the type of cis element they bind and are classified as universal activators,[2] it may be that reductionist single DBD-promoter assays capture only these universally activating sequences and not more context sensitive EDs.

From several high-throughput AD screens, machine learning models for predicting ADs have been developed. Erijman et al. developed the convolutional deep neural network ADpred using their screen of random peptides in yeast as training data.[57] ADpred uses the amino acid sequence of the AD and its secondary structure to predict the probability of AD function for a given sliding sequence window. It predicts quite accurately mutational effects on activation measured in a study of Gcn4[77] and a bit worse on a deep mutational scan of Gcn4.[59] However, Sanborn et al. found that ADpred performed worse on their dataset of all yeast TF ADs.[45] Sanborn et al. trained a neural network, termed PADDLE, on their dataset and found it performed well on data withheld from training and on a set of human ADs.[49] Machine learning holds great promise for the identification of EDs from sequence; the success of AF2 in structural biology demonstrates its applicability to difficult biological problems.[29]

4. Protein-Protein Interactions through TF IDRS

As illustrated above in the case of transcriptional EDs, a very common function of TF IDRs is binding proteins to mediate their work in the genome. IDRs are particularly well suited to the role of PPI hubs as their structural variety enables them to bind to a wide array of structured and unstructured partners.[14,78] Other qualities of IDRs facilitate these interactions. For example, IDRs can contain small motifs that steer them into various pathways.[79] IDRs also tend to have very high association rates, potentially through a fly casting mechanism[80,81] afforded by their increased capture radius, which is conducive to rapid responses in signalling networks- a functional role in which IDRs are heavily overrepresented.[14,24] The dynamic ensembles of structures that IDRs adopt affords most residues of the chain access to the solvent and thus also binding partners, allowing for a very large number of sequence contexts for binding and rendering the chain accessible for PTMs.[14,82] The dynamic nature of IDRs thus lends itself easily to roles in processes that require rapid, flexible, and integrated processing of information in the cell- they are the organic optical fibres and switches of the molecular internet.

4.1. Protein interaction motifs within TF IDRs

Short linear motifs (SLiMs) are sequences of 3–12 amino acids in IDRs that mediate binding to other proteins.[83,84] Commonly these motifs bind to surfaces of globular proteins, and may fold upon binding and/or maintain a "fuzzy binding" interface to their partner.[23,85] SLiMs can also be found in multiple copies in a single chain as in the case of Gcn4 binding to Med15.[47,83] It has been estimated that over 100 000 SLiMs exist in the human proteome, and they are especially abundant in signalling and transcription.[86]

The best studied and perhaps the most important interactions in TFs mediated by SLiMs are interactions with globular domains of cofactors. Especially formative in this field has been work on SLiM binding to the paralogous coactivators CBP and p300.[27,88,89] Another recent and illustrative example of these interactions are found in αα-hubs, identified as common folded hubs in transcriptional cofactors that interact with disordered ADs of TFs.[90,91] There are many examples of such domains in transcriptional cofactors, a particularly well studied one is the RST domain of RCD1.[92] RCD1 is a plant-specific cofactor that modulates abiotic stress by binding to various TFs through its αα-hub domain RST interactions with the TF DREB2A.[93,94] The main distinctive feature of αα-hubs is an αα-hairpin motif with one helix preceding and another following the hairpin motif. In the RCD1-RST domain the helices together form an L-shaped glove with a hydrophobic palm made up of interior surface parts of each helix (Figure 5).[90,92] When the RCD1 interacting SLiM of DREB2A binds to its RST domain, it does so with nanomolar affinity and adopts a partially α-helical structure on binding with a bend between the palm and fingers of the L-

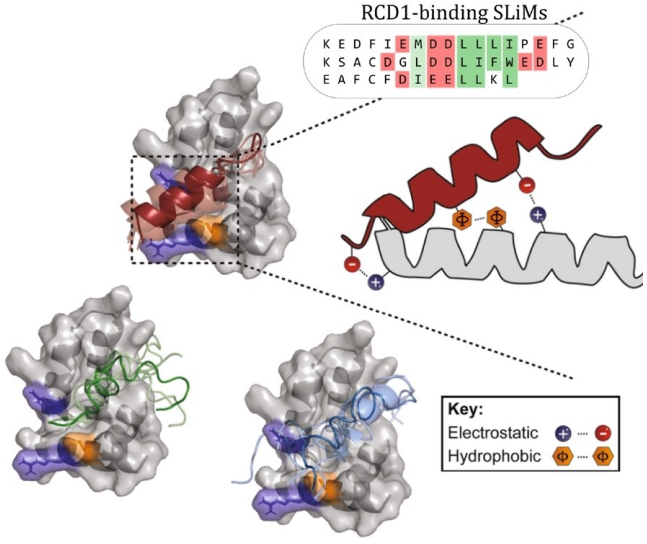


Figure 5. Simplified models of RCD1:TF interactions and representative SLiMs associated with these models. Grey surface representation shows the RCD1-RST domain with basic residues close to the binding surface highlighted in blue and hydrophobic in orange. Binding peptides are show as overlayed coloured cartoons with DREB2-like peptides in red, NAC013 in blue and NAC046 in green. Modified from Ref. [87] under Creative Commons 4.0 licence (https://creativecommons.org/licenses/by/4.0/).

shaped glove (Figure 5, red peptide).[91] This interaction is primarily dependent on just four hydrophobic residues in DREB2A and three to four aspartates at the ends of the SLiM that interact with basic residues on the rim of the hydrophobic palm.[91]

An investigation of the RCD1-RST domain showed that it tends to bind to disordered SLiMs that contain leucines, aromatic and acidic residues,[84] such as the DREB2 A SLiM. The contribution of helicity to binding was dependent on the particular SLiM, with one loosing binding on proline substitution and another unaffected.[84] These results align very well with high-throughput screens of ADs, and especially noteworthy is the observation that helicity requirement is not dependent on the coactivator but rather the SLiM as suggested by mutational studies of ADs.[45] It is perfectly plausible that some coactivators prefer a helical binding partner but this case along with studies of Med15 illustrate that helical conformations in ADs may be a simple way of exposing a hydrophobic binding surface to a binding partner and not necessarily a requirement of shape-complementarity.[45,47,60] When comparing the sequence features of RCD1 binders, a loosely defined putative binding motif was identified. The authors identified the motif in a dozen other interactors of RCD1, though a fair number did not bind in their assay.[84] Despite good predictive performance of the motif, it was not compared to the predictive power of sequence composition alone.[84] Additionally, a subsequent study found that the C-terminal domain of RCD1 acts as a ligand mimic, with increased hydrophobic and acidic content correlated with increased interaction in the absence of the putative motif.[95] This suggests that exact motif grammar is not of vital importance for most coactivator binding, as seen in high-throughput studies of TF EDs.[45,49,56,57]

4.2. SLiM specificity

Despite the flexible RCD1 binding SLiM, experiments on αα-hub binding show specific binding to RCD1 for DREB2 A over the structurally similar TAF4 αα-hub within the same species but much lower specificity when binding is examined for the orthologous coactivator from human (50x vs. 7x higher affinity, respectively), indicating that the SLiM has evolved to specifically bind RCD1 in its native context.[96] The authors suggest that small variations in the αα-hubs architecture result in different binding locations and modes of interaction for the partners, enabling some specificity despite the similar coactivator fold and residues.[92,96] These interactions are also distinguished by the contributions of enthalpy and entropy to binding with the specific interaction being more enthalpically favourable and less entropically favourable. Such differences in mode may increase the binding specificity differences in vivo where the crowded cellular milieu reduces the entropic space available and curtails the entropic benefit the non-specific binder has in vitro.[97]

4.3. IDRs involved in negative feedback of transcriptional regulation

Many TFs form homo- or heterodimeric complexes with other TFs[98] and IDRs frequently play an important role in their formation. IDR modulation of dimerization can be used to finely tune transcriptional feedback mechanisms. One such case involves PU.1, a master regulator critical in maintaining a pool of hematopoietic stem cells. PU.1 contains an N-terminal AD, a disordered anionic PEST domain, and a structured ETS DBD. Recently, Xhani et al. showed that activation of gene expression by PU.1 is regulated by two distinct dimeric states: a transcriptionally active 1:1 complex and an inactive 2:1 ternary complex involving two PU.1 molecules bound to a single DNA recognition site.[99] By monitoring the dose-dependence of PU.1 transcriptional activity in HEK293 cells, the authors showed a bell-shaped activity response to PU.1 concentrations, suggesting a negative feedback mechanism. NMR spectroscopy and tryptophan fluorescence experiments revealed that the intrinsically disordered PEST domain reduced the binding affinity of the second PU.1 molecule to form a ternary complex but also promoted homodimerization in the absence of DNA. The two dimeric forms were found to be non-equivalent, with an asymmetric DNA-bound PU.1 dimer and a symmetric homodimer in the DNA-free state. The negative feedback was reduced with phosphomimetic substitutions in the PEST domain which promoted the formation of a transcriptionally active 1:1 complex with DNA.

Another example of an IDR involvement in negative feedback control comes from a study on the TF HIF-1.[100] Hypoxic stress response in cells is mediated by HIF-1 and through the interaction of its IDR with the TAZ1 domains of CBP and p300. HIF-1 regulates the transcription of the gene CITED2, which negatively regulates its own expression by competing with HIF-1 for TAZ1 binding. Berlow and colleagues showed by NMR spectroscopy that CITED2 forms a transient ternary complex with HIF-1 and TAZ1, and subsequently competes using a common conserved motif for a shared binding site that enhances the rate of HIF-1 dissociation from the complex. This molecular threesome exemplifies how disordered and dynamic TFs can enable rapid responses in signalling, in this case terminating the hypoxic response.

4.4. Predicting functions and PPIs of IDRs

The large variety of sequences that a single IDR can bind along with the important role of context in attenuating these interactions motivates attempts to predict PPIs of IDRs in silico from curated existing data, along with physical and evolutionary principles. These predictive methods can further help in identifying possible interaction patterns that high-throughput screens may not be able to discriminate or identify. Amino acid sequence motif searches using online computational tools such as the ELM resource are powerful tools to quickly identify possible functions, though care must be taken if the context of the motif allows for the predicted function as not all matching

sequences are biologically active.[52,101] Motif conservation has also been highlighted as a predictor of active function and evolutionary analysis of IDRs in general are powerful tools to determine their conserved function.[102]

Evolutionary analysis has been a great method for assigning function to conserved folded domains. SLiMs are a promising start in this endeavour as they have been identified as "islands of evolutionary conservation in rapidly evolving IDRs".[103] Despite being better preserved than the surrounding disorder, SLiMs evolve rapidly and are lost and added far more frequently than conserved folded domains.[103] Rapid evolutionary dynamics and the observation that conserved segments of IDRs are a minority (~5%) of all IDR residues[102] significantly restrain reliable sequence to function analysis for IDRs.[11]

It has been found that even though the amino acid sequences of SLiMs and IDRs in general are not conserved, the chemical features and patterns that enable molecular interactions, such as overall amino acid composition and net charge, are constrained and conserved.[40,104–106] Several studies by the group of Alan Moses have used simulated molecular evolution to determine what chemical constraints apply to IDRs and how these constraints are related to function. Interestingly, they have found cases where diverged, yet similarly chemically constrained IDRs are functionally equivalent.[40,104] They then reasoned that if shared chemical constraints enable functional equivalence, function can be predicted by identification of these constraints.[40,107] To enable such predictions Zarin et al. developed the statistical model FAIDR that uses vector descriptions of chemical constraints to predict the IDR of a protein responsible for a given molecular function and the chemical constraints of the IDR that enable individual functions.[107] Using this approach, they identified conserved chemical features in the N-terminal IDR of the yeast protein Mfg1 that indicated it is a TF. Despite Mfg1 not being annotated as TF, the authors found strong existing experimental evidence that it is indeed a TF.[107] Such successes highlight the utility of evolutionary analysis in functional annotation of the diverse roles of TF IDRs. These roles extend well beyond PPIs and transcriptional regulation as we detail below.

5. IDRS Contribute to DNA Searching, Binding, and Specificity

Traditionally, the DBD has been considered responsible for ensuring specific binding of TFs to their cognate DNA sequence. Commonly, the recognition sequence is only a few base pairs in length (6-12 bases) and can often be found in areas bound by the factors in vivo. From these and other observations a general model has been formulated in which structured DBDs bind to defined regulatory elements to effect transcription though their IDRs. DBDs and IDRs are thus thought to have separate roles in modular fashion. Several studies largely support this view with hundreds of examples of synthetic ADs attached to native DBDs with no seemingly ill effect on DNA binding or transcriptional activation. However, the old view of IDRs functioning exclusively as recruiting hubs for the transcriptional machinery is rapidly being replaced by one far more promiscuous (Figure 6). Instead of being solely involved in PPIs, examples are now emerging of how IDRs modulate interactions or directly interact themselves with nucleic acids and chromatin units.[108,109] Albeit not a TF, the linker histone H1 provides a clear example of a bipartite protein binding to nucleosomes through both its disordered positively charged C-terminal domain and the globular DBD.[110,111] Remarkably, the C-terminal domain, which drives linker DNA collapse, remains fully disordered in the complex.[112]

5.1. IDR involvement in the search for DNA binding sites

For many TFs, IDRs contribute to the search for short recognition motifs within the vast human genome. Current models of search modes include 1-dimensional sliding and 2-

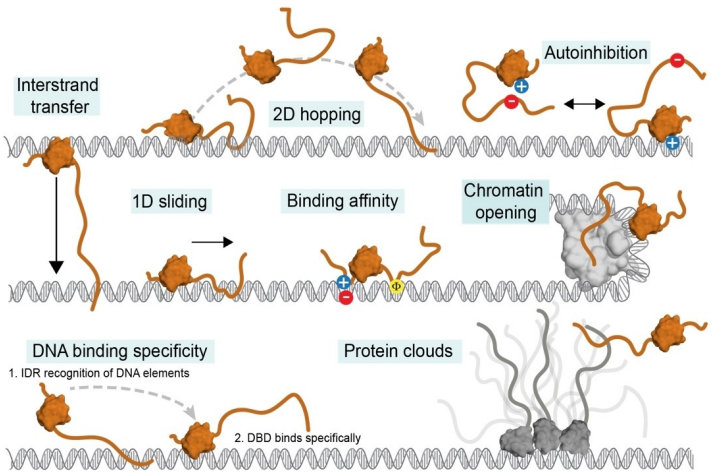


Figure 6. Schematic illustration of diverse TF IDR functions linked to DNA search, binding, and specificity (see text for details).

dimensional hopping on DNA (Figure 6), both of which have been suggested to enhance the search speed by 100-fold compared to a 3-dimensional search.[113,114] DNA and RNA binding proteins are furthermore enriched in positive charges (Figure 2B) and positive charges in IDRs can increase affinity to the negatively charged DNA backbone. Simulations using the IDRs of homeodomains have suggested this to lead to more sliding rather than 3D diffusion and hopping[115] but that comes at the cost of slower searching. IDRs can also facilitate interstrand exchange (Figure 6), whereby the large capture radius of the IDR allows it to bind another proximal DNA strand, pulling or directing the DBD towards it for a local search of DNA sequence.[115] The interstrand transfer, which has been termed a "monkey-bar" mechanism, was highly dependent on separation between segments, that is, there was an optimal position of a positively charged IDR relative to the DBD for promoting interstrand exchange. Disordered linkers between several DBDs can have a similar role; they increase DNA binding affinity over a single domain and allow for a monkey-bar mechanism whereby the strongest binding DBD serves as an anchor enabling the weaker binder to search.[115]

5.2. IDRs can affect binding affinity

DNA binding affinity is enhanced by short, often basic disordered regions flanking the DBDs,[116] with examples including the pluripotency factor Sox2,[117] the Hox proteins[118] and LEF-1.[119] The flanking regions can interact with the major groove, most often through specific interactions, or through the minor groove generally non-specifically through charged interactions since the minor groove contains mostly the phosphate backbone and offers little sequence specificity. In some cases, the DBDs are already defined by including these flanking IDRs but in other cases not. What is important to emphasize is that the definition of a DBD is not necessarily based on structured modules, as is commonly implied, but can also include disordered flanking segments of considerable length.

Interactions between different functional modules of a TF have now been observed for many proteins.[120,121] It is therefore no longer useful to define TFs as simple disconnected modular units as the entire polypeptide sequence can be synergistically linked across order and disorder. Interactions between domains can in principle affect a multitude of properties and parameters, for example, the dimensions of both DBD and IDR, binding affinity to DNA through association and/or dissociation rates, protein interactions or dimerization, degradation and turnover time, localization, and phase separating propensities. Simulations of synthetic and natural proteins have suggested that dimensions of IDRs are sensitively modulated by neighbouring folded domains,[122] which in turn would control the accessibility of the IDR, for example, for cofactors. However, pinpointing interdomain interactions and how they link to function remains challenging.

Generally speaking and perhaps intuitively, the presence of IDRs reduces binding affinity to DNA by interacting with the DBDs in absence of DNA[118,123] (Figure 6). However, the effects on DNA binding affinity can also be more complex and bring important implications for specificity. A large body of work comes from studies on the tumour suppressor TF p53, which illustrates how complex and cooperative IDR interactions can be. The functional form of p53 is tetrameric with a molecular weight of ~180 kDa which renders the protein a formidable challenge for structural studies. Kroise et al. used intein splicing to investigate the roles of the p53 IDRs with atomic resolution using NMR spectroscopy.[121] This approach allowed segmental labelling of discrete regions, which greatly facilitated resonance assignments and led to resolvable signals for almost every residue in the N-terminal transactivation domain (NTAD) within the full-size tetramer. Interestingly, the chemical shifts of the NTAD showed a different signature in the full-size protein compared to in an isolated NTAD construct, which were convincingly demonstrated to be due to direct, mainly electrostatic interactions of the NTAD with the DBD. DNA binding experiments showed that the interactions of the NTAD with the DBD regulated DNA binding by reducing the binding affinity to non-specific DNA 5-fold without compromising affinity to specific DNA. He et al. used proteolytic fragment release assays to first confirm that the N-terminal IDR interacts directly with the DBD.[10] By using point mutations to abolish the native tetramerization and produce either dimers or monomers, the authors further showed that the disordered region interacted primarily intermolecularly with a DBD from a different monomer and accelerated disassociation from DNA. Given the abundance of TFs with similar charge profiles in their DBDs and EDs, such transient electrostatic interactions are likely to be quite common but predicting the functional consequences may prove difficult and case specific.

Work from Laptenko and co-workers has shown that the C-terminal IDR of p53 (CTD) dramatically affects DNA binding.[124,125] In this case, footprinting assays suggested the CTD to increase affinity to DNA. Using ChIP-Seq and SELEX (Systematic evolution of ligands by exponential enrichment), the authors showed that alterations to the CTD reduced affinity to the bound sites and reduced binding to alternative motifs. Biochemical assays then led to the conclusion that the CTD stabilizes the cooperative interaction of the DBDs on non-cognate DNA by direct contacts. Based on the results, the authors suggested a model in which the CTD diversifies the sequences that the p53 tetramer can bind and enhances cooperativity. A similar observation was reported for the NFκB family TF p50/RelA heterodimer where the IDR of RelA was shown to increase DNA binding affinity at the cost of reducing specificity.[126] Clearly, the nature of interdomain interactions in TFs can be complex and far from trivial to predict. To complicate matters even further, the IDRs are often subject to chemical modifications in vivo that affect the conformational properties of IDRs.

5.3. Chemical modifications tune IDR interactions

Changes to TFs in the form of PTMs can extend their chemical properties beyond those defined by the 20 common amino

acids and thereby modulate protein functions. Recent work[127] expanded on the p53 study described above by monitoring the effects of phosphorylation on DNA binding. The study revealed how a DNA-damage induced dephosphorylation of Thr55 in the NTAD reduced DNA binding inhibition, allowing p53 to bind to lower-affinity elements related to the DNA damage response. Renewed phosphorylation is also linked to switching off the response and is predicted to release p53 from lower-affinity elements. The Thr55 phosphorylation effect is to remove DNA binding cooperativity and force p53 to bind initially as a dimer. As p53 binds an additional dimer, Thr55 competition for the DBD binding surface causes total detachment of the tetramer and inhibition of binding. Thr55 phosphorylation thus acts as a powerful tuner of p53 DNA binding.

Phosphorylations need not affect DNA binding affinity as seen for the TF B-MYB, an oncoprotein involved in cell cycle regulation as well as apoptosis and cancer.[128] A recent report showed that the C-terminal IDR makes direct interactions with the DBD in a phosphorylation-dependent manner.[129] Phosphorylation of a single amino acid, Ser577, was enough to dismantle the interaction yet deletion mutants showed that DNA binding was unaffected by the interaction with the IDR and is instead likely to involve interactions to an acidic patch on the side of the DBD opposite the DNA-binding surface. Thus, intramolecular interactions between domains in TFs need not affect DNA binding but may have downstream regulatory roles for the EDs.

5.4. IDRs have unexpected effects on DNA binding specificity

How TFs achieve specificity, both with regards to DNA sequence and subsequent recruitment of coactivators, is a fundamental and ill-understood problem. As mentioned before, TFs only occupy a subset of their putative motifs in vivo. Do IDRs contribute to the selection of binding-motif sites that are occupied by TFs? Brodsky et al. addressed this question and investigated the role of long IDRs in DNA binding specificity, on a genomic scale.[130] They used two budding yeast TFs; Msn2 and Yap1,both of which contain over 500 amino acid long IDRs. The binding profiles of both TFs were investigated by chromatin endonuclease cleavage followed by sequencing. Remarkably, the results revealed that the IDRs were required and even sufficient for the TFs to find their target promoter. The isolated DBDs (i.e., lacking the IDRs) localized to the same preferred motif type as the full-length TFs, but to a different subset in the genome, targeting previously unbound sites. The isolated IDRs did not occupy the same preferred motifs as the DBDs but were able to recognize most of the same promoters as the full-length TFs. This surprising specificity was found to be directed through weak multivalent interactions distributed throughout the polypeptide sequence and implied to be directly to the DNA. This type of binding site sensing through IDRs was suggested to enhance the search rate by rapidly confining TFs to specific regions containing the cognate binding site.

What if instead of interacting with the DNA, TFs directly interact with proteins already present at the promoter sites? Max Staller devised this alternative hypothesis based on data from Brodsky et al. and others[130–132] and suggested a 2-step search mechanism for some TFs[133]. The model involves an initial global search of a TF for what Staller refers to as 'protein clouds' (Figure 6), through PPIs with other TF IDRs already localized to the region, followed by a local search for their cognate DNA binding sites using their DBDs. The protein cloud is a loosely defined phenomenon but could involve homo- and heterotypic interactions between tens of TFs and/or cofactors that are then recognized by the searching TF. Staller's model provides possible solutions to some confounding results including why large regions in TFs have unknown functions, why most TFs bind only a fraction of their potential binding sites, and why many TFs seem to occupy regions that do not contain their cognate binding sites. The 2-step search mechanism implies that: i) regions with unknown functions are responsible for global searching through PPIs, and ii) searching and binding the right protein cloud would contribute to the selection of binding sites, explaining why TFs are sometimes found where there is no cognate binding sequence.

Stallers 2-step search mechanism involving protein clouds also highlights another issue with specificity; how do TFs select coactivators to bind through PPIs of their IDRs. Accumulating data suggest that these interactions are primarily non-specific and that binding modes between TFs and coactivators are inherently fuzzy. Henley et al. posit that non-specific interactions are inconsistent with the functional role played by specific activator/coactivator complexes and point out that there are many examples where transcriptional activation depends on specific interactions between TF ADs and activator binding domains of coactivators.[134] The authors studied the interaction between ADs of three members of the ETV/PEA3 TF family and the Mediator subunit Med25, which have previously been suggested to be mainly non-specifically mediated. The ADs of ETV were very sensitive to subtle changes in sequence primarily due to effects on specific sets of interactions despite being generally dynamic and fuzzy interactions. Interestingly, plasticity in the Med25 interaction site contributed to the conformational sensitivity as well. Overall, the results suggest that even in dynamic complexes that seem entirely non-specific, there may be hidden important specific interactions. The authors note that if indeed there is more specificity involved in activator/coactivator communication, the commonly referred to "undruggability" of TFs may be overcome, as pointed out by others.[135,136]

5.5. IDRs in nucleosome-binding TFs can contribute to chromatin opening

Classically, TFs require their binding sites in DNA to be accessible to enable binding, but most of genomic DNA is wrapped around core histone octamers and thus inaccessible. A systematic SELEX screen of 220 TF DBDs revealed that many can bind DNA which is condensed in nucleosomes.[137] A further screen of 593 full length TFs identified strong nucleosome binders, including a subset that is involved in cell reprogramming mechanisms.[138] This particular subset of TFs, termed

pioneer TFs, are important in development and play a crucial role in cell decisions[139]. Their function in cell reprogramming may be due to their abilities to target and open condensed chromatin, either directly or through recruitment of chromatin remodellers.[140] This ability makes pioneer TFs essential for the cell differentiation and activation of otherwise transcriptionally silent genes.

FoxA TFs were among the first to be suggested to have pioneer activity.[141] FoxA TFs bind to their DNA motifs through a winged-helix DBD that resembles the globular domain of linker histone H1,[142,143] and also bind to the core histones H3 and H4 via their C-terminal domain.[142] Binding of FoxA TFs induces local nucleosome decompaction of H1 bound nucleosomal arrays,[142] it has also been shown that FoxA TFs compete with and displace linker histone H1 from the nucleosome, thus creating local chromatin decompaction that enables DNA binding by non-pioneer TFs.[139] Recent work by Zaret and colleagues shows strong evidence of FoxA1 IDR involvement in chromatin opening. A combination of sequence analysis, cross-linking and mass spectrometry revealed a conserved 9 amino acid region within the long C-terminal IDR of FoxA1 that had helical propensity and contributed to chromatin opening through an interaction with the core histones of the nucleosome. Measurements of DNase cleavage activity showed that deletion of this region led to severely reduced chromatin opening by FoxA1. Furthermore, the same deletion in mouse embryos led to a reduction in target gene chromatin accessibility which severely impaired embryonic development.[144]

A similar contribution to chromatin opening was observed for the pioneer TF PU.1. As discussed above, a negatively charged IDR in PU.1, the PEST domain, modulates homodimerization on and off DNA, thereby regulating transcriptional activity. Interestingly, hypersensitivity assays on chromatin fibres in vitro showed that a mutant with a deleted PEST IDR performed considerably worse in rendering the chromatin fibre accessible to nucleases.[145]

The Yamanaka factors Sox2, Oct4, Klf4, and c-Myc play a key role in the in vitro induction of pluripotent stem cells.[146] The key factor Sox2 consists of two IDRs flanking an HMG-box DBD. A part of the IDR flanking the C-terminal side of DBD has been recently shown to bind RNA, even when the DBD is bound to DNA.[108,109] Deletion of this RNA binding region severely reduced pluripotency induction efficiency, indicating a connection between IDRs and cell reprogramming. The Sox2 HMG-box cooperates with the Oct4 POU-domain and this interaction is crucial for pluripotent stem cell induction and cell pluripotency.[147] Recent studies also revealed Oct4-Sox2 nucleosome ternary complex by cryo-EM,[148] and structural changes in the core nucleosome structure ranging from local DNA distortion to full DNA removal from one side, depending on the cognate binding site location upon cooperative Oct4-Sox2 binding. However, the exact mechanism that leads to chromatin opening is still unclear.

6. Phase Separation of TFs is Modulated by IDRs

6.1. The chemistry of phase separation

Liquid-liquid phase separation (LLPS) is a well-known phenomenon in polymer physics and chemistry which has been extensively studied since its first introduction by Paul Flory in 1942.[149] However, LLPS remained unconsidered for biomacromolecules despite successful implementation of polymer physics concepts to describe biomolecular structure and behaviour, for example, Markov state models for protein misfolding,[150] Worm-like chain and Freely-jointed chain models for DNA and RNA.[151,152] Some proteins, like haemoglobin were reported to undergo phase separation in vitro[153] and LLPS is quite commonly encountered during protein crystallization trials, but the significance of these observations remained unclear.[154,155]

Phase separation depends on a variety of factors and can be observed both in vivo and in vitro.[156–158] In vitro, the formation of condensates is generally promoted by molecular crowding due to either high protein concentrations or addition of crowding agents like PEG[156] and lower ionic strength solutions.[157] In vivo, LLPS are usually thought to be mediated by phosphorylation or poly(ADP-ribosylation) and multivalent interactions between and within nucleic acids and proteins, often involving interactions between hydrophobic and charged residues.[159] A study by Rosen and colleagues on phase transitions of multivalent signalling proteins suggested that phase separation might be crucial to the formation of membrane-less organelles and to the regulation of certain biological functions like transcriptional activation and gene expression.[160] Recent work has suggested that multivalency is a defining feature that drives phase transition of proteins to form membrane-less organelles.[161] IDRs are a typical example of multivalent polypeptides and their ability to form many contacts simultaneously might be important for LLPS as well.[162] But the question of how exactly IDRs drive phase separation is still not fully understood. The amino acid sequences of many IDRs are often enriched in Gly, Ala, Ser, Pro, Leu and Glu residues (as shown previously (Figure 4B)), however other residues are shown to contribute to LLPS as well. These residues do not seem to be randomly distributed but rather appear among SLiMs, as discussed above. There is considerable evidence that interactions that drive LLPS include electrostatic, dipole-dipole, π-π, cation-π, hydrophobic, and hydrogen bonding interactions (Figure 7).[163–166] The π-π interactions occur between sp2-hybridized groups that are most abundant in aromatic amino acids such as Phe, Trp and Tyr. Electron interaction of their aromatic groups can adopt off-centre parallel or edge-to-face configuration, although polarization of aromatic rings may lead to face-centred stacking.[167,168] Additionally, weak multivalent interaction throughout the entire protein sequence may arise from partial π-bonds in the peptide backbone. Cation-π interactions are usually observed between

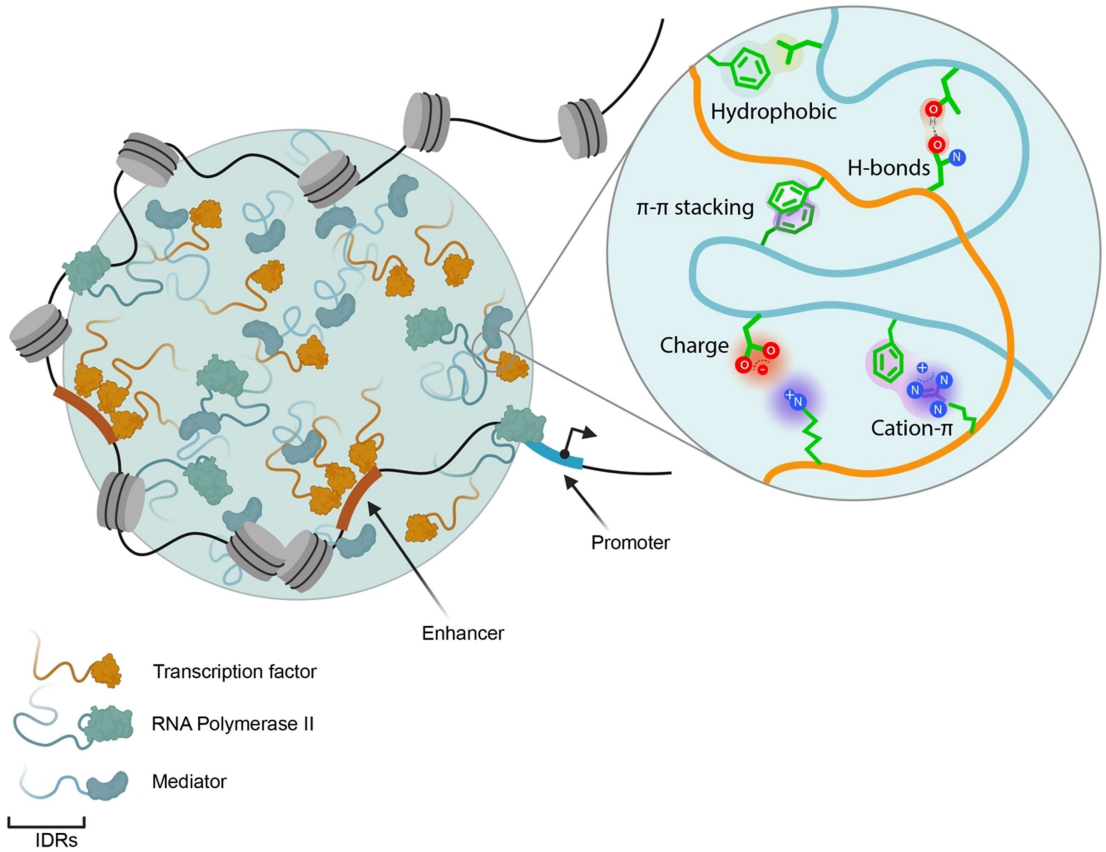


Figure 7. TF-Mediator coactivator phase separation. Model of a droplet condensate formed by DNA, TFs and coactivators. This model includes multivalent interactions between both disordered and structured regions, involving many types of dynamic non-covalent bonds. Based on figure from [26].

Arg and Tyr residues but can also occur between other residues as shown previously .[169,170]

The π -π and cation-π interactions are very prominent in the context of IDRs and unfolded nucleic acids which have their aromatic groups exposed and multiple studies have shown that single-stranded but not double-stranded DNA is incorporated into IDR-rich condensates.[171,172] Hydrophobic contacts and hydrogen bonds might be less predominant in IDRs compared to folded domains, but they might be important for maintaining protein chain disorder and keeping assemblies liquid-like rather than solid.[173,174]

6.2. Phase separation of TFs

Being rich in disordered regions, many TFs have been reported to undergo LLPS.[175,176] The pioneer TF Oct4 was shown to phase separate in vitro with Mediator which promoted a reorganization of topological-associated domains and cell fate transitions.[157,177] Mediator has been reported previously to form condensates in embryotic stem cells[178] and is thought to interact with Oct4 via Med1 subunit that also contains IDR.[179] In a recent work, Boija and colleagues used ChiP-seq and fluorescence in situ hybridization to show that ADs of Oct4 and GCN4 posses an ability to undergo phase separation in the presence of Mediator, and might be important for transcriptional activation.[26] The authors established that specific acidic residues within IDR-rich AD of Oct4 are responsible for the MED1-IDR phase separated droplet formation, and then proceeded to demonstrate phase separation of 7 more TFs (MYC, p53, NANOG, Sox2, RARa, GATA2,ER) through MED1-IDR interaction. IDR-mediated phase separation was proposed to be a general mechanism by which TF ADs affect gene expression. This attractive model also offered an explanation to observations that are difficult to reconcile with a classical lock-and-key model of PPIs, including how hundreds of TFs can interact with relatively few coactivators.[73]

Besides Oct4, other pioneer TFs have been observed to undergo phase separation, suggesting a potential link between LLPS and the key role of pioneer TFs in cell identity. Morin and colleagues used a dual magnetic trap in combination with fluorescence and controlled microfluidics to show that pioneer TFs such as Klf4 can undergo surface condensation on DNA. This transition is consistent with physical theory behind surface condensation, known as prewetting, and can provide a mechanism for sequence-dependent formation of condensates limited in size by interaction with DNA surface.[180] It has also been suggested that IDRs in Klf4 might not be necessary for phase separation in vitro, but rather might be important for LLPS in vivo by recruiting other TFs and promoting droplet

formation.[181] Yet another pioneer TF, Sox2, is known to undergo LLPS via multivalent interactions at low salt concentrations and has recently been shown to phase separate at non-physiological salt concentrations, suggested to be favoured mainly by hydrophobic and non-ionic interactions.[182]

An important issue is whether there exists a direct link between transcriptional activity and propensity for LLPS in TFs. Trojanowski and colleagues addressed this issue and showed using synthetic TFs that that multivalent interactions of IDRs (harbouring ADs) enhanced transcription but independent of droplet formation. They also observed that TF transcriptional activity was dependent on residence time in the promotor-bound state rather than binding site occupancy, suggesting that residence time can be rate limiting for transcription activation of endogenous TF.[183] The influence of LLPS on transcriptional activity remains to be fully established and may be dependent on the TF system in question.

7. Summary and Outlook

Transcription is an exquisitely complex process that involves many molecular components organized temporally and spatially. At the core of transcriptional regulation are TFs; dynamic and structurally heterogeneous proteins that synchronize the entire transcriptional machinery and are notoriously challenging to study. We have highlighted the prevalence of IDRs in TFs and recent efforts in understanding how the sequences of TF IDRs translate to structural and functional outputs. IDRs mediate diverse functions beyond recruiting coactivators, including contributing to DNA search, binding, and specificity through various mechanisms.

However, we still lack a detailed quantitative, physical description of TF IDRs and their link to transcriptional activity. It remains a major challenge to study the conformational dynamics of full-length TFs, as is evident from the complete lack of experimentally derived structures of entire TFs in the protein data bank. The modular approach to studying TFs is no longer useful since examples of molecular communication between DBDs and IDR that contribute to their functions are frequently emerging. A static view of folded domains as building blocks pieced together through shape complementarity needs to be replaced by a dynamic structural ensemble view for IDRs, with a heavier focus on statistics and distributions of conformational states. The modern view is also relevant to gene transcription in general, which is an inherently stochastic and dynamic process.

Another fundamental problem is to connect in vitro observations to in vivo behaviour. Most research on the physical properties of TFs rely on simplified in vitro experiments. To approach a realistic physical description of IDR ensembles we need to redefine research questions to fit within dynamic structural biology. For that purpose, integrative modelling using a combination of experiments and simulations has proven particularly powerful.[184] Integrating experiments and simulations needs to incorporate a solid conceptual framework driven by theory and high-throughput studies, to bridge the knowledge-gap between in vitro and in vivo research.

Genome-wide high-throughput studies are crucial to decipher the sequence alphabet of EDs and many ground-breaking studies have emerged in the last few years. It may be that studies using transcriptionally derived readout enrich for universal coactivators and fail to capture more specific binders or that alongside specific binding, some affinity for universal coactivators like Mediator is required.[2,45] However, for highly effective models to be trained and validated, a common community definition of an AD, harmonised AD activity data reporting and validation alongside blind tests of the algorithms is needed. Common reporting standards will allow the creation of large benchmark datasets for training followed by blind tests to assure generality, such as those now in place for structure and disorder prediction.[37,185]

The correlation of features important for phase separation and transcriptional activation raises the possibility that these processes are intimately linked. Many of the same features have also been strongly linked to direct co-activator binding. Distinguishing the contributions of these two mechanisms, direct recruitment and phase separation, is a difficult challenge that requires deconvoluting interlinked properties, such as solubility and phase separation. It is very probable that coactivator binding and LLPS are not distinct phenomena but one and the same, with LLPS arising from the multivalent interactions of TF EDs and coactivator binding domains. LLPS is then an emergent phenomenon when the concentrations exceed saturation.

The physical properties of TFs and the molecular phenomena they direct are ultimately intricately connected and must therefore be studied synergistically to obtain a full mechanistic map of transcription in eukaryotes. Machine learning and artificial intelligence will play an increasingly important role in integrating the diverse datasets, from low-throughput in vitro studies to high-throughput assays on living systems, promising an exciting future ahead in the fascinating field of transcription factor research.

Acknowledgements

We apologize sincerely to the authors that were not cited due to space restrictions. We acknowledge all members of the Heidarsson research group for stimulating discussions over the years. We thank Sarah F. Ruidiaz, Jordan McIvor, and Davide Mercadante for assistance with artwork. We thank Karen Skriver, Max V. Staller, Nicole DelRosso and Taraneh Zarin for comments on the manuscript. This work was supported by funding from the European Research Council (ERC StG, 101040601-PIONEER), the Icelandic Research Fund (grant nr. 217392051), and the Icelandic Cancer Society.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Keywords: biophysics · DNA recognition · protein-protein interactions · protein structures · structural biology

[1] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, M. T. Weirauch, *Cell* **2018**, *172*, 650–665.
[2] G. Stampfel, T. Kazmar, O. Frank, S. Wienerroither, F. Reiter, A. Stark, *Nature* **2015**, *528*, 147–151.
[3] F. Reiter, S. Wienerroither, A. Stark, *Curr. Opin. Genet. Dev.* **2017**, *43*, 73–81.
[4] T. I. Lee, R. A. Young, *Cell* **2013**, *152*, 1237–1251.
[5] L. F. Soto, Z. Li, C. S. Santoso, A. Berenson, I. Ho, V. X. Shen, S. Yuan, J. I. Fuxman Bass, *Mol. Cell* **2022**, *82*, 514–526.
[6] P. B. Sigler, *Nature* **1988**, *333*, 210–212.
[7] J. J. Ferrie, J. P. Karr, R. Tjian, X. Darzacq, *Mol. Cell* **2022**, S1097276522009108.
[8] S. L. Shammas, *Curr. Opin. Struct. Biol.* **2017**, *42*, 155–161.
[9] L. Staby, C. O'Shea, M. Willemoës, F. Theisen, B. B. Kragelund, K. Skriver, *Biochem. J.* **2017**, *474*, 2509–2532.
[10] X. Guo, M. L. Bulyk, A. J. Hartemink, in *Biocomput. 2012*, WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA, **2011**, 104–115.
[11] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, *Chem. Rev.* **2014**, *114*, 6589–6631.
[12] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, *Chem. Rev.* **2014**, *114*, 6561–6588.
[13] S. E. Bondos, A. K. Dunker, V. N. Uversky, *Cell Commun. Signaling* **2021**, *19 (88)*, s12964-021-00774-3.
[14] P. E. Wright, H. J. Dyson, *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29.
[15] F.-X. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P. B. Crowley, L. Gierasch, G. J. Pielak, A. H. Elcock, A. Gershenson, P. Selenko, *Chem. Rev.* **2014**, *114*, 6661–6714.
[16] B. Tsang, I. Pritišanac, S. W. Scherer, A. M. Moses, J. D. Forman-Kay, *Cell* **2020**, *183*, 1742–1756.
[17] J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, A. K. Dunker, *Biochemistry* **2006**, *45*, 6873–6888.
[18] G.-N. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, C. C. Gradinaru, *J. Am. Chem. Soc.* **2020**, *142*, 15697–15710.
[19] *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* (Eds.: I. C. Felli, R. Pierattelli), Springer International Publishing, Cham, **2015**.
[20] S. Gosavi, B. Schuler, *Curr. Opin. Struct. Biol.* **2020**, *60*, iii-iv.
[21] S. Brodsky, T. Jana, N. Barkai, *Curr. Opin. Struct. Biol.* **2021**, *71*, 110–115.
[22] R. Vancraenenbroeck, Y. S. Harel, W. Zheng, H. Hofmann, *Proc. Nat. Acad. Sci.* **2019**, *116*, 19506–19512.
[23] P. E. Wright, H. J. Dyson, *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.
[24] G. Vinterhalter, J. J. Kovačević, V. N. Uversky, G. M. Pavlović-Lažetić, *Int. J. Biol. Macromol.* **2021**, *167*, 446–456.
[25] C. J. Oldfield, A. K. Dunker, *Annu. Rev. Biochem.* **2014**, *83*, 553–584.
[26] A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes, R. A. Young, *Cell* **2018**, *175*, 1842–1855.e16.
[27] H. J. Dyson, P. E. Wright, *J. Biol. Chem.* **2016**, *291*, 6714–6722.
[28] The UniProt Consortium, *Nucleic Acids Res.* **2021**, *49*, D480–D489.
[29] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.
[30] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, *Nature* **2021**, *596*, 590–596.
[31] D. Piovesan, A. M. Monzon, S. C. E. Tosatto, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.03.03.482768.
[32] T. R. Alderson, I. Pritišanac, A. M. Moses, J. D. Forman-Kay, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.02.18.481080.
[33] R. K. Das, R. V. Pappu, *Proc. Nat. Acad. Sci.* **2013**, *110*, 13392–13397.
[34] A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. G. Richardson, R. V. Pappu, *Biophys. J.* **2017**, *112*, 16–21.
[35] W. C. Wimley, S. H. White, *Nat. Struct. Biol.* **1996**, *3*, 842–848.
[36] C. J. Oldfield, V. N. Uversky, A. K. Dunker, L. Kurgan, in *Intrinsically Disord. Proteins*, Elsevier, **2019**, 1–34.
[37] M. Necci, D. Piovesan, CAID Predictors, DisProt Curators, S. C. E. Tosatto, *Nat. Methods* **2021**, *18*, 472–481.
[38] J. Janin, M. J. E. Sternberg, *F1000 Biol. Rep.* **2013**, *5*, 10.3410/B5-2.
[39] V. N. Uversky, *F1000Research* **2020**, *9*, 101.
[40] T. Zarin, B. Strome, A. N. Nguyen Ba, S. Alberti, J. D. Forman-Kay, A. M. Moses, *eLife* **2019**, *8*, e46883.
[41] K. Kasahara, H. Terazawa, T. Takahashi, J. Higo, *Comput. Struct. Biotechnol. J.* **2019**, *17*, 712–720.
[42] R. O. J. Weinzierl, *Biomol. Eng.* **2021**, *11*, 856.
[43] S. Müller-Späth, A. Soranno, V. Hirschfeld, H. Hofmann, S. Rüegger, L. Reymond, D. Nettels, B. Schuler, *Proc. Nat. Acad. Sci.* **2010**, *107*, 14609–14614.
[44] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, R. V. Pappu, *Proc. Nat. Acad. Sci.* **2010**, *107*, 8183–8188.
[45] A. L. Sanborn, B. T. Yeh, J. T. Feigerle, C. V. Hao, R. J. Townshend, E. Lieberman Aiden, R. O. Dror, R. D. Kornberg, *eLife* **2021**, *10*, e68068.
[46] T. Firman, K. Ghosh, *J. Chem. Phys.* **2018**, *148*, 123305.
[47] L. M. Tuttle, D. Pacheco, L. Warfield, D. B. Wilburn, S. Hahn, R. E. Klevit, *Nat. Commun.* **2021**, *12*, 2220.
[48] P. Boyle, C. Després, *Plant Signaling Behav.* **2010**, *5*, 629–634.
[49] N. DelRosso, J. Tycko, P. Suzuki, C. Andrews, Aradhana, A. Mukund, I. Liongson, C. Ludwig, K. Spees, P. Fordyce, M. C. Bassik, L. Bintu, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.08.26.505496.
[50] J. Ma, M. Ptashne, *Cell* **1987**, *51*, 113–119.
[51] C. M. Drysdale, E. Dueñas, B. M. Jackson, U. Reusser, G. H. Braus, A. G. Hinnebusch, *Mol. Cell. Biol.* **1995**, *15*, 1220–1233.
[52] I. Langstein-Skora, A. Schmid, R. J. Emenecker, M. O. G. Richardson, M. J. Götz, S. K. Payer, P. Korber, A. S. Holehouse, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.02.10.480018.
[53] N. Alerasool, H. Leng, Z.-Y. Lin, A.-C. Gingras, M. Taipale, *Mol. Cell* **2022**, *82*, 677–695.e7.
[54] L. Klaus, B. P. de Almeida, A. Vlasova, F. Nemčko, A. Schleiffer, K. Bergauer, M. Rath, A. Stark, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.08.26.505062.
[55] J. Tycko, N. DelRosso, G. T. Hess, Aradhana, A. Banerjee, A. Mukund, M. V. Van, B. K. Ego, D. Yao, K. Spees, P. Suzuki, G. K. Marinov, A. Kundaje, M. C. Bassik, L. Bintu, *Cell* **2020**, *183*, 2020–2035.e16.
[56] C. N. Ravarani, T. Y. Erkina, G. De Baets, D. C. Dudman, A. M. Erkine, M. M. Babu, *Mol. Syst. Biol.* **2018**, *14*, 10.15252/msb.20188190.
[57] A. Erijman, L. Kozlowski, S. Sohrabi-Jahromi, J. Fishburn, L. Warfield, J. Schreiber, W. S. Noble, J. Söding, S. Hahn, *Mol. Cell* **2020**, *78*, 890–902.e6.
[58] C. D. Arnold, F. Nemčko, A. R. Woodfin, S. Wienerroither, A. Vlasova, A. Schleiffer, M. Pagani, M. Rath, A. Stark, *EMBO J.* **2018**, *37*, DOI 10.15252/embj.201798896.
[59] M. V. Staller, A. S. Holehouse, D. Swain-Lenz, R. K. Das, R. V. Pappu, B. A. Cohen, *Cell Syst.* **2018**, *6*, 444–455.e6.
[60] M. V. Staller, E. Ramirez, S. R. Kotha, A. S. Holehouse, R. V. Pappu, B. A. Cohen, *Cell Syst.* **2022**, *13*, 334–345.e5.
[61] L. M. Tuttle, D. Pacheco, L. Warfield, J. Luo, J. Ranish, S. Hahn, R. E. Klevit, *Cell Rep.* **2018**, *22*, 3251–3264.
[62] M. Piskacek, M. Havelka, M. Rezacova, A. Knight, *PLoS One* **2016**, *11*, e0162842.
[63] G. I. Makhatadze, in *Adv. Protein Chem.*, Elsevier, **2005**, 199–226.
[64] M. Tanaka, W. M. Clouston, *Mol. Cell. Biol.* **1994**, *14*, 10.
[65] C. Chen, *Mol. Cell. Biol.* **1999**, *19*, 10.
[66] N. G. Faux, S. P. Bottomley, A. M. Lesk, J. A. Irving, J. R. Morrison, M. G. de la Banda, J. C. Whisstock, *Genome Res.* **2005**, *15*, 537–551.
[67] N. J. Maclaine, T. R. Hupp, *Cell Cycle* **2011**, *10*, 916–921.
[68] X. Hu, L.-F. Chen, *Front. Cell Dev. Biol.* **2020**, *8*, 179.
[69] T. Collins, J. R. Stone, A. J. Williams, *Mol. Cell. Biol.* **2001**, *21*, 3609–3615.
[70] N. T. Krogan, J. A. Long, *Curr. Opin. Plant Biol.* **2009**, *12*, 628–636.

[71] J. Ke, H. Ma, X. Gu, A. Thelen, J. S. Brunzelle, J. Li, H. E. Xu, K. Melcher, *Sci. Adv.* **2015**, *1*, e1500107.
[72] M. Boulanger, M. Chakraborty, D. Tempé, M. Piechaczyk, G. Bossis, *Molecules* **2021**, *26*, 828.
[73] M. A. Gillespie, C. G. Palii, D. Sanchez-Taltavull, P. Shannon, W. J. R. Longabaugh, D. J. Downes, K. Sivaraman, H. M. Espinoza, J. R. Hughes, N. D. Price, T. J. Perkins, J. A. Ranish, M. Brand, *Mol. Cell* **2020**, *78*, 960–974.e11.
[74] C. Neumayr, V. Haberle, L. Serebreni, K. Karner, O. Hendy, A. Boija, J. E. Henninger, C. H. Li, K. Stejskal, G. Lin, K. Bergauer, M. Pagani, M. Rath, K. Mechtler, C. D. Arnold, A. Stark, *Nature* **2022**, *606*, 406–413.
[75] V. Haberle, C. D. Arnold, M. Pagani, M. Rath, K. Schernhuber, A. Stark, *Nature* **2019**, *570*, 122–126.
[76] J. Jacobs, M. Pagani, C. Wenzl, A. Stark, *BioRxiv preprint* **2022**, DOI: 10.1101/2022.11.07.515017.
[77] L. Warfield, L. M. Tuttle, D. Pacheco, R. E. Klevit, S. Hahn, *Proc. Nat. Acad. Sci.* **2014**, *111*, 10.1073/pnas.1412088111.
[78] M. M. Babu, R. W. Kriwacki, R. V. Pappu, *Science* **2012**, *337*, 1460–1461.
[79] V. N. Uversky, *Curr. Pharm. Des.* **2013**, *19*, 4191–4213.
[80] Y. Huang, Z. Liu, *J. Mol. Biol.* **2009**, *393*, 1143–1159.
[81] L. Mollica, L. M. Bessa, X. Hanoulle, M. R. Jensen, M. Blackledge, R. Schneider, *Front. Mol. Biosci.* **2016**, *3*, 10.3389/fmolb.2016.00052.
[82] A. Bah, J. D. Forman-Kay, *J. Biol. Chem.* **2016**, *291*, 6696–6705.
[83] Y. Ivarsson, P. Jemth, *Curr. Opin. Struct. Biol.* **2019**, *54*, 26–33.
[84] C. O'Shea, L. Staby, S. K. Bendsen, F. G. Tidemand, A. Redsted, M. Willemoës, B. B. Kragelund, K. Skriver, *J. Biol. Chem.* **2017**, *292*, 512–527.
[85] M. Fuxreiter, *J. Mol. Biol.* **2018**, *430*, 2278–2287.
[86] P. Tompa, N. E. Davey, T. J. Gibson, M. M. Babu, *Mol. Cell* **2014**, *55*, 161–169.
[87] L. F. Christensen, L. Staby, K. Bugge, C. O'Shea, B. B. Kragelund, K. Skriver, *Sci. Rep.* **2019**, *9*, 18927.
[88] R. B. Berlow, M. A. Martinez-Yamout, H. J. Dyson, P. E. Wright, *Biochemistry* **2019**, *58*, 1354–1362.
[89] M. W. Risør, A. L. Jansma, N. Medici, B. Thomas, H. J. Dyson, P. E. Wright, *Biochemistry* **2021**, *60*, 3887–3898.
[90] K. Bugge, L. Staby, E. Salladini, R. G. Falbe-Hansen, B. B. Kragelund, K. Skriver, *J. Biol. Chem.* **2021**, *296*, 100226.
[91] K. Bugge, L. Staby, K. R. Kemplen, C. O'Shea, S. K. Bendsen, M. K. Jensen, J. G. Olsen, K. Skriver, B. B. Kragelund, *Structure* **2018**, *26*, 734–746.e7.
[92] L. Staby, K. Bugge, R. G. Falbe-Hansen, E. Salladini, K. Skriver, B. B. Kragelund, *Cell Commun. Signaling* **2021**, *19*, 2.
[93] P. Jaspers, M. Brosché, K. Overmyer, J. Kangasjär, *Plant Signaling Behav.* **2010**, *5*, 78–80.
[94] J. P. Vainonen, P. Jaspers, M. Wrzaczek, A. Lamminmäki, R. A. Reddy, L. Vaahtera, M. Brosché, J. Kangasjärvi, *Biochem. J.* **2012**, *442*, 573–581.
[95] L. Staby, A. D. Due, M. B. A. Kunze, M. L. M. Jørgensen, K. Skriver, B. B. Kragelund, *J. Mol. Biol.* **2021**, *433*, 167320.
[96] F. Friis Theisen, E. Salladini, R. Davidsen, C. Jo Rasmussen, L. Staby, B. B. Kragelund, K. Skriver, *J. Biol. Chem.* **2022**, *298*, 101963.
[97] I. Kuznetsova, B. Zaslavsky, L. Breydo, K. Turoverov, V. Uversky, *Molecules* **2015**, *20*, 1377–1409.
[98] G. D. Amoutzias, D. L. Robertson, Y. Van de Peer, S. G. Oliver, *Trends Biochem. Sci.* **2008**, *33*, 220–229.
[99] S. Xhani, S. Lee, H. M. Kim, S. Wang, S. Esaki, V. L. T. Ha, M. Khanezarrin, G. L. Fernandez, A. V. Albrecht, J. M. Aramini, M. W. Germann, G. M. K. Poon, *Sci. Adv.* **2020**, *6*, eaay3178.
[100] R. B. Berlow, H. J. Dyson, P. E. Wright, *Nature* **2017**, *543*, 447–451.
[101] M. Kumar, S. Michael, J. Alvarado-Valverde, B. Mészáros, H. Sámano-Sánchez, A. Zeke, L. Dobson, T. Lazar, M. Örd, A. Nagpal, N. Farahi, M. Käser, R. Kraleti, N. E. Davey, R. Pancsa, L. B. Chemes, T. J. Gibson, *Nucleic Acids Res.* **2022**, *50*, D497-D508.
[102] A. N. Nguyen Ba, B. J. Yeh, D. van Dyk, A. R. Davidson, B. J. Andrews, E. L. Weiss, A. M. Moses, *Sci. Signaling* **2012**, *5*, 10.1126/scisignal.2002515.
[103] N. E. Davey, M. S. Cyert, A. M. Moses, *Cell Commun. Signaling* **2015**, *13*, 43.
[104] T. Zarin, C. N. Tsai, A. N. Nguyen Ba, A. M. Moses, *Proc. Nat. Acad. Sci.* **2017**, *114*, 10.1073/pnas.1614787114.
[105] D. Lemas, P. Lekkas, B. A. Ballif, J. O. Vigoreaux, *J. Proteomics* **2016**, *135*, 191–200.
[106] H. A. Moesa, S. Wakabayashi, K. Nakai, A. Patil, *Mol. BioSyst.* **2012**, *8*, 3262.
[107] T. Zarin, B. Strome, G. Peng, I. Pritišanac, J. D. Forman-Kay, A. M. Moses, *eLife* **2021**, *10*, e60220.
[108] L. Hou, Y. Wei, Y. Lin, X. Wang, Y. Lai, M. Yin, Y. Chen, X. Guo, S. Wu, Y. Zhu, J. Yuan, M. Tariq, N. Li, H. Sun, H. Wang, X. Zhang, J. Chen, X. Bao, R. Jauch, *Nucleic Acids Res.* **2020**, *48*, 3869–3887.
[109] Z. E. Holmes, D. J. Hamilton, T. Hwang, N. V. Parsonnet, J. L. Rinn, D. S. Wuttke, R. T. Batey, *Nat. Commun.* **2020**, *11*, 10.1038/s41467-020-15571-8.
[110] M. De, M. A. Öztürk, S. Isbaner, K. Tóth, R. C. Wade, *Biophys. J.* **2021**, *120*, 3747–3763.
[111] D. T. Brown, T. Izard, T. Misteli, *Nat. Struct. Mol. Biol.* **2006**, *13*, 250–255.
[112] P. O. Heidarsson, D. Mercadante, A. Sottini, D. Nettels, M. B. Borgia, A. Borgia, S. Kilic, B. Fierz, R. B. Best, B. Schuler, *Nat. Chem.* **2022**, *14*, 224–231.
[113] O. G. Berg, R. B. Winter, P. H. Von Hippel, *Biochemistry* **1981**, *20*, 6929–6948.
[114] P. H. von Hippel, O. G. Berg, *J. Biol. Chem.* **1989**, *264*, 675–678.
[115] D. Vuzman, Y. Levy, *Isr. J. Chem.* **2014**, *54*, 1374–1381.
[116] C. Crane-Robinson, A. I. Dragan, P. L. Privalov, *Trends Biochem. Sci.* **2006**, *31*, 547–552.
[117] A. I. Dragan, C. M. Read, E. N. Makeyeva, E. I. Milgotina, M. E. A. Churchill, C. Crane-Robinson, P. L. Privalov, *J. Mol. Biol.* **2004**, *343*, 371–393.
[118] Y. Liu, K. S. Matthews, S. E. Bondos, *J. Biol. Chem.* **2008**, *283*, 20874–20887.
[119] J. J. Love, X. Li, D. A. Case, K. Giese, R. Grosschedl, P. E. Wright, *Nature* **1995**, *376*, 791–795.
[120] G. M. Lee, L. W. Donaldson, M. A. Pufall, H.-S. Kang, I. Pot, B. J. Graves, L. P. McIntosh, *J. Biol. Chem.* **2005**, *280*, 7088–7099.
[121] A. S. Krois, H. J. Dyson, P. E. Wright, *Proc. Nat. Acad. Sci.* **2018**, *115*, 10.1073/pnas.1814051115.
[122] I. Taneja, A. S. Holehouse, *Curr. Res. Struct. Biol.* **2021**, *3*, 216–228.
[123] B. Bourgeois, T. Gui, D. Hoogeboom, H. G. Hocking, G. Richter, E. Spreitzer, M. Viertler, K. Richter, T. Madl, B. M. T. Burgering, *Cell Rep.* **2021**, *36*, 109446.
[124] O. Laptenko, I. Shiff, W. Freed-Pastor, A. Zupnick, M. Mattia, E. Freulich, I. Shamir, N. Kadouri, T. Kahan, J. Manfredi, I. Simon, C. Prives, *Mol. Cell* **2015**, *57*, 1034–1046.
[125] O. Laptenko, D. R. Tong, J. Manfredi, C. Prives, *Trends Biochem. Sci.* **2016**, *41*, 1022–1034.
[126] H. E. R. Baughman, D. Narang, W. Chen, A. C. Villagrán Suárez, J. Lee, M. J. Bachochin, T. R. Gunther, P. G. Wolynes, E. A. Komives, *J. Biol. Chem.* **2022**, *298*, 102349.
[127] X. Sun, H. J. Dyson, P. E. Wright, *Proc. Nat. Acad. Sci.* **2021**, *118*, e2021456118.
[128] Y. Cicirò, A. Sala, *Oncogenesis* **2021**, *10*, 19.
[129] E. Werwein, A. Biyanee, K. Klempnauer, *FEBS Lett.* **2020**, *594*, 4266–4279.
[130] S. Brodsky, T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, M. Carmi, N. Barkai, *Mol. Cell* **2020**, 1–13.
[131] J. Chen, Z. Zhang, L. Li, B. C. Chen, A. Revyakin, B. Hajj, W. Legant, M. Dahan, T. Lionnet, E. Betzig, R. Tjian, Z. Liu, *Cell* **2014**, *156*, 1274–1285.
[132] T. Gera, F. Jonas, R. More, N. Barkai, *eLife* **2022**, *11*, e73225.
[133] M. V. Staller, *Genetics* **2022**, *222*, iyac111.
[134] M. J. Henley, B. M. Linhares, B. S. Morgan, T. Cierpicki, C. A. Fierke, A. K. Mapp, *Proc. Nat. Acad. Sci.* **2020**, *117*, 27346–27353.
[135] A. K. Dunker, V. N. Uversky, *Curr. Opin. Pharmacol.* **2010**, *10*, 782–788.
[136] K. Tsafou, P. B. Tiwari, J. D. Forman-Kay, S. J. Metallo, J. A. Toretsky, *J. Mol. Biol.* **2018**, *430*, 2321–2341.
[137] F. Zhu, L. Farnung, E. Kaasinen, B. Sahu, Y. Yin, B. Wei, S. O. Dodonova, K. R. Nitta, E. Morgunova, M. Taipale, P. Cramer, J. Taipale, *Nature* **2018**, *562*, 76–81.
[138] M. Fernandez Garcia, C. D. Moore, K. N. Schulz, O. Alberto, G. Donague, M. M. Harrison, H. Zhu, K. S. Zaret, *Mol. Cell* **2019**, *75*, 921–932.e6.
[139] M. Iwafuchi-Doi, K. S. Zaret, *Dev. Camb.* **2016**, *143*, 1833–1837.
[140] H. W. King, R. J. Klose, *eLife* **2017**, *6*, 1–21.
[141] L. A. Cirillo, K. S. Zaret, *Mol. Cell* **1999**, *4*, 961–969.
[142] L. F. R. C. I. F. D. J. M. Z. K. S. Cirillo, L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, K. S. Zaret, *Mol. Cell* **2002**, *9*, 279–89.
[143] K. L. Clark, E. D. Halay, E. Lai, S. K. Burley, *Nature* **1993**, *364*, 412–420.
[144] M. Iwafuchi, I. Cuesta, G. Donahue, N. Takenaka, A. B. Osipovich, M. A. Magnuson, H. Roder, S. H. Seeholzer, P. Santisteban, K. S. Zaret, *Nat. Genet.* **2020**, *52*, 418–427.
[145] M. A. Frederick, K. E. Williamson, M. Fernandez Garcia, M. B. Ferretti, R. L. McCarthy, G. Donahue, E. Luzete Monteiro, N. Takenaka, J.

Reynaga, C. Kadoch, K. S. Zaret, *Nat. Struct. Mol. Biol.* **2022**, 10.1038/s41594-022-00886-5.
[146] K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, S. Yamanaka, *Cell* **2007**, *131*, 861–872.
[147] A. Reményi, K. Lins, L. J. Nissen, R. Reinbold, H. R. Schöler, M. Wilmanns, *Genes Dev.* **2003**, *17*, 2048–2059.
[148] A. K. Michael, R. S. Grand, L. Isbel, S. Cavadini, Z. Kozicka, G. Kempf, R. D. Bunker, A. D. Schenk, A. Graff-Meyer, G. R. Pathare, J. Weiss, S. Matsumoto, L. Burger, D. Schübeler, N. H. Thomä, *Science* **2020**, *368*, 1460–1465.
[149] P. J. Flory, *J. Chem. Phys.* **1942**, *10*, 51–61.
[150] A. Sirur, D. De Sancho, R. B. Best, *J. Chem. Phys.* **2016**, *144*, 075101.
[151] S. Cocco, J. F. Marko, R. Monasson, *Comptes Rendus Phys.* **2002**, *3*, 569–584.
[152] J. F. Marko, E. D. Siggia, *Macromolecules* **1995**, *28*, 8759–8770.
[153] M. L. Broide, C. R. Berland, J. Pande, O. O. Ogun, G. B. Benedek, *Proc. Nat. Acad. Sci.* **1991**, *88*, 5660–5664.
[154] A. C. Dumetz, A. M. Chockla, E. W. Kaler, A. M. Lenhoff, *Biophys. J.* **2008**, *94*, 570–583.
[155] P. R. ten Wolde, D. Frenkel, *Science* **1997**, *277*, 1975–1978.
[156] S. Milles, K. Huy Bui, C. Koehler, M. Eltsov, M. Beck, E. A. Lemke, *EMBO Rep.* **2013**, *14*, 178–183.
[157] Y. Lin, D. S. W. Protter, M. K. Rosen, R. Parker, *Mol. Cell* **2015**, *60*, 208–219.
[158] Y. Fujioka, J. Md. Alam, D. Noshiro, K. Mouri, T. Ando, Y. Okada, A. I. May, R. L. Knorr, K. Suzuki, Y. Ohsumi, N. N. Noda, *Nature* **2020**, *578*, 301–305.
[159] T. Mittag, R. Parker, *J. Mol. Biol.* **2018**, *430*, 4636–4649.
[160] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, M. K. Rosen, *Nature* **2012**, *483*, 336–340.
[161] V. N. Uversky, *Annu. Rev. Biophys.* **2021**, *50*, 135–136.
[162] B. S. Schuster, E. H. Reed, R. Parthasarathy, C. N. Jahnke, R. M. Caldwell, J. G. Bermudez, H. Ramage, M. C. Good, D. A. Hammer, *Nat. Commun.* **2018**, *9*, 2985.
[163] C. P. Brangwynne, P. Tompa, R. V. Pappu, *Nat. Phys.* **2015**, *11*, 899–904.
[164] Z. Monahan, V. H. Ryan, A. M. Janke, K. A. Burke, S. N. Rhoads, G. H. Zerze, R. O'Meally, G. L. Dignon, A. E. Conicella, W. Zheng, R. B. Best, R. N. Cole, J. Mittal, F. Shewmaker, N. L. Fawzi, *EMBO J.* **2017**, *36*, 2951–2967.
[165] R. M. Vernon, P. A. Chong, B. Tsang, T. H. Kim, A. Bah, P. Farber, H. Lin, J. D. Forman-Kay, *eLife* **2018**, *7*, e31486.
[166] J. P. Brady, P. J. Farber, A. Sekhar, Y.-H. Lin, R. Huang, A. Bah, T. J. Nott, H. S. Chan, A. J. Baldwin, J. D. Forman-Kay, L. E. Kay, *Proc. Nat. Acad. Sci.* **2017**, *114*, 10.1073/pnas.1706197114.
[167] C. R. Martinez, B. L. Iverson, *Chem. Sci.* **2012**, *3*, 2191.
[168] B. L. Schottel, H. T. Chifotides, K. R. Dunbar, *Chem. Soc. Rev.* **2008**, *37*, 68–83.
[169] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, A. A. Hyman, *Cell* **2018**, *174*, 688–699.e16.
[170] A. Shakya, J. T. King, *ACS Macro Lett.* **2018**, *7*, 1220–1225.
[171] T. J. Nott, T. D. Craggs, A. J. Baldwin, *Nat. Chem.* **2016**, *8*, 569–575.
[172] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, A. J. Baldwin, *Mol. Cell* **2015**, *57*, 936–947.
[173] S. Rauscher, R. Pomès, *eLife* **2017**, *6*, e26526.
[174] A. C. Murthy, G. L. Dignon, Y. Kan, G. H. Zerze, S. H. Parekh, J. Mittal, N. L. Fawzi, *Nat. Struct. Mol. Biol.* **2019**, *26*, 637–648.
[175] G. Laflamme, K. Mekhail, *Commun. Biol.* **2020**, *3*, 773.
[176] K. Wagh, D. A. Garcia, A. Upadhyaya, *Curr. Opin. Struct. Biol.* **2021**, *71*, 148–155.
[177] J. Wang, H. Yu, Q. Ma, P. Zeng, D. Wu, Y. Hou, X. Liu, L. Jia, J. Sun, Y. Chen, D. Guallar, M. Fidalgo, J. Chen, Y. Yu, S. Jiang, F. Li, C. Zhao, X. Huang, J. Wang, C. Li, Y. Sun, X. Zeng, W. Zhang, Y. Miao, J. Ding, *Cell Stem Cell* **2021**, *28*, 1868–1883.e11.
[178] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, R. A. Young, *Science* **2018**, *361*, eaar3958.
[179] E. Apostolou, F. Ferrari, R. M. Walsh, O. Bar-Nur, M. Stadtfeld, S. Cheloufi, H. T. Stuart, J. M. Polo, T. K. Ohsumi, M. L. Borowsky, P. V. Kharchenko, P. J. Park, K. Hochedlinger, *Cell Stem Cell* **2013**, *12*, 699–712.
[180] J. A. Morin, S. Wittmann, S. Choubey, A. Klosin, S. Golfier, A. A. Hyman, F. Jülicher, S. W. Grill, *Nat. Phys.* **2022**, *18*, 271–276.
[181] R. Sharma, K.-J. Choi, M. D. Quan, S. Sharma, B. Sankaran, H. Park, A. LaGrone, J. J. Kim, K. R. MacKenzie, A. C. M. Ferreon, C. Kim, J. C. Ferreon, *Nat. Commun.* **2021**, *12*, 5579.
[182] G. Krainer, T. J. Welsh, J. A. Joseph, J. R. Espinosa, S. Wittmann, E. de Csilléry, A. Sridhar, Z. Toprakcioglu, G. Gudiškytė, M. A. Czekalska, W. E. Arter, J. Guillén-Boixet, T. M. Franzmann, S. Qamar, P. S. George-Hyslop, A. A. Hyman, R. Collepardo-Guevara, S. Alberti, T. P. J. Knowles, *Nat. Commun.* **2021**, *12*, 1085.
[183] J. Trojanowski, L. Frank, A. Rademacher, N. Mücke, P. Grigaitis, K. Rippe, *Mol. Cell* **2022**, *82*, 1878–1893.e10.
[184] A. Borgia, M. B. Borgia, K. Bugge, V. M. Kissling, P. O. Heidarsson, C. B. Fernandes, A. Sottini, A. Soranno, K. J. Buholzer, D. Nettels, B. B. Kragelund, R. B. Best, B. Schuler, *Nature* **2018**, *555*, 61–66.
[185] J. G. Greener, S. M. Kandathil, L. Moffat, D. T. Jones, *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55.

Manuscript received: October 30, 2022
Accepted manuscript online: January 17, 2023
Version of record online: March 6, 2023