Disentangling the CHAOS of intrinsic disorder in human proteins

Ida de Vries[*#], Jitske Bak[*#], Daniel Álvarez Salmoral[#], Ren Xie, Razvan Borza, Maria Konijnenberg, Anastassis Perrakis*

[#] Shared first authors
* Corresponding authors

Key messages:

1. 83% of human proteins contain one or more intrinsically disordered regions.
2. Intrinsically disordered regions are predominantly located at protein termini.
3. The amount and location of disordered regions s differs throughout cellular compartments.
4. AlphaFold predicts secondary structure elements within intrinsically disordered regions and fully disordered proteins.

Key words
Intrinsically disordered regions, intrinsically disordered proteins, essentiality, cellular location, post-translational modifications, human proteins

Abstract
Most proteins consist of both folded domains and Intrinsically Disordered Regions (IDRs). However, the widespread occurrence of intrinsic disorder in human proteins, along with its characteristics, is often overlooked by the broader communities of structural and molecular biologists. Building on the MobiDB database of intrinsically disorder in proteins, here we develop a comprehensive dataset (**C**omprehensive analysis of **H**uman proteins **A**nd their dis**O**rdered **S**egments (CHAOS)). We implement empirical internally consistent definitions of what constitutes a disordered region, annotate general characteristics such as cellular location, essentiality, and post-translational modifications, and cross-reference to structure predictions from AlphaFold. Most proteins contain at least one disordered region, predominantly located at the protein termini. IDRs are less hydrophobic and are enriched in post-translational modifications compared to non-IDRs. Additionally, we discovered that proteins residing in different cellular locations possess distinct disorder profiles. Finally, the predicted AlphaFold models of proteins in CHAOS suggest that while protein disorder may be intrinsic, it does not have to be extrinsic. Hereby we enhance the visibility and understanding of intrinsic disorder in human proteins.

Introduction

Proteins typically consist of folded domains and intrinsically disordered regions (IDRs). Folded domains are well characterized and contain a plethora of functions. Although the roles of IDRs are less well understood, there is increasing evidence that IDRs have important functions, like protein-protein interactions (also referred to as fuzzy complexes[1] or flexible nets[2]), aid rapid evolution[3], and can be involved in disease[4]. IDRs have also been implicated in numerous roles in cellular processes, such as signaling[5,6], protein degradation[7], translation initiation[8], membrane remodeling[9], viral infection[10], and chromatin remodeling[11].

Studies focusing on IDRs typically rely on fluorescent microscopy. While these are very useful to study the spatiotemporal nature of interactions, these techniques cannot visualize proteins at the atomic level. Structural biology is well equipped to study folded domains, however the techniques are poorly equipped to visualize IDRs and their interactions *in vivo* or in a native environment. X-ray analysis[12,13] requires the crystallization of protein, which is hindered by the presence of IDRs. Cryo-EM[14,15] tolerates the presence of IDRs, however these often stay invisible or blurry due to a lower signal to noise ratio. Despite NMR[16,17] being the most suitable technique for detecting more flexible regions, protein structure determination is hindered by protein size, labeling and solvability. These experimental limitations led scientists massively ignoring the existence of disordered regions in proteins[18,19].

The growing capabilities of computational studies have enabled the identification and characterization of intrinsically disordered regions (IDRs). Previously, there has been a focus on predicting protein regions that are disordered based on the amino acid sequence of the protein. Examples of deployed methods and databases are the IUPRED[20], GLOBPLOT[18] and DISEMBL[19]. Amusingly, these servers have mostly been developed to aid protein crystallization[19,21], neglecting the potential of studying IDRs themselves. More recently, Artificial Intelligence techniques made it possible to predict relatively accurate atomic structures of proteins by algorithms like AlphaFold2[22], AlphaFold3[23] and RosettaFold[24]. Although these techniques were also developed to predict the structured elements of proteins, they drastically increased the visibility of protein disorder at atomic level[25,26].

Taking advantage of computational algorithms and available experimental data, efforts have been made to improve annotations of IDRs. A notable information-rich resource among them is MobiDB[27]. Besides curated and predicted IDR annotations, MobiDB contains annotations of post translational modifications (PTMs), as well as annotations of phase separating proteins. In this work, we build on this database, adding annotations on essentiality, cellular location (from UniProt), and PTMs (from dbPTM) to create a comprehensive dataset of IDRs in human proteins. We look at the abundance and properties of IDRs in the protein sequence and in the context of the human cell. We study IDRs in the context of predicted protein structure models, and characterize fully disordered proteins.

What are IDRs?

A problem studying IDRs is that there is no formalized definition. There is no consensus about the minimum number of consecutive disordered residues. As such, some studies define IDRs as being disordered regions in proteins longer than 5[28], 10[26], or even longer than 30 amino acids[29]. In this work, we set the minimum length for IDRs at 8 residues. The average length of unmodeled loops (regions shorter than 30 residues) in the PDB-REDO databank[30] is 7 amino acids (median 5 residues). As IDRs are supposed to be longer than loops, defining an IDR as 8 or more consecutive disordered residues seems reasonable for the purpose of this study.

Using the above definition, we build our dataset: **C**omprehensive analysis of **H**uman proteins **A**nd their dis**O**rdered **S**egments (CHAOS). First, we mined the IDR locations of human proteins as annotated in MobiDB. Additionally, we downloaded the predicted protein structure models from the AlphaFold protein structure database[31,32]. These models contain a per residue pLDDT (Predicted Local Distance Difference Test) score, which typically is low (confident) for IDRs and high (confident) for structured regions (non-IDRs)[25,26]. To annotate secondary structure elements in these protein models, we ran DSSP (Dictionary of Secondary Structure in Protein) to define if the residues in the models adopt a secondary structure or not.

To polish the data (Figure 1A, Methods Figure 1), we extend an IDR to the N-, or C-terminus when there are less than 8 residues from the start or end of the protein to the start or end position of an IDR. Further, when there are less than 8 residues in between two IDRs, we merge these two IDRs with the residues in between them into a single IDR. Next, we merge or extend IDRs with their residues in between or to the termini, when there are 8 to 16 residues, and these residues adopt little secondary structure elements or these residues show an average pLDDT score below 70. Finally, we remove regions shorter than 8 residues. The details of this data polishing procedure can be found in the methods section.

Prevalence and location of IDRs in human proteins

The resulting dataset contains 40,756 IDRs in 20,038 human proteins. From this, we define three protein categories: IDR-containing proteins, Intrinsically Disordered Proteins (IDPs) and proteins without IDRs (Figure 1B). The majority (83%) are IDR-containing proteins, which are proteins with one or multiple IDRs *and* one or multiple non-IDRs. 5% of the proteins are IDPs, these are proteins that contain a single IDR that spans the entire protein. Lastly, 11% of the proteins contain no IDRs at all. In total, the disordered residues account for 33% of all amino acids (Figure 1C).

The majority of all IDRs (57%) are located at protein termini. The abundance of IDRs leads to 82% of all proteins containing terminal IDRs. The majority (64%) contains an N-terminal and half (50%) contain a C-terminal IDR (Figure 1D). The probability that an IDR with an average length (82 residues) is positioned in either the first or the last 8 residues of a protein with average length (523 residues) is around 4% ((8/(523-82))*2). This is substantially lower than the mentioned 82%, suggesting that proteins are highly biased to having IDRs at their termini.

Noteworthy is that some terminal IDRs span nearly the entire protein (Supplemental Tables 1 and 2).

On the other hand, internal IDRs (42%) are less prevalent. Less than half (43%) of the proteins have one or more internal IDR (Figure 1D). The idea that internal IDRs could be viewed as linker regions[3], suggests that proteins with internal IDRs are longer than proteins without internal IDRs. Indeed, internal IDRs (average 739 residues) are considerably longer than proteins with N- or C-terminal IDRs (average 550 and 574 residues, respectively) (Supplemental Figure 1A). We find a correlation above the diagonal between the number of disordered residues and protein length (Pearson correlation coefficient 0.70) (Supplemental Figure 1B), suggesting long proteins contain relatively more disorder. This is mainly affected by more (Supplemental Figure 1C), rather than longer internal IDRs (Supplemental Figure 1D). However, with an average of 2.0 IDRs per protein and the bias to terminal proteins, a protein is unlikely to contain multiple internal IDRs.

Indeed, the most likely combination of IDRs in a protein are an N- and a C-terminal IDR together, without internal IDRs (Figure 1E). Containing only an N-terminal IDR is the second most abundant group, while proteins containing a C-terminal IDR or proteins without terminal IDRs are the third and fourth most abundant group. Every internal IDR added to any of these four categories reduces the likelihood of finding a protein with that particular combination.

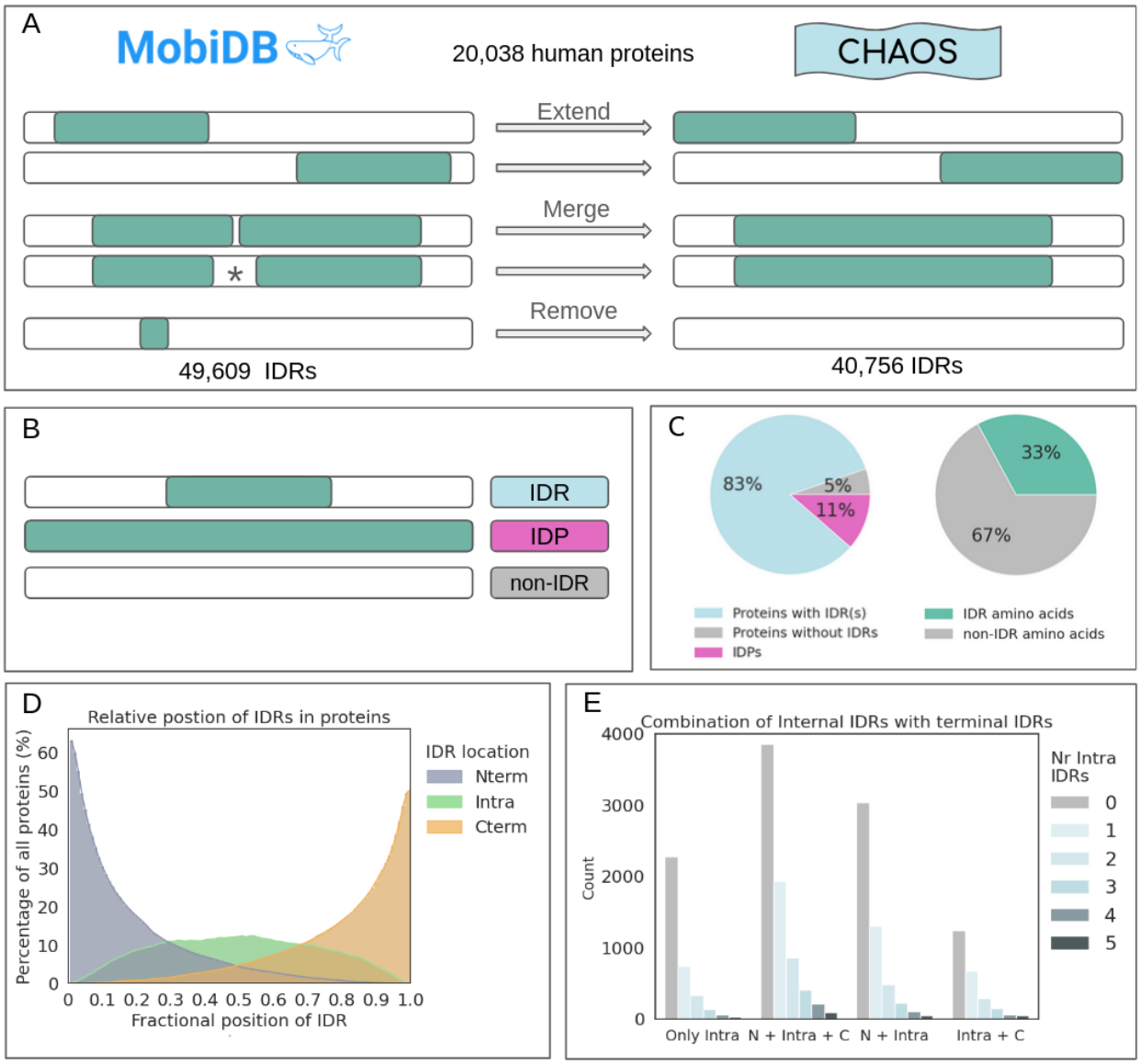


Figure 1: IDRs in human proteins. ***A)*** *Schematic overview on how the IDRs obtained from MobiDB were extended, merged or removed resulting in our dataset: Comprehensive analysis of Human proteins And their disOrdered Segments (CHAOS) with 20,038 proteins and 40,756 IDRs. *:merged under special conditions, see methods.* ***B)*** *Schematic representation of the three protein categories considered in this work. From top to bottom: IDR containing proteins, Intrinsically Disordered Proteins (IDPs), proteins without IDRs (non-IDR proteins).* ***C)*** *Left: Abundance of IDR containing proteins, proteins without IDRs and IDPs in this work. Right: The IDRs account for 33% of all amino acids in the dataset.* ***D)*** *Relative positioning of IDRs in all proteins. N-terminal (Nterm) IDRs are 32% of all IDRs, internal IDRs (Intra) account 42% and C-terminal IDRs (Cterm) are 25%.* ***E)*** *Prevalence of combinations of IDR location in proteins with and without IDRs. Only intra: protein with only internal IDRs, N + Intra + C: protein with an N-terminal, internal and C-terminal IDR, N + Intra: protein with an N-terminal and internal IDR, Intra + C: protein with an internal and C-terminal IDR. The colors indicate the number of internal IDRs.*

Where in the human cell do we find disorder?

To study whether the amount of disorder differs at different subcellular locations, we incorporated UniProt-annotated subcellular lobations and categorized each protein into different categories: plasma membrane, cytoplasm, secreted, cortex, nucleus, single organel, multiple organelles or unknown localization (see methods for details) (Figure 2A). Note that each protein has been assigned to a single category, but in reality proteins can shuttle between multiple cellular locations[33]. Our classification roughly agrees with published numbers[34,35], indicating the categorization is fair.

Proteins at different subcellular locations have different fractions of disorder (Figure 2B). Cortical, organellar and secreted proteins contain similar amounts of disorder (33-34%), while cytoplasmic and plasma membrane proteins contain less disorder (27%). Proteins without a known location contain more disorder (44%). The latter is likely due to experimental difficulties with highly disordered proteins, underlining the importance of this study. Within the organel category we also find substantial differences in fractions of disorder (Supplemental Figure 2A). For example, autophagosome and endosomes contain a higher fraction of disorder, while e.g. peroxisomes, lysosomes and mitochondrial proteins have less disorder. It is possible that proteins in peroxisomes and lysosomes benefit from less disorder to prevent proteolysis as an alternative to glycosylation. In mitochondria the low disorder might be related to their evolutionary descent from bacteria, which contain less protein disorder than human proteins[3].

When investigating the number of IDRs per protein (Figure 2B), we observe that proteins located at the cortex uniquely contain more IDRs per protein (2.7 IDRs per protein) than the average IDR containing protein (2.4 IDRs per protein). Plasma membrane and secreted proteins contain least IDRs per protein (1.9, 1.8 IDRs per protein respectively), while the cytoplasm, and organelles are just slightly below average (2.1 and 2.1 IDRs per proteins, respectively). The protein length at different subcellular locations also differs, for example, cortical proteins are particularly long (Supplemental Figure 2B).

These data show that protein characteristics at different subcellular locations contribute to a unique environment at their specific locations. For example, long, highly disordered cortical proteins form a uniquely disordered environment just below the cell surface. The high amount of disorder at the cell cortex might be important for sensing and remodeling membranes[36–38].

Cellular location determines IDR location

Most secreted and plasma membrane proteins contain N-terminal IDRs (90%, 73%, respectively) (Figure 2B). The high likelihood for N-terminal IDRs might be due to the annotation of N-terminal localization peptides as IDRs. Such peptides are required to secrete or translocate proteins into cell membranes. As localization peptides are often proteolytically removed[39]**,** these numbers likely do not reflect the functional composition of IDRs in secreted and plasma membrane proteins, but rather the complete protein before any processing. Further, less secreted proteins contain C-terminal IDRs (33%), while more proteins at the plasma membrane

have a C-terminal IDR (61%). Secreted and plasma membrane proteins both contain less internal IDRs (34%, 36%). These might render secreted and cell membrane proteins susceptible to proteolysis, altering their function or regulating activity levels.

Proteins that localize at the cytoplasm, organelles or cortex contain fewer N-terminal IDRs (50%, 65%, 66%) (Figure 2B). In turn, these proteins are enriched in either or both C-terminal or internal IDRs. Cytoplasmic and cortical proteins are enriched for internal IDRs (49%, 60%), while organellar and cortical proteins are enriched in C-terminal IDRs (41% and 65%, respectively). From these data it becomes clear that the location of the IDR in the protein is biased, which is reflected in differences per cellular location and protein length. Whether this is cause or causality remains unclear.

Essentiality

We investigated how IDRs manifest in the human essentialome. This is the collection of genes or proteins that is required for survival and proliferation in cells, highlighting fundamental pathways. Note that essentiality annotations we used stem from single cell studies[40,41] and might differ from essentiality in whole organisms. In our dataset we find 3431 (17%) essential proteins, of which half are found in the nucleus (Figure 2C). Essential proteins that contain IDRs, have a similar fraction of disorder (32%) than all IDR containing proteins (34%) (Figure 2C, Supplemental Figure S2C). Also the average number of IDRs (2.5) does not substantially differ from non-essential IDR containing proteins (2.4) (Figure 2C, Supplemental Figure S2C). Note that essential proteins are approximately equally likely to contain a N-terminus, C-terminus or internal IDR (50%, 44%, 49%) (Figure 2C).

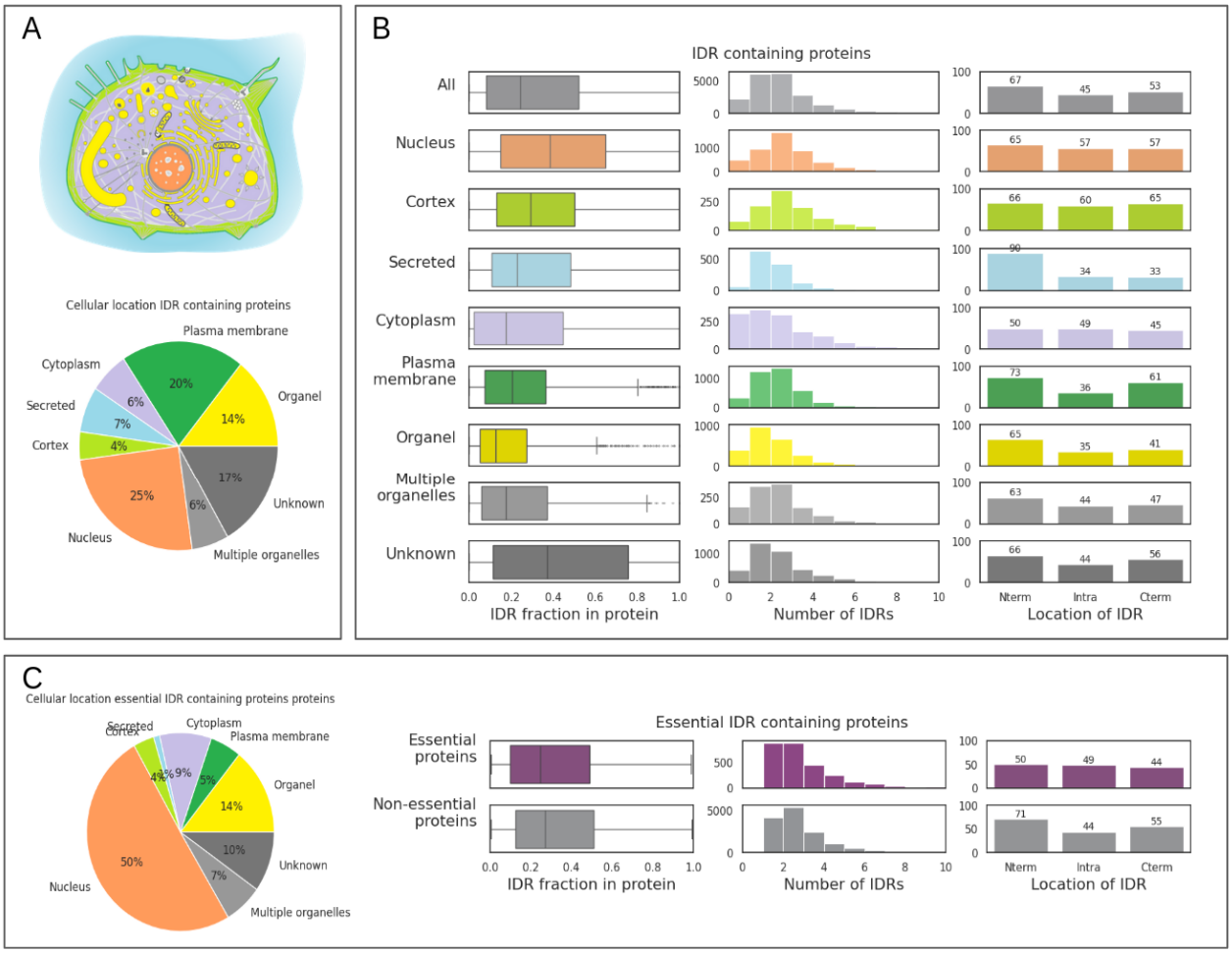


Figure 2: Cellular localization of IDR containing proteins. ***A)*** *Graphic representation of the cellular locations defined for this work, wherein each color indicates a category (background figure obtained from https://www.swissbiopics.org/name/Animal_cell). The pie plot shows the distribution of proteins amongst these cellular locations.* ***B)*** *Boxplots of the fraction of IDR residues in IDR containing proteins (left). Distribution of the number of IDRs in IDR-containing proteins, wherein the y-axis represents the number of proteins (middle). The percentage (y-axis) of IDR containing proteins that contains an N-terminal (Nterm), internal (Intra) and C-terminal (Cterm) IDR (right). The row 'All' contains all IDR containing proteins, the following rows hold IDR containing proteins found at the mentioned cellular location.* ***C)*** *Cellular location of essential IDR containing proteins (left). Comparison of essential and non-essential IDR containing proteins for the fraction of IDR residues in the proteins, the number of IDRs in the protein and where the IDRs locate, similar as in B.*

What are intrinsic properties of IDRs?

How long are IDRs?

The length of IDRs might influence their function and properties[42–45]. On average, IDRs are 82 residues long, which is shorter than their non-IDR counterparts (197 residues) (Figure 3A). Cortical and organelles contain the longest IDRs while secreted and membrane proteins contain the shortest IDRs. Overall, N-terminal (74 residues) and internal IDRs (77 residues) are of similar length, while C-terminal IDRs are longer (98 residues) (Supplemental Figure 3A). At different cellular locations we observe that membrane and secreted proteins have shorter N-terminal IDRs than other locations. Cortical C-terminal IDRs are the longest while the length of internal IDRs is relatively constant at different locations (Supplemental Figure 3B).

IDRs are not hydrophobic

The chemical composition of IDRs and non-IDRs shows a different profile (Figure 3B). IDRs are less hydrophobic and slightly enriched in polar and special residues compared to non-IDRs. Cationic and anionic (charged) residues occur in equal fractions. These findings are in line with other studies[45–48]. When investigating the chemical properties of IDRs based on their position in the protein, we find that N-terminal IDRs have the most distinctive chemical profile (Supplemental Figure 4). These IDRs are considerably more hydrophobic than internal and C-terminal IDRs. The increased hydrophobicity could be due to localization peptides, which are known to contain a hydrophobic region. Such localization peptides are used to direct proteins to organelles, into the secretory pathway and translocate membrane proteins into the membrane[49]. Confirmingly, we find increased hydrophobicity of N-terminal IDRs in secreted, organellar and plasma membrane proteins (Figure 3C).

IDRs are enriched in PTMs

Just as the amino acid sequence of IDRs yields a unique chemical profile, post-translational modifications might (PTMs) do the same, while exerting important regulatory functions[50,51]. Similar to other studies[52], we find that residues that can be modified are relatively more prevalent in IDRs (7%) than in non-IDRs (3.6%) (Figure 3D). This could be explained by surface accessibility, amino acid proportion or the regulatory role of IDRs. The number of PTMs that can be added to a particular residue is similar in IDRs and non-IDRs (2.0 PTMs per PTM residue)(Figure 3D). Be aware that more PTMs and/or PTM residues in either IDRs or non-IDRs might be found in the future.

The most enriched modifications in IDRs relative to non-IDRs are lactylation (polar), butyrylation and crotonylation (hydrophobic modifications) (Figure 3D). All these modifications occur on lysines and fall under the acylation umbrella[53]. The fourth most enriched PTM type in IDRs is hydroxylation. The latter seems particularly enriched in secreted proteins (Supplemental Figure 5), where it might aid the formation of supramolecular structures and lower the susceptibility to

proteolysis[54,55]. Essential proteins contain relatively more IDR PTM sites (11%), which is in line with a previous study that shows more PTMs are found in essential proteins[56].

The most commonly found PTM in our dataset is phosphorylation, which is not clearly enriched in IDRs or non-IDRs. However, phosphorylation on serine residues is overrepresented in IDRs (Figure 3D), while phosphorylation on tyrosine residues is enriched in non-DRs. Threonine phosphorylation is not enriched in either category. Of note, both acetylation and ubiquitination are considerably enriched in non-IDRs. The latter could indicate IDRs are susceptible to different degradation pathways than non-IDRs, which possibly is an important factor in cell homeostasis[57].

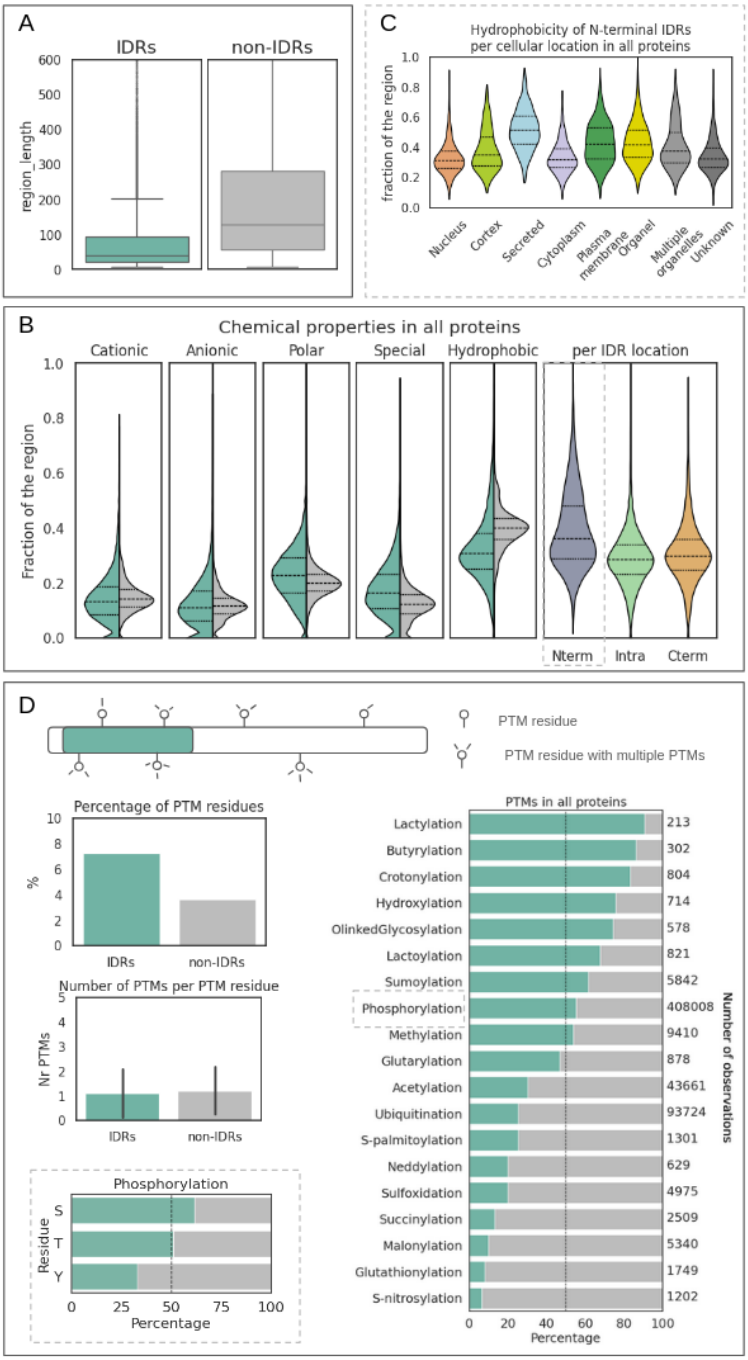


Figure 3: Properties of IDRs and non-IDRs. ***A)*** *Length in the number of residues for IDRs and non-IDRs in our dataset.* ***B)*** *Chemical properties of IDRs and non-IDRs divided into cationic (R,H,K), anionic (D,E), polar (S,T,N,Q), hydrophobic (A,V,I,L,M,F,Y,W) and special (C,P,G) residues. Hydrophobicity is also shown per IDR location in the protein (Nterm: N-terminal, Intra: internal, Cterm: C-terminal IDR).* ***C)*** *Hydrophobicity of N-terminal IDRs found in IDR containing proteins at different cellular locations.* ***D)*** *Schematic representation of PTMs in a protein, wherein a residue can obtain a single or multiple PTM (top). Top bar plot illustrates the percentage of IDR and non-IDR residues that can receive one or multiple PTMs, lower bar plot shows the number of PTMs per PTM site in IDRs and non-IDR residues, the error bar represents the standard deviation. Right: Relative presence of PTMs in IDRs vs. non-IDRs for all proteins in the dataset. The dotted vertical lines mark the 50% border, and the numbers on the right hand sides represent the number of PTMs found in total for that bar. For clarity only PTM types that were found >=200 times in our dataset are shown. Phosphorylation is highlighted in the dotted box, where the rows represent the amino acids that can get phosphorylated (S: serine, T: threonine, Y: tyrosine).*

What have we learned from protein structure prediction methods?

To capture the discrepancies between AlphaFold models and our data, we calculated the percentage of residues that adopt a secondary structure, and the average pLDDT value for each IDR and non-IDR (Figure 4A). Expectedly, we find lower pLDDT values and lower percentages of residues that adopt a secondary structure for IDRs than for non-IDRs (Figure 4B). Typically the pLDDT of IDRs is low (51 on average), however the percentage of residues that adopt a secondary structure averages around 38%. This is considerably higher than the naive expectation that IDRs contain 0% secondary structure. Strikingly, AlphaFold predicts some IDRs to be completely ordered. When we select such regions, we find 465 proteins containing 577 IDRs with a high percentage of secondary structure in total, of which 85% is an internal IDR. Based on cellular localization, IDR and non-IDR length, and chemical properties, they are not similar to non-IDRs (Supplemental Figure 6).

Unexpectedly, AlphaFold also predicts some non-IDRs to be fully disordered. These have low pLDDT and adopt little secondary structure. Selecting these non-IDRs, we find 338 non-IDRs in 304 proteins, which are enriched in N-terminal non-IDRs (41%). Based on their length and chemical properties these non-IDRs are similar to IDRs, and could therefore be incorrectly annotated (Supplemental Figure 6). Overall, AlphaFold increased the visibility of IDRs, and highlights that IDRs might not be fully disordered at all times.

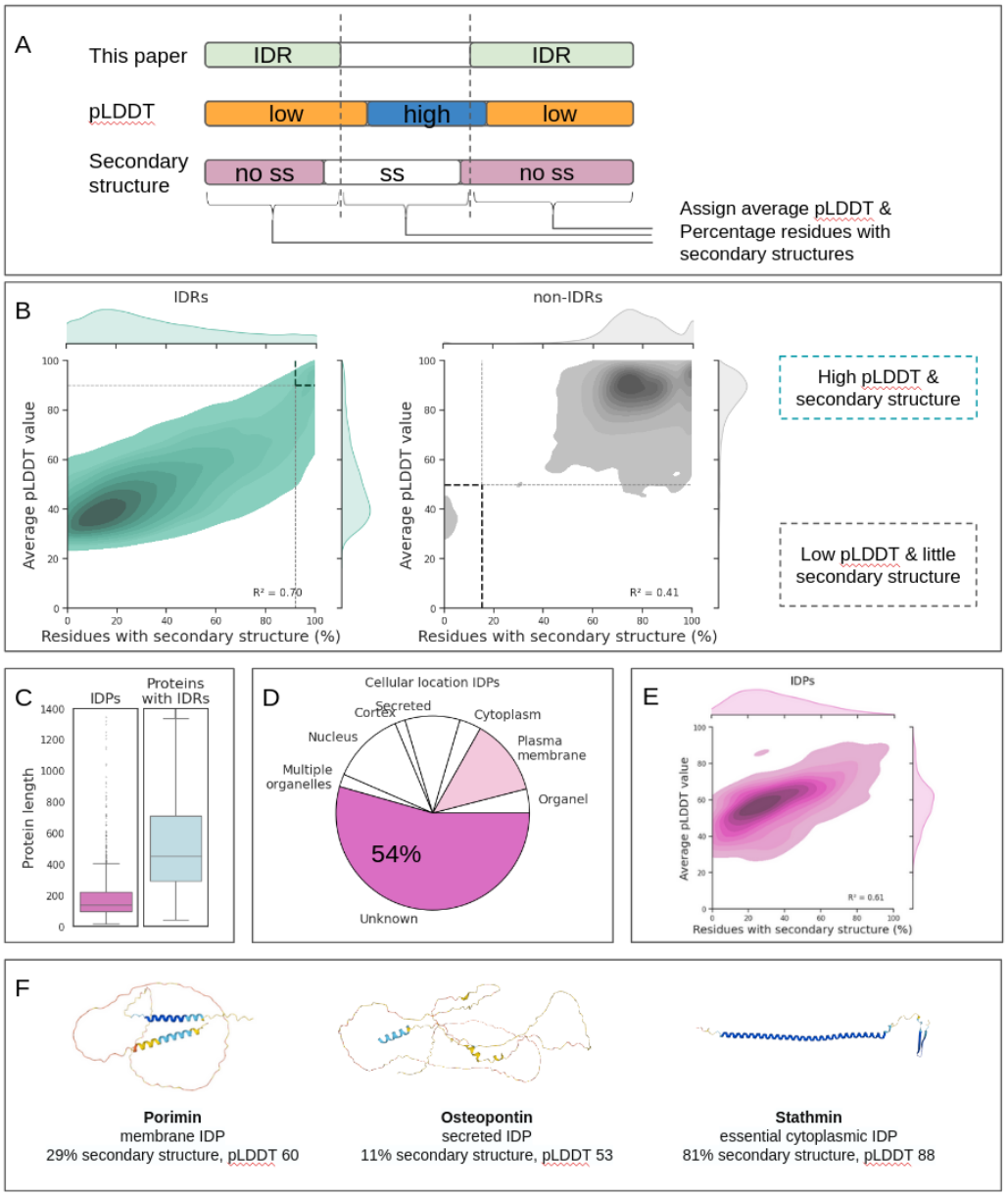


Figure 4: IDRs and IDPs are not fully disordered. ***A)*** *Schematic representation of IDRs in a protein sequence. The pLDDT score of the corresponding AlphaFold model and the percentage of residues with secondary structure do not perfectly align to the CHAOS dataset. Average values are annotated per region.* ***B)*** *Correlation between the percentage residues that adopt secondary structure and the average pLDDT score of the IDRs (left) and non-IDRs (right) found in our dataset. R^2 marks the Pearson correlation coefficient. The green dotted lines in the IDR plot (pLDDT >=90, percentage of residues that adopt secondary structure >=92%) mark IDRs with high pLDDT and high percentage of secondary structure. The black dotted lines in the non-IDR plot (pLDDT <=50, percentage of residues that adopt secondary structure <15%) mark non-IDRs with low pLDDT and little secondary structure.* ***C)*** *The length of IDPs and IDR containing proteins.* ***D)*** *The cellular localization of IDPs in our dataset.* ***E)*** *Correlation between the percentage residues that adopt secondary structure and the average pLDDT score of IDPs. R^2 marks the Pearson correlation coefficient.* ***F)*** *3D representation of example cases of IDPs to illustrate IDPs are not predicted to be fully disordered. Coordinates originate from the AlphaFold database (Porimin: ID Q8N131, Osteopontin: ID P10451, Stathmin: ID Q9H169), and the models are colored by pLDDT (dark blue: pLDDT >90, light blue: 90 > pLDDT > 70, yellow: 70 >= pLDDT >50, orange pLDDT <=50).*

Are IDPs more ordered then they should be?

In this work, we define Intrinsically Disordered Proteins (IDPs) as proteins that contain a single IDR spanning from the first to the last residue. This strict annotation results in 1046 (5%) IDPs. These proteins are on average 200 residues long (Figure 4C), which is considerably longer than the average IDRs (82 residues), but considerably shorter than average proteins (523 residues).

More than half of the IDPs (569, 54%) lack subcellular location annotation (unknown) (Figure 4D). This is likely due to experimental difficulties and underlines the importance of a broad characterization. Strikingly, we also find a considerable amount of IDPs located in the plasma membrane (13%). Membrane proteins typically contain a transmembrane spanning helix (typically 20 amino acids long[58]), which implies IDPs in the plasma membrane cannot be fully disordered at all times. Most likely such proteins form β-barrels, or similar secondary structures that only exist when multiple strands are lined up. This implies that membrane-IDPs should be considered in their native environment and with their native stoichiometry. It is also possible, that protein disorder may enhance the fluidity of the plasma membrane.

We propose defining IDPs as a distinct class of proteins, with membrane IDPs as a specific subclass. To our knowledge only MINAR-like proteins (e.g. P59773 and Q9UPX6) have been characterized as a distinct membrane-IDP protein family in Pfam, however IDPs and membrane IDPs as such, have not been classified[59].

Although the chemical properties of IDPs and IDRs are more similar than that of non-IDRs, IDPs contain relatively more special residues (C,P,G) (Supplemental Figure 7A). We also find less PTMs sites on IDPs (5%) than in IDRs (7%). This may be due to a lack of studies on IDPs, similar to the absence of cellular localization annotations. Further, we note that only 1.4% (49/3431) of essential proteins is an IDP, whereas 4.7% (985/18,992) of non-essential proteins is an IDP. This substantial difference suggests that combining the characteristics of essentiality and IDPs is evolutionary disadvantageous.

While IDPs are expected to be fully disordered, the AlphaFold-predicted that 35% of IDP residues adopt a secondary structure with an average pLDDT score of 58 (Figure 4E). Both values are higher for IDPs than for IDRs. The pLDDT and percentage of residues that adopt a secondary structure are even higher for plasma membrane IDPs (Supplemental Figure 7B). For example, Porimin, a membrane-IDP, contains helical parts accounting for 29% residues with secondary structure (Figure 4F). On the other hand, Osteopontin, a secreted IDP, looks more like the expected IDPs, having almost no secondary structure elements (11%) (Figure 4F). Exemplifying the complexity of IDPs, we highlight Stathmin, an essential cytoplasmic IDP that regulates microtubule dynamics. The protein is predicted with almost solely secondary structure elements (81%) (Figure 4F). Although in solution Stathmin indeed behaves as an IDP, when forming a protein complex with tubulin, in its native environment, Stathmin adopts a well-defined structure[60]. These examples illustrate that while IDPs can be disordered on their own, they could adopt secondary structure elements depending on their biological environment. Thus,

IDPs are a unique class of short proteins that display more secondary structure than their name suggests.

Conclusion and Outlook

The knowledge of protein structures and structure-related functions is rapidly expanding. Despite the increasing attention for IDRs in proteins[61], systemic characterization is lacking. Our analysis provides a unique view, as we combine publically available data and predicted structures of human proteins to form a comprehensive dataset: **C**omprehensive analysis of **H**uman proteins **A**nd their dis**O**rdered **S**egments (CHAOS), describing IDRs and proteins.

Overall**,** we find that the vast majority of proteins contain one or more IDRs. Strikingly, terminal IDRs are a lot more prevalent than internal IDRs. We also find that proteins containing internal IDRs are considerably longer. Furthermore, the cellular location of a protein determines the preferred IDR location in the protein sequence. N-terminal localization peptides are likely annotated as IDRs. Therefore it might be worth considering localization peptides as a subclass of IDRs. However, more detailed studies on differences and similarities of IDRs and localization peptides are complicated due to their wide variety[62] and lack of experimental validation[63].

IDRs themselves are enriched in PTMs and less hydrophobic than non-IDRs. To characterize IDRs in more detail, our dataset could be further enriched with conservation scores. Traditional conservation algorithms alone might not be particularly informative on IDRs[64]. Adding other protein characteristics like linear motifs, genetic diversity or conformational ensembles[65], could aid to assign informative conservation scores to IDRs and their characteristics. Together these could lead to a more detailed classification of IDRs in proteins, which ultimately sheds light on how differences in IDR composition relate to functionality.

Disordered regions are known to be able to adopt secondary structure elements depending on their biochemical context, e.g. proximity of another protein[66]. Proteins containing such conditionally folding regions might have a distinctive IDR profile. However, our analyses on proteins containing one or multiple disorder-to-order regions, did not indicate any noteworthy differences compared to IDRs in proteins that do not have a conditionally folding annotation. The analysis of pLDDT and secondary structure elements of IDRs in the predicted protein structure models did not show a distinctive profile, which one might have expected[25,26].

As IDRs have been proposed as the driving force of phase separation[44], we identified phase separating proteins in our dataset based on annotations in MobiDB with the intention to understand how their IDRs are different from IDRs that do not induce phase separation. We found that phase separating proteins contain a higher percentage of disorder and their IDRs are longer than proteins that are not annotated to phase separate (Supplemental Figure 8A and B). We also found relative enrichment of methylation in IDRs in phase separating proteins (Supplemental Figure 8C). Methylation reduces charge and increases the hydrophobicity, and might be acting as an on/off switch to phase separating behavior[67]. However, our data only contains 173 phase separating proteins, which we do not consider sufficient to draw robust

conclusions. Further we find 12 proteins without IDRs that are annotated to be involved in phase separation (Supplemental Figure 8D). These annotations could simply be a glitch in our data. However, they might also have been annotated because the discrepancy between a protein *driving* phase separation or *being* in the phase separation bubble is not always clear. Further, it begs the question ‘Can proteins without IDRs also phase separate?’.

Lastly, we find that both IDRs and IDPs are predicted to be more ordered than they should be in their AlphaFold protein structure models. Protein-protein interactions or PTMs[65,66] might aid the formation of structured elements, which could serve as initial contact points for interactions with other proteins. Indeed, IDPs can have a packing density and core hydration[65] or adopt 3D structures[66]. This emphasizes the complex nature of IDRs and IDPs, highlighting the need for further research.

Methods

Data collection

The data of the human genome was downloaded from MobiDB[68] on 2022-11-02 in JSON format. We chose MobiDB because it is user-friendly both for humans and computers, and was recently updated (2022-07) at the time the project started. Swiss-prot annotated human protein information was downloaded from UniProtKB[69] on the same day.

Dictionary of Secondary Structure in Proteins (DSSP)[30] was used to define the secondary structure elements that AlphaFold2[22] predicted for the human proteins (AlphaFold_human, Reference Proteome: UP000005640). The secondary structure elements were stored per residue. Turn (T), Bend (S) and Other (“ ”) were defined as disordered residues, whereas Alpha-helix (H), Beta-bridge (B), Strand (E), Helix_3 (G), Helix_5 (I) and Helix_PPI (P) were considered as ordered residues.

UniProt primary accession codes for which no AlphaFold prediction was available in the database (4 proteins) were deleted from the dataset, as no DSSP information could be retrieved. Additionally, the proteins for which the coordinates are stored in multiple coordinate files in the AlphaFold database (208 proteins), were removed from the dataset as these are hard to handle. Also the members of Endogenous retrovirus group K (37 proteins) were removed from the dataset as these proteins were not present in the AlphaFold_human dataset, so no DSSP information could be obtained. An additional set of UniProt primary accession codes (34 proteins) was deleted from the data, as these did not appear to be present in the AlphaFold_human data. The UniProt primary accession codes of the removed proteins can be found in Supplemental Methods 1. This led to a final selection of 20,038 human UniProt primary accession codes.

To be able to incorporate information about PTMs (post-translational modifications), the PTM annotations from MobiDB were stored and enriched with all PTMs present in dbPTM[70], downloaded on 2023-02-01]. PTMs found in our UniProt primary accession code selection of the human genome were selected and stored with the PTM type. This resulted in a list of 584,236 PTMs.

Dataset construction

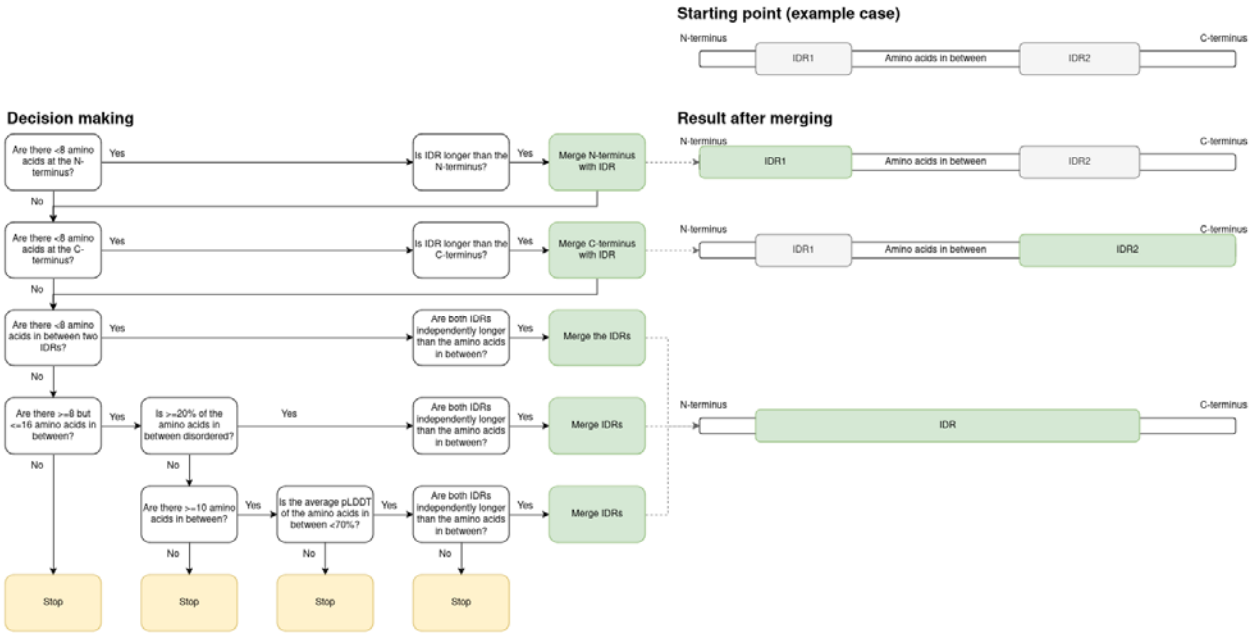
To define IDRs in the found proteins, IDRs annotated by MobiDB were mined and stored with their start and end position. To avoid two IDRs with no or small sequence distance, the IDRs annotated by MobiDB were re-addressed based on their length, the number of amino acids in between two IDRs, and the DSSP annotation of the secondary structure that AlphaFold2 predicts for the IDR regions. The predicted local distance difference test (pLDDT) score was used as a justification metric. Namely for an ordered region in between two IDRs with high pLDDT (>70%), it is most likely an ordered region[26,68]. On the other hand, if pLDDT of these

residues is low (<70%), it is most likely a disordered region[26,68], so we can merge the two IDRs with the amino acid in between them into a single IDR. Based on the average length of unmodeled regions in PDB-REDO structure models, we decided to keep a minimal IDR length of 8 amino acids. We also set this requirement, because 8 amino acids could form a two-turn helix, which we consider the smallest possible secondary structure element.

The decision making process is as follows:

1. If there are equal to or less than 8 amino acids present between two IDRs, the two IDRs are merged with the amino acids in between into one single IDR when both IDRs are independently longer or equally long as the number of amino acids in between them.
2. If there are more than 8 up to including 16 amino acids between two IDRs, those IDRs will be merged with the amino acids in between into one single IDR when there are less than 80% ordered residues annotated by DSSP for the region in between *and* both IDRs are independently longer or equally long as the number of amino acids in between them. The two IDRs will also be merged when AlphaFold predicted more than 80% ordered residues being present in the region between the two IDRs, *and* this 80% holds at least 8 amino acids, *and* the average pLDDT value of the region between the two IDRs is below 70%. Again, merging only happens when the region in between and both IDRs are independently longer or equally long as the number of amino acids in between them.
3. The same logic as described in 1 and 2 is applied for the N-terminus to the first IDR in the protein, and between the end of the last IDR in a protein to the C-terminus.
4. If an IDR is shorter than 8 amino acids in length, it is removed from the dataset.

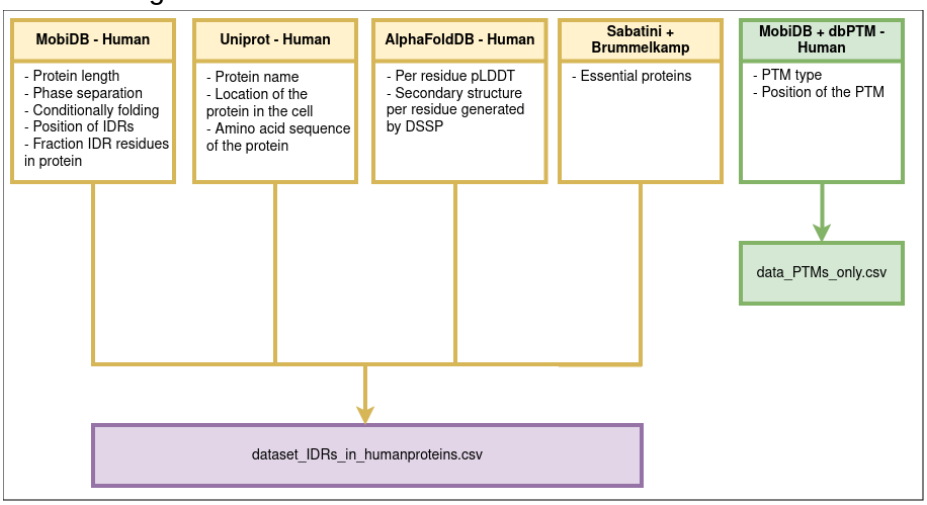
If above criteria are not met, IDRs will not be merged nor deleted and kept as annotated by MobiDB. This process is schematically described in Methods Figure 1.



Methods Figure 1: Decision tree for extending and merging IDRs.

IDR dataset

For each of the unique UniProt primary accession codes in the dataset, several parameters and characteristics of the human proteins were mined or calculated. A summary can be found in Methods Figure 2.



Methods Figure 2: Summary of the obtained data for human proteins, their IDRs and characteristics.

Protein characteristics - length, IDR presence and cellular localization

On the protein level, for each UniProt primary accession code we state the protein length, the name of the protein as annotated in UniProt, the number of IDRs in the protein and the number of IDR residues. We calculated the IDR fraction per protein by dividing the number of IDR residues in a protein by the protein length. Additionally we defined the cellular localization of each protein based on the annotation found in UniProt. We define the categories:

- Nucleus: proteins that localize at the nucleus.
- Organelles: proteins that localize to one of the following organelles or highly specialized cell compartments: mitochondria, endoplasmic reticulum, endosome, golgi, autophagosome, phagosome, peroxisome, melanosome, lysosome or microsome.
- Multiple organelles: Proteins that localize to more than one of the organelles.
- Cortex: proteins that localize to places in the cell that are involved in sensing the environment, but do not localize to organelles. This category entails locations such as gap junctions, cilia, protrusions the cell cortex, collectively called interactions.
- Plasma membrane: transmembrane proteins and membrane-anchored proteins. Note that this category does not include proteins in the membranes of organelles, those are classified into the organelles.
- Secreted: Proteins that are extracellular.
- Cytoplasm: Proteins that are annotated to localize in the cytoplasm *and* do not fall in any of the above categories.
- Unknown: proteins without any known cellular location annotation.
- Miscellaneous: proteins that have a cellular location, but do not belong to any of the above categories (15 proteins). These proteins are not shown in further analyses.

We mark the proteins involved in phase separation based on the in MobiDB integrated annotations for this feature, which are originally retrieved from PhaSepDB[71] and PhaSePro[72]. In order to determine if a protein is associated with an essential gene, the genes marked as essential in one or more cell-lines by Blomen *et al.*[40] or Wang *et al.*[41] were mapped to corresponding UniProt identifiers using the UniProt IDmapping tool [https://www.uniprot.org/id-mapping].

We use MobiDB-annotated binding modes (originating from FuzDB[73] and FuzPred[74] and distinguish (1) interacting proteins that remain disordered in their bound state (disorder-to-disorder), and (2) proteins that fold in their bound state (disorder-to-order). We also find 9 proteins that have a binding mode annotated that depends on the context, which we do not consider as a separate binding mode category.

Pairwise Alignment of the protein sequences of the different source databases (Uniprot, AlphaFoldDB, dbPTM) were mapped to our main source database (MobiDB) to ensure consistency while comparing amino acid positions and types. To score the number of mismatches and gaps in protein alignments, the BLOSUM62 substitution matrix was used in combination with default scoring settings as in BlastP[75] .

IDR and non-IDR characteristics

For each IDR and non-IDR region we annotated the location of the region in the protein based on their amino acid positions. Regions starting at the first residue of the protein are annotated as N-terminal (Nterm) regions, regions that end at the last residue of the protein are marked as C-terminal (Cterm) regions. Regions that start and end at any location that is not the first or last residue of the protein, are marked as internal (Intra) regions. Proteins that contain no IDR residues are marked as ordered protein (OP) and proteins solely containing IDR residues are annotated as IDPs.

For each IDR and non-IDR region of the proteins in our dataset, we calculated the average pLDDT score based on the predicted protein structures by AlphaFold2. Additionally, we define the fraction of ordered residues in the region. For this purpose, we characterized the amino acids into five main categories: cationic (R,H,K), anionic (D,E), polar (S,T,N,Q), hydrophobic (A,V,I,L,M,F,Y,W) and special (C,P,G) (Supplemental Table 3). For each region, we counted how many residues occur in each of the categories and calculated their relative percentage by dividing this count by the total of amino acids in the region.

Final dataset

The procedures above resulted in a final IDR dataset containing 20,038 proteins with 40,756 IDRs in total. In this set 1046 (5%) UniProt primary accession codes are annotated to be IDP (100% disordered), 16,707 (83%) have at least one IDR, and 1046 (11%) proteins have no IDRs annotated.

Analyses

The datasets were loaded using Pandas 2.0.1 [https://zenodo.org/records/10957263] and plots were generated using Seaborn 0.13.2, all using Python 3.11.3. Box-, violin- and kde-(correlation) plots were created with a minimum of 30 data points[76], if the desired sub-dataset was smaller, a stripplot was made to represent the data more realistically. Mean and median values are calculated for datasets containing at least 5 observations[76].

Used software and packages

BioPython 1.81 [77]
DSSP 4.3 [30]
Numpy 1.24.3 [78]
Pandas 2.0.1 [79]
Pdbecif 1.5 [80]
Python 3.11.3
Seaborn 0.13.2 [81]

Contributions

Conceptualization: JB IdV DAS RB RX MK

Methodology: IdV DAS JB RB RX MK

Formal analysis: JB IdV DAS

Investigation: JB IdV DAS

Resources: AP, RHPC facility NKI

Data Curation: IdV DAS

Writing - Original Draft: JB & IdV (all), DAS (methods)

Writing - Review & Editing: DAS, RB, BE, MU, MK, RX, AP

Visualization: JB & IdV

Supervision: -

Funding acquisition: AP

Acknowledgements

We would like to express our gratitude to our department's freedom, scientific interest and thoughtful comments, that all contributed to the shaping of this project. Specifically we have had fruitful discussions on diverse implications of this study with Robbie P. Joosten, Hans Wienk, Maarten L. Hekkelman, Shun Lee, Felipe R. Rodriguez, Gian-Luca McLelland and Tim Schröder. Specifically we like to thank Boğaç Erçığ and Michael Uckelmann for reviewing the text. We thank the Research High Performance Computing facility of the Netherlands Cancer Institute for providing and maintaining computation resources.

Funding

This work has not been formally supported by dedicated funding; it has been made possible by the freedom offered by the institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport to the NKI and by the team of AP being supported by then Oncode Institute. Authors of this grant have been employed over the course of this research by different sources, including Oncode Accelerator, NWO ENW (OCENW.M20.324), the Horizon2020 EC projects iNEXT-Discovery (871037), FragmentScreen (101094131) and ESPERANCE (101087215).

References

1. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
2. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148 (2005).
3. Basile, W., Salvatore, M., Bassot, C. & Elofsson, A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput. Biol.* **15**, e1007186 (2019).
4. Babu, M. M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185–1200 (2016).
5. Pinet, L., Assrir, N. & van Heijenoort, C. Expanding the Disorder-Function Paradigm in the C-Terminal Tails of Erbbs. *Biomolecules* **11**, 1690 (2021).
6. Alowolodu, O. *et al.* Intrinsic disorder in spondins and some of their interacting partners. *Intrinsically Disord. Proteins* **4**, e1255295 (2016).
7. Bhattacharya, S. *et al.* The disordered regions of the methyltransferase SETD2 govern its function by regulating its proteolysis and phase separation. *J. Biol. Chem.* **297**, 101075 (2021).
8. Chaves-Arquero, B. *et al.* eIF4G1 N-terminal intrinsically disordered domain is a multi-docking station for RNA, Pab1, Pub1, and self-assembly. *Front. Mol. Biosci.* **9**, 986121 (2022).
9. Di Bartolo, A. L., Caparotta, M. & Masone, D. Intrinsic Disorder in α-Synuclein Regulates the Exocytotic Fusion Pore Transition. *ACS Chem. Neurosci.* **14**, 2049–2059 (2023).
10. Dyson, H. J. Vital for Viruses: Intrinsically Disordered Proteins. *J. Mol. Biol.* **435**, 167860 (2023).

11. El Hadidy & Uversky. Intrinsic Disorder of the BAF Complex: Roles in Chromatin Remodeling and Disease Development. *Int. J. Mol. Sci.* **20**, 5260 (2019).

12. Brooks-Bartlett, J. C. & Garman, E. F. The Nobel Science: One Hundred Years of Crystallography. *Interdiscip. Sci. Rev.* **40**, 244–264 (2015).

13. Klyshko, E. *et al.* Functional protein dynamics in a crystal. *Nat. Commun.* **15**, 3244 (2024).

14. Fukuda, Y., Stapleton, K. & Kato, T. Progress in spatial resolution of structural analysis by cryo-EM. *Microscopy* **72**, 135–143 (2023).

15. Kühlbrandt, W. The Resolution Revolution. *Science* **343**, 1443–1444 (2014).

16. Kirnarsky, L., Shats, O. & Sherman, S. Improving the efficiency of protein structure determination from NMR. *J. Mol. Struct. THEOCHEM* **419**, 213–220 (1997).

17. Markley, J. L. & Westler, W. M. Biomolecular NMR: Past and future. *Arch. Biochem. Biophys.* **628**, 3–16 (2017).

18. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).

19. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Struct. Lond. Engl. 1993* **11**, 1453–1459 (2003).

20. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinforma. Oxf. Engl.* **21**, 3433–3434 (2005).

21. Murachelli, A. G., Damaskos, G. & Perrakis, A. CCD2: design constructs for protein expression, the easy way. *Acta Crystallogr. Sect. Struct. Biol.* **77**, 992–1000 (2021).

22. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

23. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

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

25. Piovesan, D., Monzon, A. M. & Tosatto, S. C. E. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci. Publ. Protein Soc.* **31**, e4466 (2022).

26. Alderson, T. R., Pritišanac, I., Kolarić, Đ., Moses, A. M. & Forman-Kay, J. D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc. Natl. Acad. Sci.* **120**, e2304302120 (2023).

27. Piovesan, D. *et al.* MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res.* **51**, D438–D444 (2022).

28. Necci, M., Piovesan, D. & Tosatto, S. C. E. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein Sci. Publ. Protein Soc.* **25**, 2164–2174 (2016).

29. Tesei, G. *et al.* Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).

30. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411–D419 (2011).

31. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

32. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

33. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016).

34. Chou, K.-C. & Elrod, D. W. Protein subcellular location prediction. *Protein Eng. Des. Sel.* **12**, 107–118 (1999).

35. Pan, X. *et al.* Identification of Protein Subcellular Localization With Network and

Functional Embeddings. *Front. Genet.* **11**, (2021).

36. Fakhree, M. A. A., Blum, C. & Claessens, M. M. A. E. Shaping membranes with disordered proteins. *Arch. Biochem. Biophys.* **677**, 108163 (2019).

37. Has, C., Sivadas, P. & Das, S. L. Insights into Membrane Curvature Sensing and Membrane Remodeling by Intrinsically Disordered Proteins and Protein Regions. *J. Membr. Biol.* **255**, 237–259 (2022).

38. Kjaergaard, M. & Kragelund, B. B. Functions of intrinsic disorder in transmembrane proteins. *Cell. Mol. Life Sci.* **74**, 3205–3224 (2017).

39. Liaci, A. M. *et al.* Structure of the human signal peptidase complex reveals the determinants for signal peptide cleavage. *Mol. Cell* **81**, 3934-3948.e11 (2021).

40. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

41. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).

42. Tompa, P., Dosztanyi, Z. & Simon, I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J. Proteome Res.* **5**, 1996–2000 (2006).

43. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).

44. Ibrahim, A. Y. *et al.* Intrinsically disordered regions that drive phase separation form a robustly distinct protein class. *J. Biol. Chem.* **299**, (2023).

45. Radivojac, P. *et al.* Intrinsic disorder and functional proteomics. *Biophys. J.* **92**, 1439–1456 (2007).

46. Dunker, A. K. *et al.* Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).

47. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).

48. Williams, R. M. *et al.* The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 89–100 (2001).

49. Liaci, A. M. & Förster, F. Take Me Home, Protein Roads: Structural Insights into Signal Peptide Interactions during ER Translocation. *Int. J. Mol. Sci.* **22**, 11871 (2021).

50. Duan, G. & Walther, D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLoS Comput. Biol.* **11**, e1004049 (2015).

51. Lee, J. M., Hammarén, H. M., Savitski, M. M. & Baek, S. H. Control of protein stability by post-translational modifications. *Nat. Commun.* **14**, 201 (2023).

52. Pejaver, V. *et al.* The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **23**, 1077–1093 (2014).

53. Jiang, G., Li, C., Lu, M., Lu, K. & Li, H. Protein lysine crotonylation: past, present, perspective. *Cell Death Dis.* **12**, 1–11 (2021).

54. Bochicchio, B., Laurita, A., Heinz, A., Schmelzer, C. E. H. & Pepe, A. Investigating the Role of (2S,4R)-4-Hydroxyproline in Elastin Model Peptides. *Biomacromolecules* **14**, 4278–4288 (2013).

55. Rappu, P., Salo, A. M., Myllyharju, J. & Heino, J. Role of prolyl hydroxylation in the molecular interactions of collagens. *Essays Biochem.* **63**, 325–335 (2019).

56. Singh, A. K., Amar, I., Ramadasan, H., Kappagantula, K. S. & Chavali, S. Proteins with amino acid repeats constitute a rapidly evolvable and human-specific essentialome. *Cell Rep.* **42**, (2023).

57. van der Lee, R. *et al.* Intrinsically Disordered Segments Affect Protein Half-Life in the Cell and during Evolution. *Cell Rep.* **8**, 1832–1844 (2014).

58. Whitley, P., Grau, B., Gumbart, J. C., Martínez-Gil, L. & Mingarro, I. Folding and Insertion of Transmembrane Helices at the ER. *Int. J. Mol. Sci.* **22**, 12778 (2021).

59. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

60. Chui, A. J. *et al.* Multiple structural states exist throughout the helical nucleation sequence of the intrinsically disordered protein stathmin, as reported by electron paramagnetic resonance spectroscopy. *Biochemistry* **54**, 1717–1728 (2015).

61. Bondos, S. E., Dunker, A. K. & Uversky, V. N. On the roles of intrinsically disordered proteins and regions in cell communication and signaling. *Cell Commun. Signal.* **19**, 88 (2021).

62. Kapp, K., Schrempf, S., Lemberg, M. K. & Dobberstein, B. Post-Targeting Functions of Signal Peptides. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).

63. Imai, K. & Nakai, K. Tools for the Recognition of Sorting Signals and the Prediction of Subcellular Localization of Proteins From Their Amino Acid Sequences. *Front. Genet.* **11**, 607812 (2020).

64. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

65. Choi, U. B., Sanabria, H., Smirnova, T., Bowen, M. E. & Weninger, K. R. Spontaneous Switching among Conformational Ensembles in Intrinsically Disordered Proteins. *Biomolecules* **9**, 114 (2019).

66. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026 (2004).

67. Borcherds, W., Bremer, A., Borgia, M. B. & Mittag, T. How do intrinsically disordered protein regions encode a driving force for liquid–liquid phase separation? *Curr. Opin. Struct. Biol.* **67**, 41–50 (2021).

68. Piovesan, D. *et al.* MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res.* **51**, D438–D444 (2023).

69. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic*

Acids Res. **51**, D523–D531 (2023).

70. Li, Z. *et al.* dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.* **50**, D471–D479 (2022).

71. Hou, C. *et al.* PhaSepDB in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res.* **51**, D460–D465 (2023).

72. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).

73. Hatos, A., Monzon, A. M., Tosatto, S. C. E., Piovesan, D. & Fuxreiter, M. FuzDB: a new phase in understanding fuzzy interactions. *Nucleic Acids Res.* **50**, D509–D517 (2021).

74. Hatos, A. *et al.* FuzPred: a web server for the sequence-based prediction of the context-dependent binding modes of proteins. *Nucleic Acids Res.* **51**, W198–W206 (2023).

75. Park, Y., Sheetlin, S., Ma, N., Madden, T. L. & Spouge, J. L. New finite-size correction for local alignment score distributions. *BMC Res. Notes* **5**, 286 (2012).

76. Krzywinski, M. & Altman, N. Visualizing samples with box plots. *Nat. Methods* **11**, 119–120 (2014).

77. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

78. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

79. The pandas development team. Pandas 2.0.1.

80. van Ginkel, G. *et al.* PDBeCIF: an open-source mmCIF/CIF parsing and processing package. *BMC Bioinformatics* **22**, 383 (2021).

81. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).