JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
2021, VOL. 39, NO. 8, 2885–2893
https://doi.org/10.1080/07391102.2020.1756410

Unique and exclusive peptide signatures directly identify intrinsically disordered proteins from sequences without structural information

Aditya Mittal[a,b], Anandkumar Madhavjibhai Changani[a] and Sakshi Taparia[c]

[a]Kusuma School of Biological Sciences, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India; [b]Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India; [c]Department of Mathematics (Bachelors Program in Mathematics & Computing), Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

Intrinsically disordered proteins are now widely accepted to play crucial roles in biological functions. Identification of signatures of intrinsic disorder is one of the key steps towards building a proper repertoire for their occurrence in proteomes. In this work, systematic computational synthesis of a library of all possible (3368400) dipeptides, tripeptides, tetrapeptides and pentapeptides using the natural 20 amino acids allowed us to identify 36 unique tetrapeptides present exclusively in intrinsically disordered proteins and absent in the complete primary sequence space of naturally occurring structured proteins. Further, out of more than 530000 known naturally occurring primary sequences without any structural information, 1349 sequences contain the above identified unique signatures of intrinsic disorder. These sequences, having cellular functions varying from housekeeping to metabolic to transport, more than double the number of the currently known intrinsically disordered proteins. On similar lines, we report that 26577 pentapeptide signatures exclusive to intrinsically disordered proteins, and absent in naturally occurring structured proteins, identify ∼50% of more than half-a-million curated protein sequences without structural information to be intrinsically disordered. The results reported are a major leap forward in exploring functional manifestations of intrinsically disordered proteins.

ARTICLE HISTORY
Received 19 March 2020
Accepted 13 April 2020

KEYWORDS
Intrinsically disordered proteins; IDPs; peptide; proteins; sequence

Introduction

The concept of sequence-to-structure-to-function continues to dominate the protein folding problem as a grand challenge. However since the turn of the current century, some pioneering observations on existence of "hybrid proteins" (sequences that only partly formed structured domains with other parts resulting in intrinsically disordered protein regions), with some biological activities, have led to an increase in the interest on protein-disorder-related research (Dunker et al., 2001, 2002; Dyson & Wright, 2005; Tompa, 2002; Uversky, 2002; Uversky et al., 2000; Wright & Dyson, 1999). Some of the earlier work provided indications for existence of differences between amino acid compositions, as well as their arrangements in primary sequences, of ordered proteins/domains and "natively unfolded" proteins/regions. For example, Uversky et al. (2000) pointed out that "natively unfolded proteins" may possess high content of similarly charged residues and low content of hydrophobic residues. In a comprehensive review, Dunker et al. (2001) proposed that residues can be grouped into order- and disorder- promoting categories based on comparative analysis of primary sequences. These observations led to development of "*exploratory data mining*" tools towards building "*rough, light-weight visual classifiers*" during comparative profiling of primary sequences (Vacic et al., 2007). In fact, arguably fuelled by the above studies laying the foundations for exploring the possibilities of specific biases in sequences of ordered and disordered proteins/regions, investigations on presence of protein sequences that do not fold into stable structures and yet play key functional roles in biology have become substantially significant over the last decade (Babu et al., 2012; Chouard, 2011; Jensen & Blackledge, 2014; Ozenne et al., 2012; Uversky, 2019; Zhang et al., 2018). These protein sequences, if present as localized disordered-regions within folded proteins are called intrinsically disordered regions (IDRs). Alternatively, complete protein sequences resulting in flexible conformers without a stable structure are called Intrinsically Disordered Proteins (IDPs). While the efforts towards solving "the protein folding problem" (i.e. obtaining stable structures from primary sequences) have dominated the literature over more than half-a-century (Mittal et al., 2010 and references therein), the last decade has seen a major thrust towards appreciating the importance of IDPs and proteins with IDRs (Davey et al., 2012; Dinkel et al., 2014; Gouw et al., 2017, 2018; van der Lee et al., 2014). As of today, over 3000 IDRs and ∼1500 IDPs have been identified after careful curation of literature, and based on experimental data, in the database called DisProt (Hatos et al.,

CONTACT Aditya Mittal amittal@bioschool.iitd.ac.in Kusuma School of Biological Sciences, Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07391102.2020.1756410.

2020; Piovesan et al., 2017). While these appear to be a miniscule fraction of total un-reviewed (over 177,750,000) primary sequences and even manually annotated & reviewed (~562000) primary sequences (The UniProt Consortium, 2019), there is substantial appreciation for the continuous increase observed in the number of IDPs and IDRs being identified in proteomes (Chouard, 2011; Ozenne et al., 2012; Zarin et al., 2019). This is primarily fuelled by the various roles (direct and indirect) of IDPs that have been demonstrated not just in crucial cellular and biological functions (Berlow et al., 2018; Communie et al., 2014; Oldfield & Dunker, 2014; Parigi et al., 2014; van der Lee et al., 2014; Zhang et al., 2018), but also in pathologies resulting from changes specifically in IDPs (Li & Babu, 2018; Meyer et al., 2018).

Identification of IDPs gained substantial momentum with the growing interest in their functional mediation, along with that of IDRs, attributable to sequence-disorder relationships (Babu et al., 2012). A modular view on protein sequences in genomes, with combinations of primary sequences resulting in a mix of structured and disordered ensembles, has started to emerge via a classification of short linear motifs (SLiMs, or eukaryotic linear motifs ELMs) over the years (Davey et al., 2012; Dinkel et al., 2014; Gouw et al., 2017, 2018; van der Lee et al., 2014). Additionally, efforts have also been directed towards identifying short contiguous peptide sequences having functional manifestations in context of whole proteins (Mi et al., 2012) along with search for sequence determinants of intrinsic disorder (Martin et al., 2016; Ota & Fukuchi, 2017). Beyond sequences, IDPs are also viewed in terms of collections of conformations ("ensemble descriptions") that can result in important functionalities resulting from flexible switching between conformers (Estaña et al., 2019; Jensen & Blackledge, 2014; Salvi et al., 2019; Uversky, 2019). The above efforts, while appreciating the divergence in the primary sequence space of IDPs, have largely been directed towards extracting sequence-dependent consensus features. However, these consensus-based approaches, instead of certainty of universal applicability, result in a case-by-case basis of functional understanding and inclusion into the world of IDPs. In addition, a multitude of previous studies, with careful assembling of datasets of ordered and disordered proteins, have been dedicated to the development of computational tools for predicting intrinsic disorder and functional IDRs (reviewed in Katuwawala et al., 2019) – these studies obtained composition and sequence based sets of rules aimed at minimizing misclassification of proteins; again these are limited by case-specific rather than universal applicability. Thus, a suggestion of more than a million instances of peptide motifs in the human proteome in the context of IDPs (Tompa et al., 2014) appears to be proving almost prophetic. More recently, a comprehensive assessment over 60 sequence-based disorder prediction methods highlighted the fact that in spite of these methods being theoretically applicable to datasets of proteins, the predictions are limited to individual proteins only, with a case-by-case assessment that is not universally applicable (Katuwawala et al., 2019). Considering the limitations of methods based on differences in amino acid compositions and primary sequences between ordered proteins/domains and IDPs/IDRs found based on rigorous and thorough analyses, Katuwawala et al. (2019) concluded "*Thus novel tools that accurately identify the hard-to-predict proteins and that make accurate predictions for these proteins are needed.*"

Therefore, in this work, we apply an approach completely different from the earlier efforts, to extract peptide signatures of IDPs – note that our approach focuses primarily on IDPs and not IDRs; thus while we extract peptide signatures of IDPs, we do not necessarily assume that these peptides are IDRs by themselves. In our approach, by computationally synthesizing a library of all possible peptides up to pentapeptides, using the standard 20 amino acids, we identified peptides that did not occur, i.e. were absent in primary sequences of all known structured/folded proteins, by counting occurrences of every possible peptide in the primary sequences of all known structured/folded proteins. Next, we specifically searched for the above absent peptides for their presence in primary sequences of IDPs – again, we considered only those sequences that are classified as IDPs while ignoring sequences only classified as IDRs (these contain IDRs and may be partially structured proteins) but not classified as IDPs. In spite of the number of primary sequences of IDPs being ~ 5% of the number of primary sequences of structured/folded proteins, we report the remarkable discovery of 36 unique tetrapeptides and 26577 pentapeptides that are exclusively present only in IDPs. Finally, we found that these 36 unique tetrapeptides are present only in about 1349 sequences out of over 532000 naturally occurring primary sequences without any structural information. Identification of these 1349 new sequences (since they contain the unique signatures of IDPs) almost doubles the number of known IDPs. Even more remarkably, we found that the 26577 unique pentapeptides specifically present only in IDPs, but absent in sequences with structural information, are also present in a staggering 265407 primary sequences without any structural information (out of the over 532000). This implies that almost 50% of the known protein sequences are structurally disordered proteins. While the whole of naturally occurring protein sequence space has been hypothetically expected to have over 35% IDPs (Chouard, 2011; Salvi et al., 2019), we provide the first direct computational evidence, completely independent of approaches searching for consensus sequences or specific motifs in IDPs, supporting this idea. Importantly, the new sequences are known to have cellular functions varying from housekeeping to metabolic to material transport. We are very hopeful that our findings will provide (a) a new dimension to research on importance of disordered sequences with the identified peptides serving as direct, unambiguous and universally applicable signatures of IDPs, and, (b) a leap forward in building the repertoire of functional manifestations of intrinsically disordered proteins in biology.

Methods

Complete sequence data was downloaded from Uniprot (Swiss-prot) and DisProt on 21 September 2019 as per instructions provided for offline analyses. Results presented are from this dataset. Two earlier downloads of complete sequence data on 15 February 2019 and 28 July 2019 were done to develop and test the analytical codes (Mittal et al., 2020). Negligible differences were observed in the overall results obtained in the three datasets – this provides evidence for the robustness of not only the Uniprot database but also the analytical algorithms developed in this work. Coding was done in Python for counting the number of occurrences of peptides. Independent coding was done in Java to confirm accuracy of the results up to the occurrences of tripeptides. The final data analyses were done in MATLAB (Mathworks, Inc.) and MS-Excel.

Results and discussion

Searching for peptides in the curated primary sequence space of natural proteins

In order to search for unique peptide signatures of IDPs, the first step was careful collection the available primary sequence data. The UniProtKB knowledgebase (The UniProt Consortium, 2019) has two primary databases – Swiss-Prot containing over 560823 curated (manually annotated and reviewed) sequences, and, TrEMBL containing over 177,750,000 un-reviewed sequences. To ensure reliability of investigations, we first downloaded all curated sequences available in the UniProtKB: Swiss-Prot database. Cross-referencing the downloaded sequences, we found that 26960 of the curated sequences in Swiss-Prot have well defined structures (∼104000 with varying resolutions) in the Protein Data Bank, PDB (Berman et al., 2000, 2007) – we called this dataset as "StrucSeq" (given in supplementary Table S1). We also found that over 1225 of the sequences in Swiss-Prot are listed in the independent database for IDPs called DisProt (Hatos et al., 2020; Piovesan et al., 2017) – we called this dataset as "IDPsSeq" (given in supplementary Table S2). Here, it is important to note that length of Swiss-Prot entries may exceed the length of corresponding PDB sequences since structural analyses are often based on truncated fragments (of varying sizes) of proteins due to experimental limitations. In fact, sequences may have IDRs which are removed during structural analyses. In addition, structures reported in PDB may have regions of missing electron density attributable to IDRs. Thus, it may be argued that the above may inflate the "StrucSeq" dataset by including potential IDRs. Thus, in order to avoid ambiguities related to IDRs while focussing only on IDPs, we purposefully classified all sequences in Swiss-Prot with entries corresponding to PDB as "StrucSeq" – in essence, any sequence with "partial" structure or complete structure is regarded as a member of "StrucSeq". In this way, we ensured focussing on IDPs only in comparison to structured proteins (even if they have IDRs). Now, it might be argued that such classification can introduce some analytical biases inclusion of possible "hybrid" Swiss-Prot entries into the "StrucSeq" may mask the pure order-specific sequence signatures, by "diluting" them with some disorder-specific signatures - however, this is not the case in this work since the approach developed here specifically explores disorder-specific signatures only without any emphasis on possible order-specific signatures (see below, next section onwards). The remaining (532638) curated sequences in Swiss-Prot were found to be without any structural information – we called this dataset as "OnlySeq" (due to the large file size, this dataset is provided as a .CSV file at https://web.iitd.ac.in/~amittal/Data_Mittal_etal_IDPs_JBSD.html). Thus, the complete Swiss-Prot dataset was classified into three mutually exclusive datasets called "OnlySeq", "SturcSeq" and "IDPsSeq" represented by light-grey, dark-grey and black regions respectively in Figure 1A. In addtion to the above, we also found 171 primary sequences in DisProt that do not belong to Swiss-Prot, but to TrEMBL in UniProtKB – we call this dataset of unreviewed sequences as "IDPsUnRev" (given in supplementary Table S3): it is represented by the black region outside the Swiss-Prot light-grey region in Figure 1A.

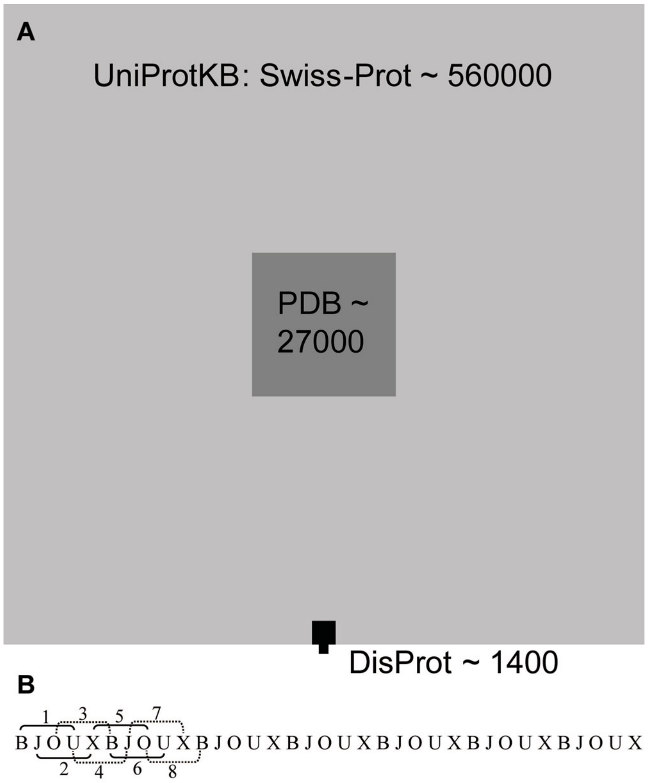
A
UniProtKB: Swiss-Prot ~ 560000
PDB ~ 27000
DisProt ~ 1400
B
1 3 5 7
B J O U X B J O U X B J O U X B J O U X B J O U X B J O U X B J O U X
2 4 6 8

Figure 1. Searching for peptides in the curated primary sequence space of natural proteins. (A) Complete primary sequence space of natural proteins, in terms of "manually annotated and reviewed" (curated) sequences is represented by the light grey square. Of these more than half-a-million curated sequences in UniProtKB (Swiss-Prot), over 27000 sequences (represented by the dark grey square) are known to have over one hundred thousand structures in the Protein Data Bank (PDB) – these sequences are listed in Table S1. Over 1225 of the curated sequences (represented by the black square within the light grey square) are known to be Intrinsically Disordered Proteins (IDPs) in DisProt – these sequences are listed in Table S2. DisProt also has an additional 171 sequences classified as IDPs but are not manually annotated and reviewed (i.e. are not in the UniprotKB: Swiss-Prot database; hence represented by the small black square outside the light grey square) – these sequences are listed in Table S3. In terms of covered area by each class of sequences, the figure is drawn to scale. (**B**) Counting peptides in primary sequences – example is shown for searching tetrapeptides in a given sequence using four reading frames. The first 8 tetrapeptide reads are shown (1-BJOU, 2-JOUX, 3-OUXB, 4-UXBJ, 5-XBJO, 6-BJOU, 7-JOUX, 8-OUXB; thus in the first 8 reads, BJOU, JOUX and OUXB occur twice each). Non-standard amino acid letters, i.e. BJOUXZ, were not counted for compiling the number of times a given peptide occurs in a given sequence – the total occurrence of non-standard amino acid letters was negligible ($\ll$ 0.5%) in each dataset.

Table 1. Counting peptides absent in sequences from UniProt (Swiss-Prot) and DisProt.

		Curated sequences (UniProt: Swiss-Prot)			Non-curated
	Total number	OnlySeq	StrucSeq	IDPsSeq	IDPsUnRev
Sequences	560823 + (171)$^{\$}$	532638*	26960**	1225***	171****
Dipeptide	400	–	–	–	–
Tripeptide	8000	–	–	06	3267
Tetrapeptide	160000	02	493	32177	148360
Pentapeptide	3200000	90515	874725	2708458	3186870

$^{\$}$Sequences which are listed as IDPs in Disprot but are not a part of Swiss-Prot and are found as unreviewed sequences in TrEMBL.

*Listed at http://web.iitd.ac.in/~amittal/Data_Mittal_etal_IDPs_JBSD.html.

**Listed in Table S1.

***Listed in Table S2.

****Listed in Table S3.

Table 2. Number of unique peptides exclusively present in IDPs only.

	Absent in StrucSeq	Absent in StrucSeq but Present in IDPsSeq	Absent in StrucSeq but Present in IDPsUnRev
Tetrapeptide	493	**36***	0
Tetrapeptide	874725	**26577****	513***

*Listed in Table 3.

**Listed in Table S4.

***Listed in Table S5.

Table 3. Unique tetrapeptide signatures exclusive to IDPs.

GYWC	CCWW	CWWH	MWCH	RIWW	WQWH
SWMC	CNNW	IMFW	MWQC	WPMQ	WKWW
PYWC	CQHW	NNWC	HCMW	WPMM	WECF
VMCW	CKCW	NWFW	HIWW	WCII	WMGH
TKMW	CMHW	MCTW	HWTF	WCMQ	WYWP
THMW	CMWW	MDQW	HWTW	WLMM	WWFS

Now using the standard 20 amino acids, we computationally synthesized libraries of dipeptides (20 × 20 = 400), tripeptides (20x20x20 = 8000), tetrapeptides (20x20x20x20 = 160000) and pentapeptides (20x20x20x20x20 = 3200000). Thus, the complete library contained 3368400 peptides (available in Mittal et al., 2019). The next step was to count the number of occurrences of each of these peptides in the four mutually exclusive primary sequence sets (OnlySeq, StrucSeq, IDPsSeq and IDPsUnRev). Occurrence of a given peptide was counted by reading each sequence from the amino to carboxy termini as shown in Figure 1B. The number of reading frames for each sequence was equal to the size of the peptide. Thus, e.g. to count the number of times a given tetrapeptide occurs in a given sequence, four reading frames starting from the N-terminal amino acid were utilized as shown in Figure 1B. While sophisticated/comprehensive statistical analyses of the number of occurrences of different peptides in the four mutually exclusive sequence datasets is very appealing (and will be pursued in future), here we decided to focus specifically only on absolute inferences that could be extracted.

Therefore, we simply counted how many peptide sequences did not occur even once in each of the datasets (i.e. identification of "zero-occurrence" peptides). Table 1 shows (i) all 400 dipeptides were present in all datasets, (ii) all tripeptides were present in "OnlySeq" and "StrucSeq", but 06 & 3267 tripeptides were absent in "IDPsSeq" & "IDPsUnRev" respectively, (iii) 02, 493, 32177 & 148360 tetrapeptides were absent in "OnlySeq", "StrucSeq", "IDPsSeq" & "IDPsUnRev" respectively, and, (iv) 90515, 874725, 2708458 & 3186870 pentapeptides were absent in "OnlySeq", "StrucSeq", "IDPsSeq" & "IDPsUnRev" respectively.

The increasing number of absent peptides with decrease in the size of sequence dataset was obviously as expected. However, the key question was that in spite of total IDPs being a very small fraction of all sequence data (e.g. IDPs are ~5% of "StrucSeq" and only ~0.15% of "OnlySeq"), were there peptides that were absent in sequences with structures but present in intrinsically disordered proteins? To answer the above question, we specifically collected the number of peptides that were absent in "StrucSeq" to check their presence in "IDPsSeq" and "IDPsUnRev" as shown in Table 2.

Remarkably, of the 493 tetrapeptides and 874725 pentapeptides absent in "StrucSeq", 36 tetrapeptides and 26577 pentapeptides were found to be present in IDPs. In spite of number of IDPs being miniscule fraction of number of sequences with structures, identification of these unique peptides exclusive to only IDPs provides a first-of-its-kind sequence signature set for IDPs.

Unique tetrapeptide signatures exclusively present in IDPs

Table 3 lists all unique tetrapeptides exclusively present only in IDPs (i.e. absent in StrucSeq and present in IDPs).

Two interesting observations emerge from these tetrapeptides – (i) The residue A (i.e. alanine) is missing, i.e. compositionally these tetrapeptides exclusive to IDPs utilize only 19 of the 20 amino acids: Why this is so remains unanswered as of now – however it is interesting to note that alanine scanning/replacement by alanine is a routine protocol followed in molecular biology for functional studies of protein sequences; and, (ii) all of the unique tetrapeptides have at least one tryptophan (W). Thus, all reviewed sequences classified as IDPs have at least one "W" – this may have important functional implications. The conformational flexibility rendered by tryptophan residues especially at the aqueous and non-aqueous interfaces does provide an appealing functional feature of these tetrapeptide signatures of IDPs – this is supported

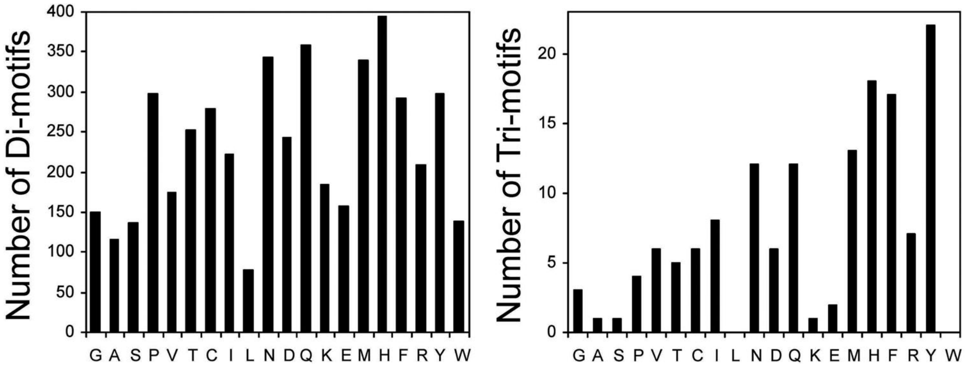

Figure 2. Homo-motifs in exclusive pentapeptide signatures of IDPs. Number of unique pentapeptides that serve as exclusive signatures of IDPs with Di-Motifs (i.e. with "XX" at any of the positions; left panel) and Tri-Motifs (i.e. with "XXX" at any of the positions; right panel). Peptides containing Di- and Tri- Motifs are listed in Table S7 and Table S8 respectively.

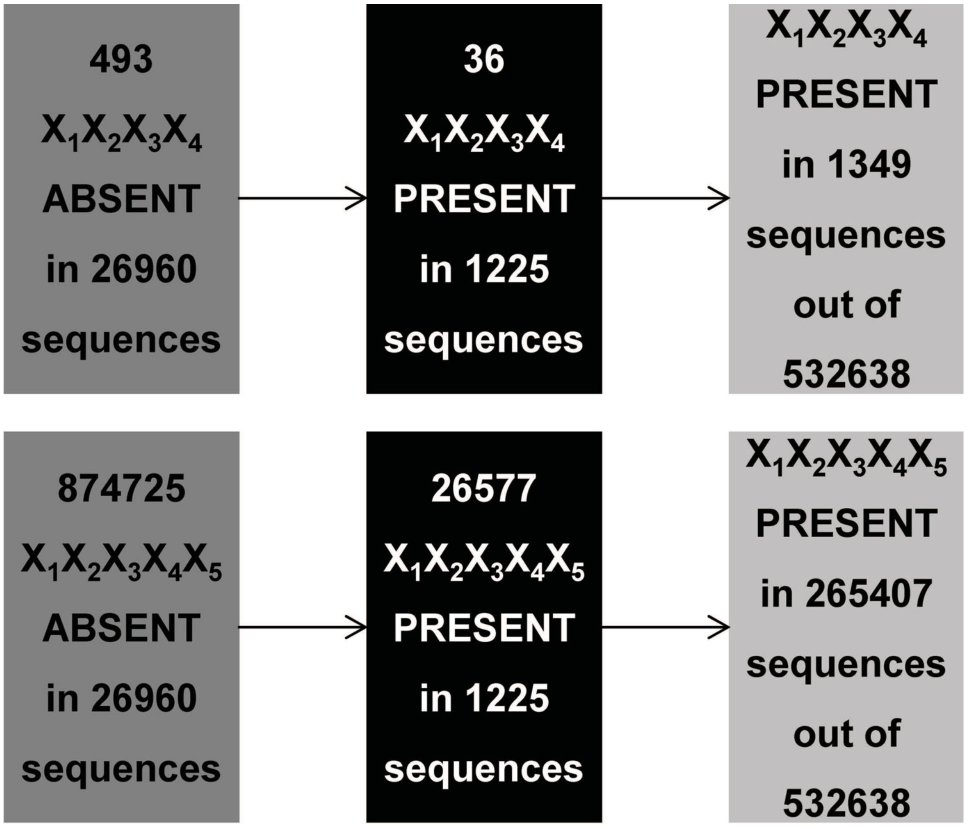

Figure 3. Unique and exclusive signatures identify new IDPs from sequences without structural information. Out of the 493 tetrapeptides absent in sequences of structured proteins (represented by dark grey box), 36 were present in IDPs (represented by black box) – these 36 peptides were present in only 1349 out of 532638 sequences without structural information (represented by light grey box). The 1349 newly identified IDPs are listed in Table S6. Similarly, out of the 874725 pentapeptides absent in sequences of structured proteins (represented by dark grey box), 26577 were present in IDPs (represented by black box) – these 26577 peptides were present in a staggering 265407 out of 532638 sequences without structural information (represented by light grey box). Due to the large size of the dataset, the 265407 newly identified IDPs are listed at http://web.iitd.ac.in/~amittal/Data_Mittal_etal_IDPs_JBSD.html.

by the recent findings on solvent-dependent dynamics of IDPs (Salvi et al., 2019).

Here, it is pertinent to mention that earlier work has attempted to classify amino acid residues into two groups – order-promoting and disorder-promoting, based on their abundance in structured proteins and IDPs/IDRs (Uversky, 2013). For example, W and C, in spite of being classified by Uversky (2013) as having lowest propensities for disorder, are dominantly present in Table 3. Similarly one may note that all unique tetrapeptides in Table 3 contain at least one order-promoting residue; many of these tetrapeptides contain several order-promoting residues; and a few tetrapeptides are entirely composed of order-promoting residues. Thus, it may be argued that these peptides are short regions with high order propensity (i.e. sequence stretches containing order-promoting residues), located within the longer disordered regions or complete IDPs. They may serve as recognition motifs for interaction with specific partners which undergo disorder-to-order transition as a result of the partner binding. Such disorder-based binding regions are known as molecular recognition elements/features - "MoREs"/"MoRFs" (Bourhis et al., 2004; Cheng et al., 2007;

Cukier, 2018; Mohan et al., 2006; Oldfield et al., 2005; Vacic et al., 2007). If this is the case, then the complete absence of these peptides from StrucSeq is indeed intriguing. However, it is also important to consider that the classification of amino acids into order-promoting and disorder-promoting groups may have serious statistical limitations and possible errors. This is due to the fact that relative abundance of amino acids in structured proteins and IDPs/IDRs must be viewed not only from the perspective of their mean occurrences in sequences but also from the perspective of the respective standard deviations of the mean occurrences (Mittal et al., 2010, 2020; Mittal & Jayaram, 2011a, 2011b).

Unique pentapeptide signatures exclusively present in IDPs

On similar lines as above, supplementary Table S4 lists all unique pentapeptides exclusively present only in IDPs (i.e. absent in StrucSeq and present in IDPs). Here it is important to note that at this point we focus only on the pentapeptides listed in Table S4 and not those listed in Table S5 as exclusive pentapeptide signatures for IDPs – this is so because Table S5 is based on non-curated protein sequences only. It was important to explore possible physico-chemical features that could be extracted from the pentapeptide signatures exclusive to IDPs. Following key observations emerged - (i) these pentapeptides utilize all 20 amino acids, (ii) each of the 20 amino acids occur with very similar frequency in these pentapeptides regardless of their specific position (i.e. the frequency of occurrence of any given residue is similar in any position from 1 to 5 in any given pentapeptide containing that residue), and (iii) the frequency of occurrence individual residues, regardless of their position in a pentapeptide is independent of their (a) physical properties such as molecular weight and hydropathy index (b) chemical properties such as pI and pKa. Thus, compositionally, there do not appear to be any physico-chemical signatures of the unique pentapeptides exclusive to IDPs. That said, we noted that very recently, the role of IDPs in pathology was discovered in form of mutational "dileucinopathies" (Li & Babu, 2018; Meyer et al., 2018). Thus, we also analyzed the pentapeptides listed in Table S4 for the presence of di- and tri-homo-motifs. Of the 26577 unique pentapeptides exclusive to IDPs, a total of 4659 pentapeptides had di-motifs (listed in Table S7) and 144 had tri-motifs (listed in Table S8). Figure 2 shows residue-wise distributions of these homo-motifs (both di- and tri-).

Interestingly, di-leucine motif is the least occurring amongst all the di-motifs while di-methionine and di-histidine are the highest occurring. Of the many interpretations possible, the one emerging out appears to be that the discovered role of "di-leucinopathies" is reflected in the minimal occurrence of this motif – dileucine motifs are least preferred and hence mutations leading to formation of these motifs could lead to malfunction. This provides an appealing homo-motif role – the more mutational occurrences of homo-motifs found to be naturally less occurring would lead to homo-motif-pathies.

Unique and exclusive signatures identify new IDPs from 532638 sequences without structural information

We believe it is important to summarize the findings to emphasize the importance of the results obtained by an algorithmic approach that is in complete contrast to the standard approach of investigating consensus sequences as functional signatures. Figure 3 shows that utilizing the 36 unique tetrapeptides as exclusive signatures of IDPs identifies 1349 new IDPs (in addition to ∼1400 already listed in DisProt) out of more than half-a-million naturally occurring primary sequences without any structural information; simultaneous utilization of 26577 unique pentapeptides serving as exclusive signatures of IDPs identify that ∼50% of the curated sequences without structural information are actually intrinsically disordered.

Based on exclusive tetra- and penta-peptide signatures, signatures, the newly identified IDPs show the following appealing observations –

i. Only a small fraction of the newly identified IDPs belong to sequences from *Homo sapiens*; in fact the identified sequences belong to a wide variety of species ranging from prokaryotic to eukaryotic origins (but not to Archaea). Interestingly, a few viruses (>20) with mammalian, invertebrate and even plants as hosts are also found to contain sequences of the identified IDPs (e.g. Feline foamy virus; African swine fever virus; Human adenovirus 5; Human Herpesvirus; Invertebrate iridescent virus 6-IIV-6-Chilo iridescent virus; Nigerian sorghum potyvirus-Dawa mosaic virus; Mirafiori lettuce big-vein virus). E.g. based on tetrapeptides, the complete list of species is provided in column G of Table S6.

ii. There is a large variety of functions of the cellular proteins identified as IDPs. E.g., from column E of Table S6, it is clear that the newly identified IDPs play varying roles in metabolism, material transport, housekeeping (e.g. chaperones, polymerases, amino acid-tRNA ligases, ribosomal), secretion and signaling (kinases and phosphatases) along with others.

iii. The list of identified IDPs also contains certain organelle specific proteins in eukaryotes and specific proteins across several species (such grouping inferences are open to interpretations and thus, instead of making such inferences ourselves, we provide the actual data only). This is made clear, e.g., by close inspection of columns E and F in Table S6.

At this point, it is important to point out that the data on exclusive tetra- and penta- peptide signatures of IDPs result in discovery different IDP sets rather than one being a subset of the other, possibly due to the novelty of the approach used in this work (Figure 3). On one hand this may be perceived as an interpretational challenge - e.g. is there any significance of exclusive pentapeptide signatures identifying a very large number of new IDPs compared to the relatively small number of new IDPs identified by exclusive tetrapeptide signatures? On the other hand, our data clearly shows that the usual extrapolations of peptide motifs to primary sequences towards functional inferences are not as trivial as considered normally. Results obtained here are based on explicit identification of sequences exclusively present in IDPs only - these sequences are not found in proteins where even a part of the primary sequence of a protein are ordered/structured. Thus, the actual reason(s) behind the exclusive peptide signatures of IDPs still remain elusive. Although beyond the scope of this work, an interesting follow-up could be splitting current hybrid "StrucSeq" dataset into subsets containing "pure order" (i.e. whole Swiss-Prot entries with complete structure and parts of Swiss-Prot entries with "partial" PDB structures corresponding to ordered regions) and "putative PDB disorder" (i.e. remaining parts of Swiss-Prot entries with "partial" PDB structures). It may then be interesting to see if disorder-specific signatures derived for the current "IDPsSeq" dataset would be different from the disorder-specific signatures derived from the "putative PDB disorder" subset. Further, it may be also interesting to see if order-specific signatures that can be derived for the current hybrid "StrucSeq" dataset would be different from the order-specific signatures that can be derived from the "pure order" subset.

Conclusions

We have discovered a library of 36 tetrapeptides and 26577 pentapeptides that are exclusively present in intrinsically disordered proteins. In spite of the primary sequence space of IDPs being ~5% of the sequence space of structured proteins, discovery of unique peptides exclusively present in the former (and absent in the latter) is expected to provide a considerable boost to efforts in not only identifying IDPs in sequences without any structural information, but also in specific investigations on functional roles of IDPs based on their specific and exclusive signatures (i.e. tetra- and pentapeptide sequences). Such direct and unambiguous identification of a particular class of proteins, in this case IDPs, based on exclusive signatures, in contrast to the standard approach of investigating consensus sequences, is a first to our knowledge.

Acknowledgements

AMC is grateful to IIT Delhi for fellowship support. The authors also thank IIT Delhi for providing access to the HPC facility. AM is grateful to Kusuma Trust (UK) for their generous funding support towards assisting him in establishing the teaching and research programs of the School of Biological Sciences (subsequently renamed as the Kusuma School of Biological Sciences) at IIT Delhi. AM is also grateful to Dept. of Biotechnology, Government of India and the National Supercomputing Mission, Government of India for their support to the Supercomputing Facility for Bioinformatics & Computational Biology at IIT Delhi.

Author contributions

AMC and ST collected the data. AMC collected the complete peptide count data and ST independently confirmed the dipeptide and tripeptide count data. AMC also analyzed some of the data. AM designed the study, analyzed the data, prepared the figures and wrote the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Aditya Mittal http://orcid.org/0000-0002-4030-0951

References

Babu, M. M., Kriwacki, R. W., & Pappu, R. V. (2012). Versatility from protein disorder. *Science*, *337*(6101), 1460–1461. https://doi.org/10.1126/science.1228775

Berlow, R. B., Dyson, H. J., & Wright, P. E. (2018). Expanding the paradigm: Intrinsically disordered proteins and allosteric regulation. *Journal of Molecular Biology*, *430*(16), 2309–2320. https://doi.org/10.1016/j.jmb.2018.04.003

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, *35*(Database), D301–3. https://doi.org/10.1093/nar/gkl971

Bourhis, J. M., Johansson, K., Receveur-Bréchot, V., Oldfield, C. J., Dunker, K. A., Canard, B., & Longhi, S. (2004). The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Research*, *99*(2), 157–167. https://doi.org/10.1016/j.virusres.2003.11.007

Cheng, Y., Oldfield, C. J., Meng, J., Romero, P., Uversky, V. N., & Dunker, A. K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*, *46*(47), 13468–13477. https://doi.org/10.1021/bi7012273

Chouard, T. (2011). Breaking the protein rules. *Nature*, *471*(7337), 151–153. https://doi.org/10.1038/471151a

Communie, G., Ruigrok, R. W., Jensen, M. R., & Blackledge, M. (2014). Intrinsically disordered proteins implicated in paramyxoviral replication machinery. *Current Opinion in Virology* , *5*, 72–81. https://doi.org/10.1016/j.coviro.2014.02.001

Cukier, R. I. (2018). Conformational ensembles exhibit extensive molecular recognition features. *ACS Omega* , *3*(8), 9907–9920. https://doi.org/10.1021/acsomega.8b00898

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Molecular Biosystems*, *8*(1), 268–281. https://doi.org/10.1039/C1MB05231D

Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kubań, M., Strumillo, M., Uyar, B., Budd, A., Altenberg, B., Seiler, M., Chemes, L. B., Glavina, J., Sánchez, I. E., Diella, F., & Gibson, T. J. (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Research*, *42*(D1), D259–66. https://doi.org/10.1093/nar/gkt1047

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, *41*(21), 6573–6582. https://doi.org/10.1021/bi012159+

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., & Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, *6*(3), 197–208. https://doi.org/10.1038/nrm1589

Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J., & Bernadó, P. (2019). Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. *Structure*, *27*(2), 381–391.e2. https://doi.org/10.1016/j.str.2018.10.016

Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., Gürth, C.-M., Huber, A.-K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg, B., Reményi, A., Dinkel, H., & Gibson, T. J. (2018). The eukaryotic linear motif resource – 2018 update. *Nucleic Acids Research*, *46*(D1), D428–D434. https://doi.org/10.1093/nar/gkx1077

Gouw, M., Sámano-Sánchez, H., Van Roey, K., Diella, F., Gibson, T. J., & Dinkel, H. (2017). Exploring short linear motifs using the ELM database and tools. *Current Protocols in Bioinformatics*, *58*, 8.22.1–8.22.35. https://doi.org/10.1002/cpbi.26

Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., … Piovesan, D. (2020). DisProt: Intrinsic protein disorder annotation in 2020. *Nucleic Acids Research*, *48*(D1), D269–D276. https://doi.org/https://doi.org/10.1093/nar/gkz975.

Jensen, M. R., & Blackledge, M. (2014). Testing the validity of ensemble descriptions of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences of the United States of America* , *111*(16), E1557–8. https://doi.org/10.1073/pnas.1323876111

Katuwawala, A., Oldfield, C. J., & Kurgan, L. (2019). Accuracy of protein-level disorder predictions. *Brief Bioinform*, 46, 48. https://doi.org/10.1093/bib/bbz100.

Li, X. H., & Babu, M. M. (2018). Human Diseases from Gain-of-Function Mutations in Disordered Protein Regions. *Cell*, *175*(1), 40–42. https://doi.org/10.1016/j.cell.2018.08.059

Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V., & Mittag, T. (2016). Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *Journal of the American Chemical Society*, *138*(47), 15323–15335. https://doi.org/10.1021/jacs.6b10272

Meyer, K., Kirchner, M., Uyar, B., Cheng, J. Y., Russo, G., Hernandez-Miranda, L. R., Szymborska, A., Zauber, H., Rudolph, I. M., Willnow, T. E., Akalin, A., Haucke, V., Gerhardt, H., Birchmeier, C., Kühn, R., Krauss, M., Diecke, S., Pascual, J. M., & Selbach, M. (2018). Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell*, *175*(1), 239–253.e17. https://doi.org/10.1016/j.cell.2018.08.019

Mi, T., Merlin, J. C., Deverasetty, S., Gryk, M. R., Bill, T. J., Brooks, A. W., Lee, L. Y., Rathnayake, V., Ross, C. A., Sargeant, D. P., Strong, C. L., Watts, P., Rajasekaran, S., & Schiller, M. R. (2012). Minimotif Miner 3.0: Database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Research*, *40*(D1), D252–60. https://doi.org/10.1093/nar/gkr1189

Mittal, A., & Jayaram, B. (2011a). Backbones of Folded Proteins Reveal Novel Invariant Amino AcidNeighborhoods. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 443–454. https://doi.org/10.1080/073911011010524954

Mittal, A., & Jayaram, B. (2011b). The newest view on protein folding: stoichiometric and spatial unity in structural and functional diversity. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 669–674. https://doi.org/10.1080/073911011010524984

Mittal, A., Changani, A. M., & Taparia, S. (2019). What limits the primary sequence space of natural proteins? *Journal of Biomolecular Structure and Dynamics*, 1–5. https://doi.org/10.1080/07391102.2019.1682051

Mittal, A., Changani, A. M., Taparia, S., Goel, D., Parihar, A., & Singh, I. (2020). Structural disorder originates beyond narrow stoichiometric margins of amino acids in naturally occurring folded proteins. *Journal of Biomolecular Structure and Dynamics*, 1–12. https://doi.org/10.1080/07391102.2020.1751299

Mittal, A., Jayaram, B., Shenoy, S. R., & Bawa, T. S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding? *Journal of Biomolecular Structure and Dynamics*, *28*(2), 133–142. https://doi.org/10.1080/07391102.2010.10507349

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., & Uversky, V. N. (2006). Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology*, *362*(5), 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087

Oldfield, C. J., & Dunker, A. K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual Review of Biochemistry*, *83*(1), 553–584. https://doi.org/10.1146/annurev-biochem-072711-164947

Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., & Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, *44*(37), 12454–12470. https://doi.org/10.1021/bi050736e

Ota, H., & Fukuchi, S. (2017). Sequence conservation of protein binding segments in intrinsically disordered regions. *Biochemical and Biophysical Research Communications*, *494*(3–4), 602–607. https://doi.org/10.1016/j.bbrc.2017.10.099

Ozenne, V., Bauer, F., Salmon, L., Huang, J. R., Jensen, M. R., Segard, S., Bernadó, P., Charavay, C., & Blackledge, M. (2012). Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, *28*(11), 1463–1470. https://doi.org/10.1093/bioinformatics/bts172

Parigi, G., Rezaei-Ghaleh, N., Giachetti, A., Becker, S., Fernandez, C., Blackledge, M., Griesinger, C., Zweckstetter, M., & Luchinat, C. (2014). Long-range correlated dynamics in intrinsically disordered proteins. *Journal of the American Chemical Society*, *136*(46), 16201–16209. https://doi.org/10.1021/ja506820r

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., … Tosatto, S. C. (2017). DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Research*, *45*(D1), D219–D227. https://doi.org/10.1093/nar/gkw1056

Salvi, N., Abyzov, A., & Blackledge, M. (2019). Solvent-dependent segmental dynamics in intrinsically disordered proteins. *Science Advances*, *5*(6), eaax2348. https://doi.org/10.1126/sciadv.aax2348

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*, D506–515.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, *27*(10), 527–533. https://doi.org/10.1016/S0968-0004(02)02169-2

Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A million peptide motifs for the molecular biologist. *Molecular Cell*, *55*(2), 161–169. https://doi.org/10.1016/j.molcel.2014.05.032

Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, *11*(4), 739–756. https://doi.org/10.1110/ps.4210102

Uversky, V. N. (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins*, *1*(1), e24684. https://doi.org/10.4161/idp.24684

Uversky, V. N. (2019). Intrinsically disordered proteins and their "mysterious" (meta) physics. *Frontiers in Physics*, *7*, 10. https://doi.org/10.3389/fphy.2019.00010

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, *41*(3), 415–427. https://doi.org/10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., & Dunker, A. K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *Journal of Proteome Research*, *6*(6), 2351–2366. https://doi.org/10.1021/pr0701411

Vacic, V., Uversky, V. N., Dunker, A. K., & Lonardi, S. (2007). Composition Profiler: A tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*, *8*(1), 211. https://doi.org/10.1186/1471-2105-8-211

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., & Babu, M. M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, *114*(13), 6589–6631. https://doi.org/10.1021/cr400525m van der Lee, R., Lang, B., Kruse, K., Gsponer, J., Sánchez de Groot, N., Huynen, M. A., Matouschek, A., Fuxreiter, M., & Babu, M. M. (2014). Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Reports*, *8*(6), 1832–1844. https://doi.org/10.1016/j.celrep.2014.07.055

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, *293*(2), 321–331. https://doi.org/10.1006/jmbi.1999.3110

Zarin, T., Strome, B., Nguyen Ba, A. N., Alberti, S., Forman-Kay, J. D., & Moses, A. M. (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife*, *8*, e46883. https://doi.org/10.7554/eLife.46883

Zhang, L., Li, M., & Liu, Z. (2018). A comprehensive ensemble model for comparing the allosteric effect of ordered and disordered proteins. *PLOS Computational Biology*, *14*(12), e1006393. https://doi.org/10.1371/journal.pcbi.1006393