JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
2021, VOL. 39, NO. 7, 2364–2375
https://doi.org/10.1080/07391102.2020.1751299

Structural disorder originates beyond narrow stoichiometric margins of amino acids in naturally occurring folded proteins

Aditya Mittal[a,b], Anandkumar Madhavjibhai Changani[a], Sakshi Taparia[c], Deepanshu Goel[d†], Animesh Parihar[d†] and Ishan Singh[e†]

[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; [d]Department of Biochemical Engineering and Biotechnology (Bachelors program), Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India; [e]Department of Computer Science & Engineering (Bachelors program Computer Science), Indian Institute of Technology Delhi (IIT Delhi), New Delhi, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

Rigorous analyses of Euclidean distances between non-peptide bonded residues in structures of several thousand naturally occurring folded proteins yielded a surprising "margin of life" for percentage occurrence of individual amino acids in naturally occurring folded proteins. On one hand, the concept of "margin of life", referring to lower than expected variances in average stoichiometric occurrences of individual amino acids in folded proteins, remains unchallenged since its discovery a decade ago. On the other hand, within this past decade there has been a strong emergence of a gradual paradigm shift in biology, from sequence-structure-function in proteins to sequence-disorder-function, fuelled by discoveries on functional implications of intrinsically disordered proteins (primary sequences that do not form stable structures). Thus the applicability of "margin of life" to peptide-bonded residues in all known natural proteins, adopting stable structures vis-à-vis intrinsically disordered needs to be explored. Therefore in this work, we analyze compositions of the complete naturally occurring primary sequence space (over 560000 sequences) after dividing it into mutually exclusive subsets of structured and intrinsically disordered proteins along with a subset without any structural information. While finding that occurrence of different peptides (up to pentapeptides) is a direct consequence of the relative occurrences of their constituting residues in folded proteins, we report that structural disorder in natural proteins originates beyond the narrow stoichiometric margins of amino acids found in structured proteins.

ARTICLE HISTORY
Received 20 September 2019
Accepted 20 March 2020

KEYWORDS
Disorder; IDPs; peptides; proteins; structure; sequence; synthesis; computational biology; bioinformatics

Introduction

The narrow margins of values of physico-chemical variables such as temperature and pH in which different living systems exist are well accepted (Dill et al., 2011; Ghosh et al., 2016; Ghosh & Dill, 2010). In fact, it is recognized that the narrow range of values of the physico-chemical variables at which molecular components function together for assembling and maintaining living systems are often different from the "optimum" conditions applicable individually to the molecular components (Ghosh & Dill, 2010). Further, along with the environmental factors, species-specific compositional constraints for the above molecular components are also well established. The simplest examples being those of codon bias in genetic codes (Sharma et al., 2008; Sun & Caetano-Anollés, 2008; Zhang et al., 2019), variations in mesophilic and archaeal proteins (Caetano-Anollés et al., 2012) and prevalence of different membrane lipids in biological membranes of specific organisms (Bansal & Mittal, 2015). Interestingly, while discovery of the Chargaff's rules in the last century (Chargaff, 1950) recognizing the compositional constraints on DNA regardless of classification of living systems played a key role in structural elucidation of DNA (Watson & Crick, 1953), appreciation of such species-independent compositional constraints on protein primary sequences has not yet fully matured since their recent discovery (Mittal et al., 2010).

About a decade ago, several thousands of high resolution ($\leq$ 2.5 Å) structures of naturally occurring folded proteins in the PDB were analyzed for extracting the presumed preferential interactions between specific residues in folded proteins (e.g. negatively and positively charged side chains, non-polar side chains with each other) by measuring Euclidian distances between non-peptide-bonded residues (Mittal et al., 2010). From analyses of the largest structural dataset at that time, it was found that Euclidian distances between α-carbon atoms of residues with oppositely charged polar side chains were no different from distances between polar and non-polar/hydrophobic residues or non-polar/hydrophobic and

CONTACT Aditya Mittal amittal@bioschool.iitd.ac.in
†Equal contribution.

non-polar/hydrophobic residues (Mittal et al., 2010; Mittal & Jayaram, 2011a; 2011b). In addition, regardless of the protein size or 3 D structure, the probability of any two non-peptide bonded residues being close together was found to depend primarily on their percentage occurrence in primary sequences rather than the physico-chemical nature of their side chains (Mittal et al., 2010; Mittal & Jayaram, 2011a; 2011b). The findings were termed as controversial, "bordering on revolutionary" (Sarma, 2011).

In order to contest the findings, an independent test of the methodology developed was applied on DNA structures (Galzitskaya et al., 2011). Remarkably, the results showed that, if existent, preferential interactions (i.e. A-T and G-C in case of DNA) were indeed extracted from structural data (Galzitskaya et al., 2011; Mittal & Jayaram, 2011b) by applying the methodolgy. In contrast, the only preferential interaction found in protein structures was Cys-Cys – the agreement between prediction (Agutter, 2011) and results (Mittal & Jayaram, 2011a; 2011b) further strengthened the reliability of the methodology utilized. Experimental validation of the apparently" blasphemous" results on percentage occurrence of amino acids (i.e. their stoichiometry) in protein sequences being responsible for their 3 D structures rather than the presumed preferential interactions between non-peptide bonded residues further came independently from (a) computational experiments showing jumbled protein sequences with the same amino-acid stoichiometry arriving at identical thermodynamic conformations in different simulation times (Song et al., 2011), and, (b) wet experiments showing that a mixture of non-peptide bonded amino acids in the same proportion as amino acids of a natively folded polypeptide sequence have the same conformational signatures (Schirò et al., 2011). Thus, structural data had serendipitiously uncovered stoichiometric constraints on amino acids constituting protein sequences as a primary feature in protein folding. Finally, an independent discovery of the fact that standard deviations of average percentage occurrence of amino acids in all structured protein sequences were found to be much lower than standard deviations expected from random normal distributions of amino acids (Mezei, 2011) led to defining of the" stoichiometric margins of life" for the constrained amino acid compositions resulting in structured proteins (Mittal & Jayaram, 2011b; 2012)

On one hand, it may appear in retrospect that simply calculating percentage occurrence of amino acids in all known protein sequences could have been a straight forwad approach towards insights into stoichiometric constraints in naturally occurring protein sequences. On the other hand, the elucidation of stoichiometric margins of life from structural data on non-peptide-bonded residues in structured protein sequences provided a completely new view on protein folding. At that time, the area of disordered proteins was still in its infancy with extremely limited data on intrinsically disordered proteins (IDPs), i.e. sequences that do not fold into a stable structure (Chouard, 2011). Since then, not only several discoveries establishing important biological roles of IDPs have been reported (Berlow et al., 2018; Li & Babu, 2018; Meyer et al., 2018; Salvi et al., 2019; Tompa et al., 2014; van der Lee et al., 2014), but the amount of data on IDPs has also increased more than ten-fold (Piovesan et al., 2017). Therefore, time and data are now appropriate to explore whether compositional constraints applicable to structured proteins are similar to or different from those of IDPs. The key question is –"are stoichiometric margins of amino acids in structured proteins applicable to IDPs and hence all naturally occurring primary sequences?" To answer this question, the straight forward approach of simply calculating occurrence of amino acids in primary sequences to different classes of proteins (e.g. Structured vs IDPs) was applied.

Methods

Complete sequence data was downloaded from Uniprot (Swiss-prot) on 28th July 2019 as per instructions provided for offline analyses (The UniProt Consortium, 2019). Results presented are from this dataset. The list of all computationally synthesized peptides and the exact number of times they occur in all curated sequences downloaded from Uniprot (Swiss-Prot) are available upon request. An earlier download of complete sequence data on 15th February 2019 was done to develop and test the analytical codes. Datasets were also downloaded on 21st September 2019 and 30th October 2019. Negligible differences (almost nil) in the results were obtained from analyses of all the datasets. 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. Data analyses was done in MATLAB (Mathworks Inc.) and MS Excel. Mean and standard deviation of occurrence for each of the dipeptides were calculated based on the following equations:

$$\mu = \frac{\sum_{i=1}^{N}\left(\frac{n_i}{R_i - 1} \times 100\right)}{N} \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(\frac{n_i}{R_i - 1} \times 100\right) - \mu\right]^2}{N - 1}} \tag{2}$$

where, N = total number of sequences, n_i = number of occurrences of a given dipeptide in the i^{th} sequence, and R_i = number of total residues in the i^{th} sequence. Similarly, mean and standard deviation of occurrence for each of the tripeptides were calculated based on the following equations:

$$\mu = \frac{\sum_{i=1}^{N}\left(\frac{n_i}{R_i - 2} \times 100\right)}{N} \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(\frac{n_i}{R_i - 2} \times 100\right) - \mu\right]^2}{N - 1}} \tag{4}$$

Following the above, mean and standard deviation of occurrence of each of the tetrapeptides and pentapeptides were calculated by using "$R_i - 3$" and "$R_i - 4$" respectively.

Here it is also important to state a key assumption in this work on comparing amino acid frequencies in primary

Table 1. Distribution of individual amino acids in different datasets.

Amino acid	StrucSeq ($n = 27199$)			OnlySeq ($n = 532553$)			IDPsSeq ($n = 707$)			IDPsUnRev ($n = 94$)		
L	9.23	±	2.81	9.60	±	3.10	8.08	±	2.94	8.03	±	3.33
A	7.87	±	3.30	8.48	±	3.67	7.80	±	3.79	8.09	±	4.13
G	7.03	±	2.68	7.19	±	2.86	7.23	±	3.71	6.99	±	3.69
E	6.83	±	2.70	6.59	±	2.83	7.87	±	3.89	7.37	±	3.36
V	6.76	±	2.25	7.08	±	2.44	5.87	±	2.29	6.04	±	2.22
S	6.73	±	2.66	6.21	±	2.62	7.53	±	3.00	6.89	±	3.09
K	6.38	±	3.12	6.16	±	3.44	7.09	±	3.95	7.94	±	4.68
R	5.43	±	2.63	5.81	±	3.09	5.49	±	3.35	5.16	±	2.78
D	5.41	±	1.90	5.28	±	2.11	5.64	±	2.54	5.85	±	2.45
T	5.41	±	1.95	5.24	±	1.94	5.43	±	2.09	5.82	±	2.25
I	5.40	±	2.30	6.11	±	2.68	4.57	±	2.34	4.94	±	2.33
P	4.97	±	2.39	4.50	±	2.30	5.79	±	3.52	4.70	±	3.58
N	4.10	±	1.95	3.92	±	2.13	3.96	±	2.13	4.73	±	2.47
Q	3.95	±	1.91	3.77	±	2.01	4.69	±	2.79	5.13	±	4.19
F	3.76	±	1.66	3.88	±	2.04	3.27	±	1.77	3.24	±	1.89
Y	3.06	±	1.51	2.87	±	1.58	2.56	±	1.48	2.89	±	1.88
M	2.39	±	1.19	2.56	±	1.32	2.38	±	1.19	2.29	±	1.39
H	2.23	±	1.26	2.23	±	1.37	2.12	±	1.58	1.63	±	1.28
C	1.85	±	2.55	1.45	±	1.99	1.63	±	2.11	1.43	±	2.52
W	1.20	±	0.99	1.05	±	1.06	1.00	±	0.91	0.83	±	0.97

sequences of all known naturally occurring proteins. Since the primary sequence data utilized is manually curated, it has been assumed that annotation (from genomes to proteomes) is accurate in Swiss-Prot. While assuming the same, we also carefully inspected the different species from which the data has been compiled in Swiss-Prot and Disprot. This is important since earlier work on comparisons between different species of living organisms has provided evidence for the dependence of amino acid frequencies on the genomic GC content (Lightfield et al., 2011; Zhou et al., 2014). Neither did we find any species specificity in Swiss-Prot vs. Disprot, nor did we find any species specificity while cross-referencing the species sources of our data with the earlier studies. This simply allowed us to safely assume that the genomic GC content resulting in the primary sequence datasets used in this work are (at least) very similar, if not exactly the same.

Results

Amino acid distributions in different classes of primary sequences and narrow stoichiometric margins of life

We first collected curated (manually reviewed) primary sequence data from the Uniprot: Swiss-Prot database (The UniProt Consortium, 2019). Then, by careful cross-referencing with the Protein Data Bank: PDB (Berman et al., 2000; 2007) and DisProt (Piovesan et al., 2017), we divided the Swiss-Prot data into three mutually exclusive datasets of primary sequences – (i) the first dataset containing all primary sequences with folded protein structures (of varying resolutions) was called" StrucSeq", (ii) the second dataset containing all primary sequences classified as IDPs was called" IDPsSeq", and (iii) the third dataset containing all primary sequences without any structural information was called" OnlySeq". In addition to the above, we found some sequences in DisProt that were not present in Swiss-Prot but present in TrEMBL by cross-referencing. We called this dataset of sequences as" IDPsUnRev" (unreviewed sequences, not part of Swiss-Prot, classified as IDPs in DisProt). Thus, the complete primary sequence space of naturally occurring proteins was divided into four mutually exclusive datasets. Table 1 shows percentage occurrence of amino acids, in form of mean ± standard deviation, in the above datasets.

Since the original discovery of compositional constraints was based on protein structures, amino acids are shown in the order of decreasing percentage occurrence in the" StrucSeq" dataset. Two primary observations emerge from inspecting Table 1 – (a) standard deviations for all amino acids are apparently lower in structured proteins compared to IDPs, and (b) there appears to be a difference between average percentage occurrence of amino acids in" StrucSeq" and IDPs. In order to explore these apparent observations, we plotted the occurrence statistics of amino acids in the different datasets in comparison to" StrucSeq".

Figure 1a and b show that both mean and standard deviation of percentage occurrence of amino acids of "OnlySeq" are highly correlated with "StrucSeq". On the contrary, Figure 1c–f show that the relative correlations of the parameters of IDPs with "StrucSeq" are lower. Thus Figure 1a–f collectively indicate – (a) stoichiometric constraints on occurrence of amino acids in primary sequences are applicable, with varying degrees, to folded/structured proteins as well as IDPs, (b) "OnlySeq" contains a large number of the sequences that could result in structured proteins as well as sequences that must be IDPs since the correlations with "StrucSeq" are very high but are still < 1.0, and (c) IDPs have compositional variations distinct from structured proteins. Statistically, the apparently high value of regression coefficients between IDPs and "StrucSeq" indicates that the stoichiometric constraints on amino acids in primary sequences that result in a stable structure are quite narrow – however, outside the narrower margins of structured proteins, the resulting primary sequences would result in IDPs. As a measure of these narrow stoichiometric margins, we further decided to look at "standard deviation/mean" for individual amino acids in the four mutually exclusive data sets. Figure 1g clearly shows that not only are the stoichiometric margins for relative occurrence of each amino acid in primary sequences the

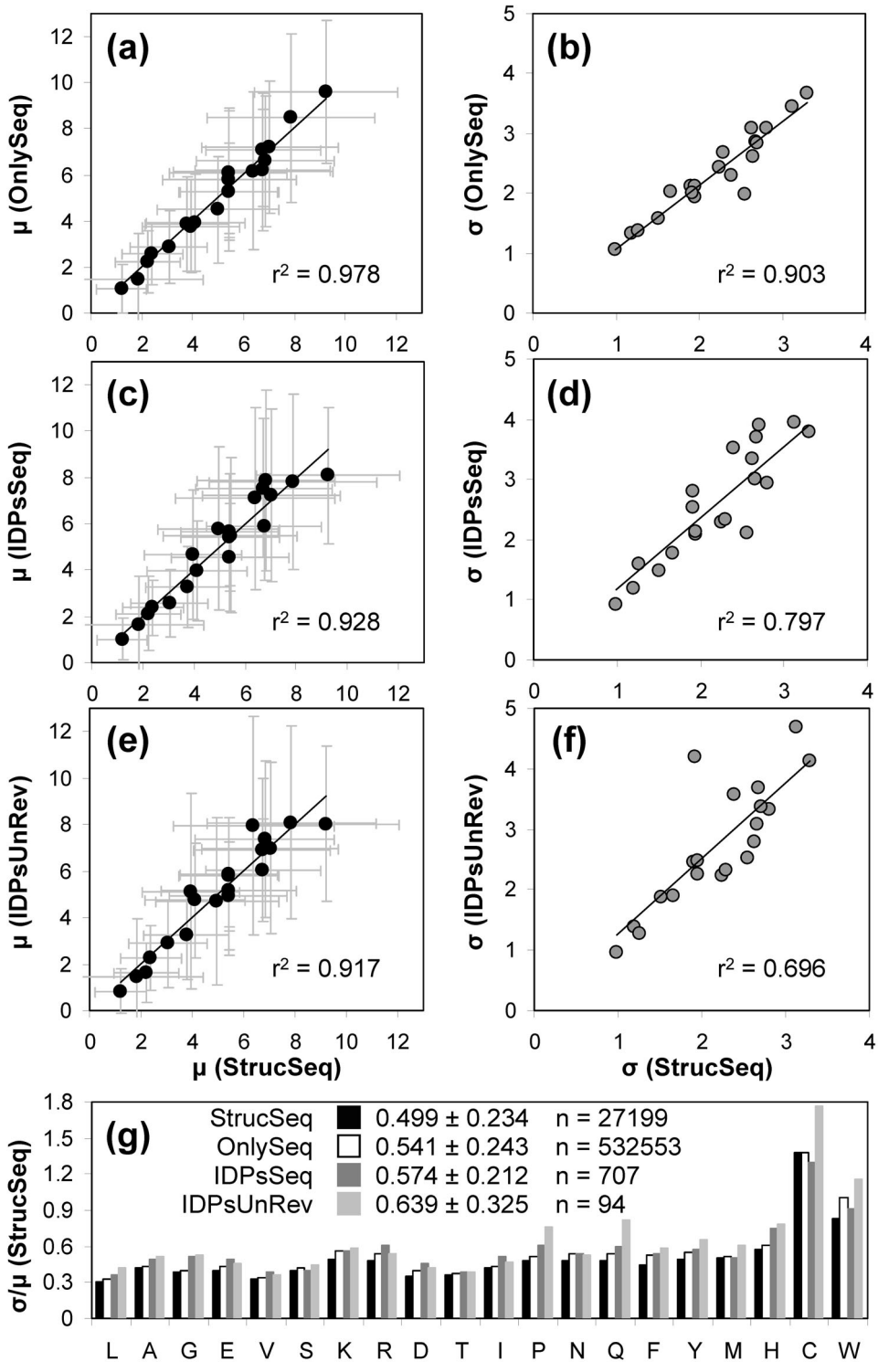

Figure 1. Narrow stoichiometric margins of life. Relationship of Mean (μ) & Standard-deviation (σ) of percentage occurrence of individual amino acids in all curated sequences without structure (dataset: OnlySeq) in Swiss-Prot – (a) & (b), all curated Intrinsically Disordered Protein sequences (dataset: IDPsSeq) in Swiss-Prot (cross-referenced with DisProt) – (c) & (d), and, un-reviewed Intrinsically Disordered Protein sequences (dataset: IDPsUnRev) in DisProt (cross-referenced with UniProt TrEMBL) – (e) & (f), with μ & σ of percentage occurrence of individual amino acids in all structured protein sequences (dataset: StrucSeq) in the Protein Data Bank (cross-referenced with Swiss-Prot). (g) σ/μ for each of the twenty amino acid residues in the four different datasets – values for all the amino acids are lowest for the dataset StrucSeq indicating the stoichiometric margins of life. Also shown are the mean ± std of the values of σ/μ for all twenty amino acids in all the four datasets along with the number of sequences in each dataset. Note that the X-axis shows amino acids in the decreasing order of mean percentage occurrence in all structured protein sequences (dataset: StrucSeq), as given in Table 1, from left to right, i.e. Leucine occurs most frequently and Tryptophan occurs least frequently, on an average, in structured protein sequences.

lowest in structured proteins, but also that these margins are distinctly higher in IDPs. Here it is pertinent to mention that an earlier assessment of non-curated protein sequences obtained from PSORT, eSLDB and Refseq databases had reported percentage mean compositions in different eukaryotes (Gaur, 2014) – interestingly, the data presented in

Table 1 here appears to be very similar to non-membrane eukaryotic proteins of the earlier report (Gaur, 2014). Thus, in natural proteins regardless of species, narrow margins of amino acid occurrences in primary sequences is similar to the well accepted and observed narrow ranges of several physico-chemical variables required for supporting life.

Stoichiometric constraints applicable to dipeptides

Having established the narrow stoichiometric margins for occurrence of individual amino acids in naturally occurring protein sequences, the next obvious step was to explore whether these are applicable to peptide-bonded partners also. While the original discovery of the stoichiometric margins was an "undesired" consequence of searching for preferential interactions between non-peptide-bonded residues in naturally occurring folded proteins (Mittal et al., 2010), here we asked whether there are any particular preferences for peptide-bonded partners in the four mutually exclusive protein datasets. To do so, we first synthesized a library of all possible dipeptides ($20 \times 20 = 400$) that can result from the 20 natural amino acids. Then we counted the occurrence of each of the dipeptides in all the sequences of StrucSeq, IDPsSeq, IDPsUnRev and OnlySeq datasets; two reading frames of each sequence starting from the amino terminus were read to count the occurrence of dipeptides as shown on the top of Figure 2. From the counted occurrences of the dipeptides, we were able to calculate the mean and standard deviation of occurrence for each of the dipeptides using Eq. 1 and Eq. 2 (see Methods) respectively.

Figure 2 shows the mean and standard deviation of occurrence of each of the dipeptides, represented by heatmaps, in all four sequence datasets. The darker the color, the higher is the value (i.e. higher mean occurrence or higher standard deviation). Rows represent the first amino acid of the dipeptide and columns represent the second amino acid. The order of amino acids is the same as Table 1, i.e. in the order of decreasing mean percentage occurrence of individual amino acids in StrucSeq. The heatmap for mean occurrence of dipeptides in StrucSeq clearly shows a gradient – dark color at the top left corner starts becoming lighter as one moves towards the bottom right corner. This clearly shows that occurrence of dipeptides is primarily dictated by percentage occurrence of individual amino acids, i.e. there are no preferences for any particular dipeptide formation. On the other hand, the heatmaps for the occurrences of dipeptides show some clear deviations in the color gradient from the top left to bottom right, thereby indicating a deviation in occurrence of dipeptides from that dictated by the percentage occurrence of individual amino acids in StrucSeq. Thus, the data on occurrence of dipeptides in IDPs supports the findings observed for individual amino acids seen in Figure 1. Similarly, data on occurrence of dipeptides in OnlySeq also support the findings observed for individual amino acids seen in Figure 1. Clearly, stoichiometric distribution of individual amino acids "largely" dictates occurrence of dipeptides, with possible exceptions in IDPs that need to be explored further. These possible exceptions indicate preferential dipeptide occurrences and detailed investigations on these exceptions as signatures of IDPs are beyond the scope of this work (they are being pursued separately). That said, it is pertinent to mention the results reported earlier (Caetano-Anolles et al., 2013) show a remarkable agreement with our findings in terms of non-random patterns of dipeptides observed by them independently in different family folds of protein structures with varying flexibility. Clearly, it may be possible to extract dipeptide signatures (with a high propensity) for increasing flexibility or creating disorder in protein structures based on the differences in the heatmaps of StrucSeq and IDPs. Interestingly, heatmaps of dipeptides obtained by us for the complete set of natural protein sequences show similar trends to a highly selective set of natural proteins (Santoni et al., 2016), especially w.r.t. dipeptides involving specific residues (e.g. C, H, W). This strongly indicates stoichiometric constraints applicable to dipeptides based on relative abundance of individual residues may be even a more generalized feature than appreciated earlier.

Stoichiometric constraints applicable to tripeptides

The next step was to investigate whether the results found above applied to tripeptides also. In order to do so, we synthesized a library of all possible tripeptides ($20x20x20 = 8000$) and developed a way to visualize the relative occurrence of tripeptides using heatmaps as shown in Figuer 3.

Each of the 400 blocks in the heatmaps shown in Figure 2 were further subdivided into 20 rows as shown in Figure 3. The first residue of a given tripeptide was represented by the letter given on the left, the second residue was represented by one of the 20 sub-row of each row in Figure 2 and the third residue of the tripeptide was represented by the column. In an ideal scenario, i.e. if tripeptide occurrence was a direct result of probabilistic occurrence of individual amino acids (i.e. probability of occurrence of a tripeptide = - product of probabilities of occurrences of the 3 individual amino acids it is composed of), the heatmap for a given tripeptide would be expected to appear as shown in the lower panel of Figure 3 since the rows, sub-rows and columns all are in the order of decreasing mean occurrence of individual amino acids.

Having developed a method to visualize the relative occurrences of tripeptides, we counted the occurrence of each of the tripeptides in all the sequences of StrucSeq, IDPsSeq, IDPsUnRev and OnlySeq datasets; three reading frames of each sequence starting from the amino terminus were read to count the occurrence of tripeptides as shown on the top of Figure 4. From the counted occurrences of the tripeptides, we were able to calculate the mean and standard deviation of occurrence for each of the tripeptides using Eq. 3 and Eq. 4 (see Methods) respectively.

Figure 4 shows the heatmaps representing mean and standard deviation of occurrence of each of the tripeptides in all four sequence datasets. The heatmap for mean occurrence of tripeptides in StrucSeq again clearly shows a gradient as predicted in Figure 3 – dark color at the top left corner starts becoming lighter as one moves towards the bottom right corner, with each sub-row showing gradients similar to those expected in the case of tripeptide occurrence being a direct result of probabilistic occurrence of its

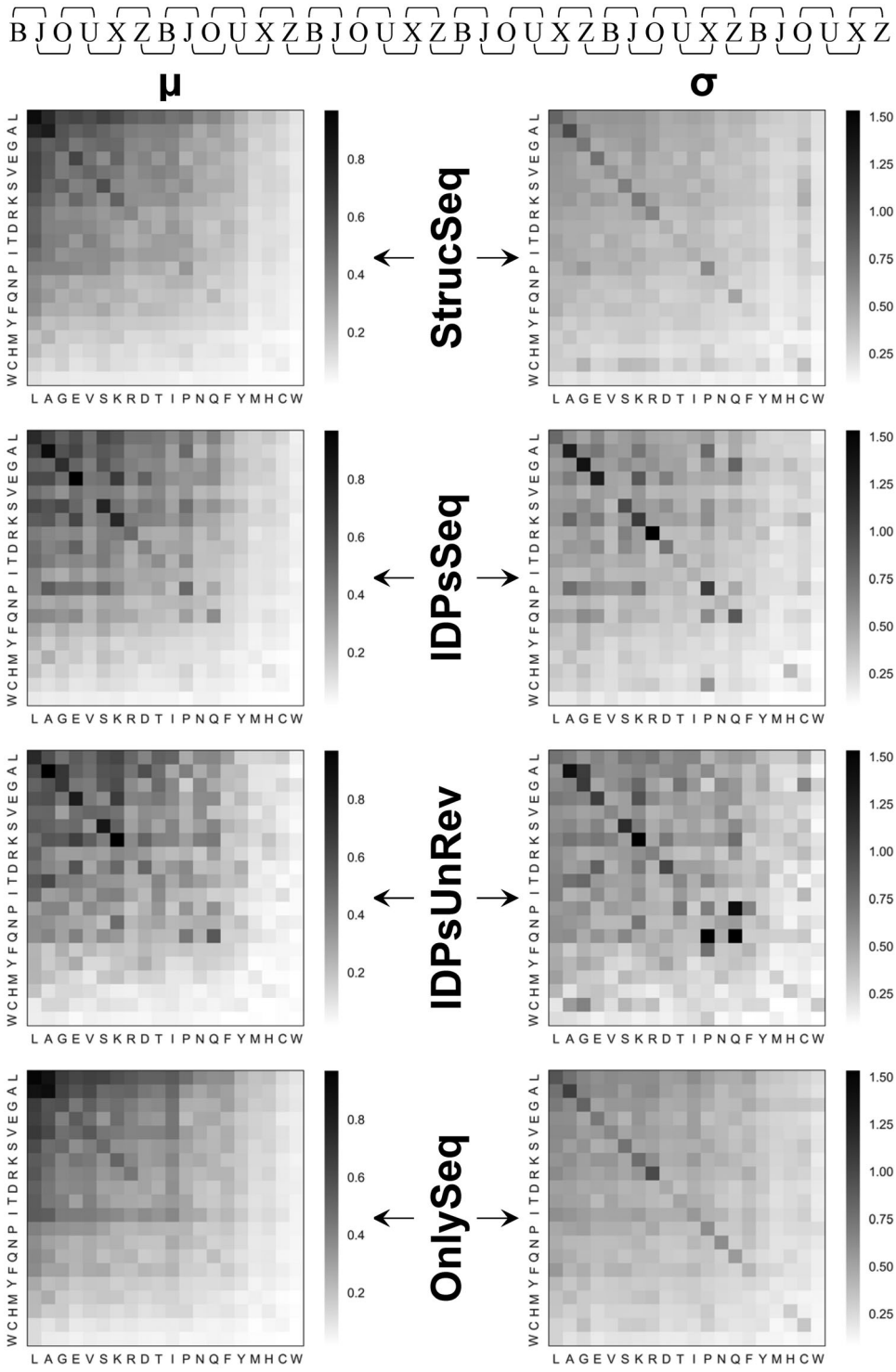

Figure 2. Stoichiometric distribution of individual amino acids "largely" dictates occurrence of dipeptides. A given primary sequence was read using two reading frames, starting from the first two residues of its amino-terminus, to count the occurrence of any dipeptide as shown at the top. The number of occurrences of all possible dipeptides (20x20 = 400) in a given set of primary sequences was recorded. The Mean (μ) & Standard-deviation (σ) of percentage occurrence of each of the dipeptides were calculated based on Equation 1 (see text). Heat maps representing μ & σ in the complete primary sequence space (divided into the exclusive four datasets – see text and Figure 1 for details) of natural proteins are shown. In each heat map, the first residue of a dipeptide is represented row-wise and the second residue is represented column-wise. Thus, there are a total of 400 blocks in each heat map, with each block representing a unique dipeptide – e.g. blocks in the top row of each heat map, starting from left, represent the dipeptides LL, LA, LG, LE, LV, LS, LK, LR, LD, LT, LI, LP, LN, LQ, LF, LY, LM, LH, LC and LW respectively. Dark color (towards black) represents high percentage occurrence and light color (towards white) represents low percentage occurrence. The dipeptide residue order, i.e. from top to bottom (representing the first residue) and left to right (representing the second residue) is the same as that in Figure 1.

individual constitutents. Thus, occurrence of tripeptides is also primarily dictated by percentage occurrence of individual amino acids at least in StrucSeq, i.e. there are no preferences for any particular dipeptide formation. At the same time, as observed earlier, the heatmaps for the occurrences of tripeptides show some clear deviations in the color

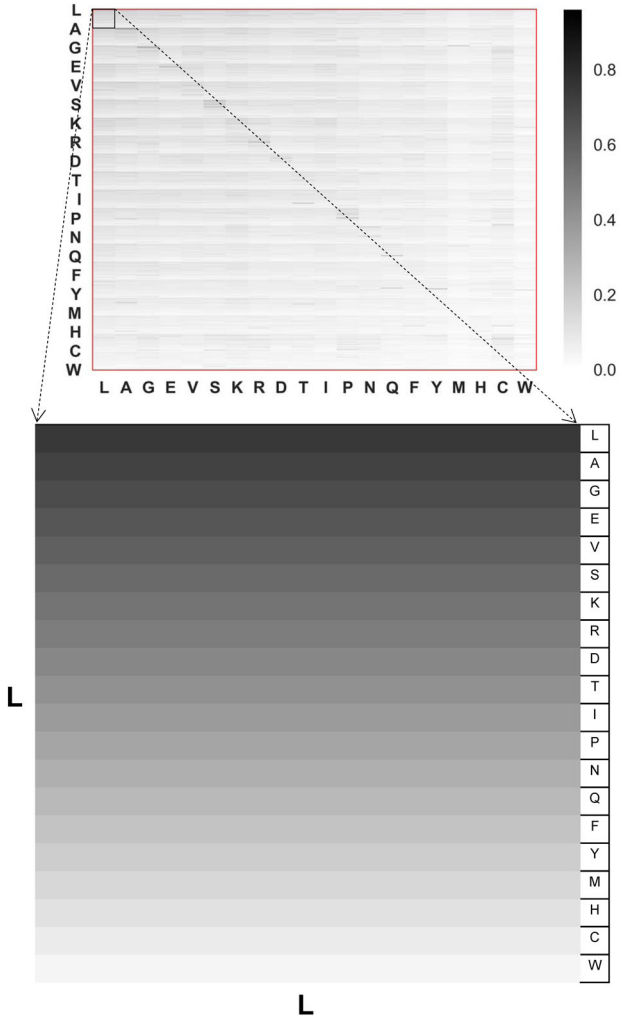

Figure 3. Visualizing relative occurrence of tripeptides using heat maps. Each row of the heat maps in Figure 2 was sub-divided into 20 more rows, corresponding to the second residue of the tripeptide. Thus, as shown here for the first block of the first row corresponding to the top left block in Figure 2, the first residue of a tripeptide is "L" and the third residue is "L". 20 sub-rows are created within this block, each sub-row representing the second residue of the tripeptide. Thus, the rows in the single block shown, starting from top to bottom, represent the tripeptides LLL, LAL, LGL, LEL, LVL, LSL, LKL, LRL, LDL, LTL, LIL, LPL, LNL, LQL, LFL, LYL, LML, LHL, LCL and LWL respectively. A total of 8000 rectangular blocks are there in each heat map, with each block representing a unique tripeptide. Dark color (towards black) represents high percentage occurrence and light color (towards white) represents low percentage occurrence. The tripeptide residue order, i.e. from top to bottom (representing the first and second residues) and left to right (representing the third residue) is the same as that in Figure 1.

gradient from the top left to bottom right for IDPs in spite of an apparent maintenance of the overall gradient. Thus, the data on occurrence of tripeptides in IDPs supports the findings observed for individual amino acids seen in Figure 1 and dipeptides seen in Figure 2. Similarly, data on occurrence of tripeptides in OnlySeq also support the findings from Figures 1 and 2. Clearly, stoichiometric distribution of individual amino acids "largely" dictates occurrence of tripeptides, with possible exceptions in IDPs that need to be explored further. These possible exceptions indicate preferential tripeptide occurrences in IDPs; as stated earlier detailed investigations on these exceptions as signatures of IDPs are beyond the scope of this work (they are being pursued separately).

Stoichiometric constraints applicable to longer peptides and amplification of deviation from the margins in disordered proteins

Having discovered (a) the occurrences of di- and tri- peptides being primarily governed by stoichiometric occurrences of individual amino acids in StrucSeq, and, (b) emergence of deviations of di- and tri- peptide occurrences from those expected based on individual amino acids in IDPs, we decided to check whether these findings are applicable to tetra- and penta- peptides also. Therefore, we computationally synthesized libraries of all possible tetrapeptides (20x20x20x20 = 160000) and pentapeptides (20x20x20x20x20 = 3200000). Then, as done earlier, we counted the occurrences of each of the tetrapeptides and pentapeptides in the sequences of all four datasets. Figure 5a shows the correlation between the actual number (frequency) of occurrences and the expected number (frequency) of occurrences of individual amino acids, dipeptides, tripeptides, tetrapeptides and pentapeptides in all the four datasets.

The expected numbers of occurrences were calculated by simply multiplying the frequency of occurrences of individual amino acids constituting the respective peptides. Figure 5b shows the correlation between the mean occurrence and the expected mean occurrence – the expected mean occurrence was calculated by simply multiplying the mean occurrence of individual amino acids constituting the respective peptides. Note that at the single amino acid level, all four data sets show $R = 1$. This is because, the frequency of occurrence or mean occurrence of each amino acid of each of the four datasets was independently considered (instead of considering w.r.t. StrucSeq as done earlier). Thus, reference for each of the four data sets in Figure 5a and b were the occurrences of amino acids specific to each of the four respective datasets. Remarkably in IDPs, the deviations of expected occurrences start showing highly significant differences from the actual occurrences, in terms of significantly lower values or R, with increasing peptide size. Clearly, the stoichiometric constraints on amino acids and peptides applicable to structured proteins are not applicable to disordered protein sequences. The same results are observed when reference for each of the four datasets is taken as amino acid occurrences only in StrucSeq, as shown in Figure. 5c and d. Therefore, a clear conclusion emerges – structural disorder originates beyond narrow stoichiometric margins of amino acids in structured proteins in all naturally occurring protein sequences regardless of their species-based classifications.

Interestingly, these results are extremely well supported by another very recent and independent study (Mezei, 2019). Firstly, Mezei (2019) established "*importance of the sequence following the known AA propensities*". Secondly, in terms of "*adjacency propensities*" larger differences were observed in distributions of tri- and tetra-peptides compared to individual residues and dipeptides. Thus, in spite of the fact that a propensity score for peptide bonded neighbors correlating with protein foldability/stability was elusive in Mezei (2019), the overall results did show an amplification of differences between naturally folded and artificial sequences for the longer peptides.

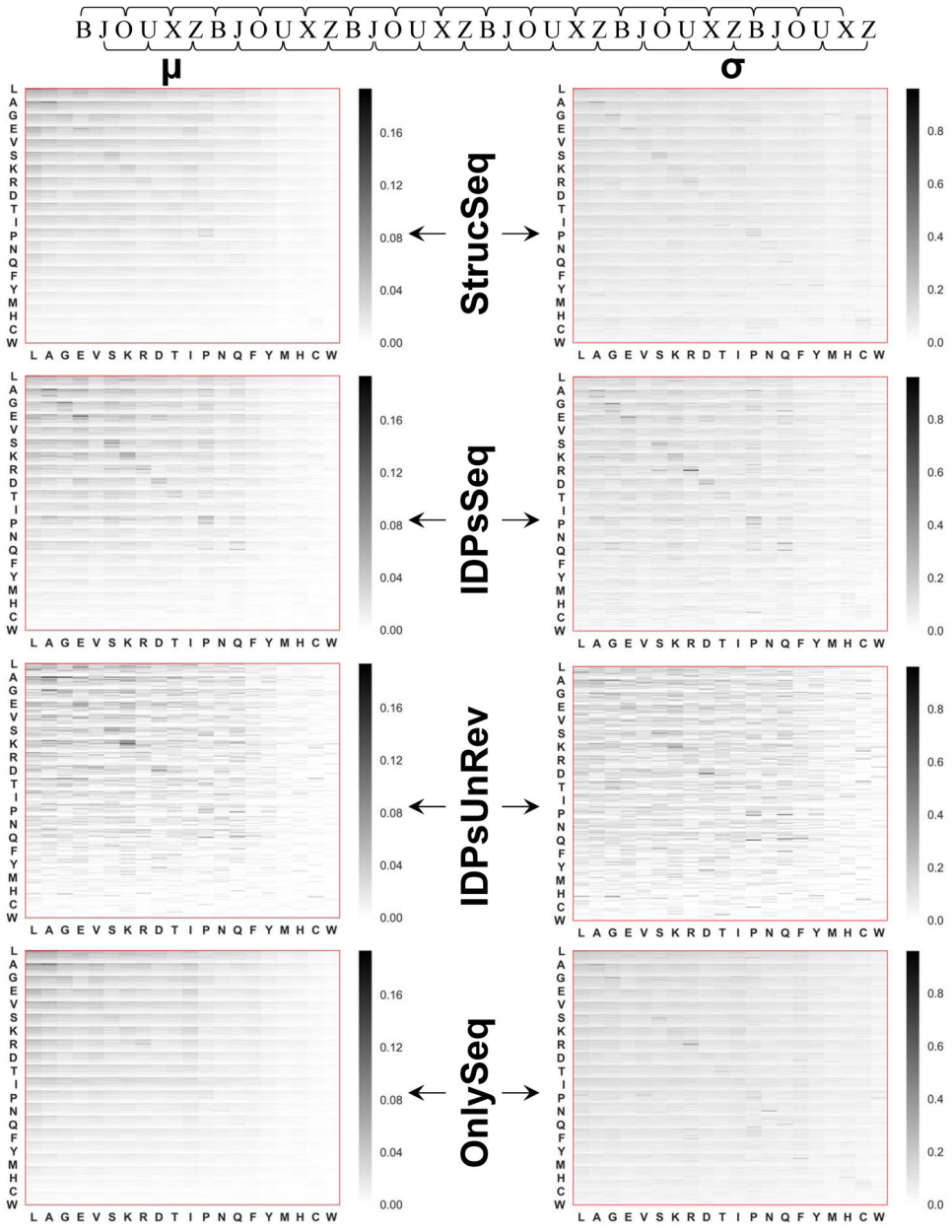

Figure 4. Stoichiometric distribution of individual amino acids "largely" dictates occurrence of tripeptides. A given primary sequence was read using three reading frames, starting from the first three residues of its amino-terminus, to count the occurrence of any tripeptide as shown at the top. The number of occurrences of all possible tripeptides (20x20x20 = 8000) in a given set of primary sequences was recorded. The Mean (μ) & Standard-deviation (σ) of percentage occurrence of each of the tripeptides were calculated based on Equation 2 (see text). Heat maps representing μ & σ in the complete primary sequence space (divided into the exclusive four datasets – see text and Figure 1 for details) of natural proteins are shown. Each row of the heat maps in Figure 2 was sub-divided into 20 more rows, corresponding to the second residue of the tripeptide (see Figure S1). Thus, as shown here in each heat map, the first residue of a tripeptide is shown row-wise and the third residue is shown column-wise. Therefore, a total of 8000 blocks are there in each heat map, with each block representing a unique tripeptide – e.g. blocks in the top row of each heat map, starting from left, represent the tripeptides LLL, LLA, LLG, LLE, LLV, LLS, LLK, LLR, LLD, LLT, LLI, LLP, LLN, LLQ, LLF, LLY, LLM, LLH, LLC and LLW respectively. Dark color (towards black) represents high percentage occurrence and light color (towards white) represents low percentage occurrence. The tripeptide residue order, i.e. from top to bottom (representing the first and second residues) and left to right (representing the third residue) is the same as that in Figure 1.

Stoichiometric constraints on occurrences of amino acids and peptides are not related to their size and hydrophobicity

The original discovery of the stoichiometric margins of life (Mittal et al., 2010; Mittal & Jayaram, 2011b) based on structural analyses of non-peptide-bonded neighbors had shown an unexpected independence of Euclidian distances between residues from their chemical properties (e.g. polar, non-polar, positively or negatively charged side chains). Therefore, here we tested whether occurrences of individual amino acids and peptides depended on their size or hydrophobicity.

Figure 6 shows that occurrence of longer peptides is independent of their molecular weights and hydrophobicities (hydrophobicity of a given peptide was calculated by adding the hydropathy index of the individual amino acid constituents; the GES scale developed by Engelman et al. (1986) was used). While there appear to be indications that (a) larger

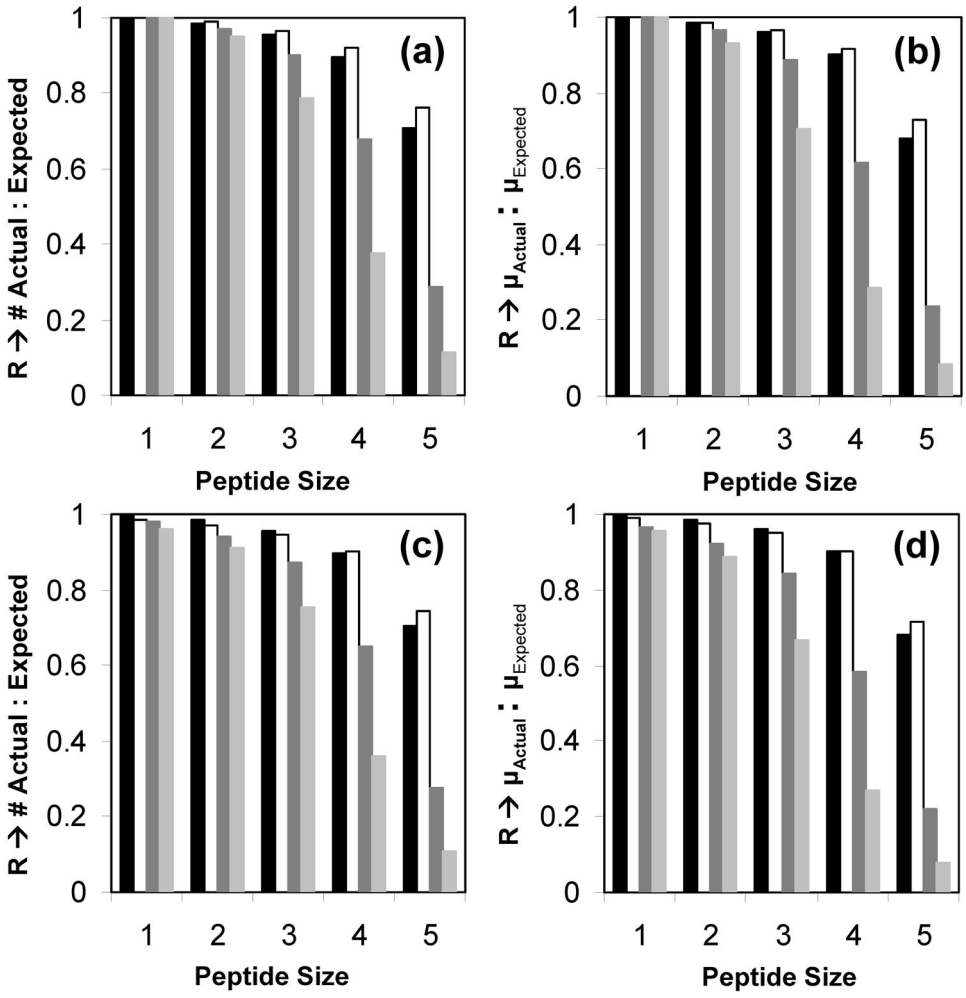

Figure 5. Amplification of deviations from narrow stoichiometric margins of life in disorder with increasing peptide size. (a) Pearson's correlation coefficient (R) between the number of actual occurrences of a peptide of a given length and the expected number of occurrences of the peptide based on the frequency of the single residues constituting the peptide in each respective dataset – black bars represent StrucSeq, white bars represent OnlySeq, dark gray bars represent IDPsSeq and light gray bars represent IDPsUnRev. (b) R between the mean percentage occurrence of a peptide of a given length and the expected percentage occurrence of the peptide based on the mean percentage occurrence of the single residues constituting the peptide in each respective dataset. (c) R between the number of actual occurrences of a peptide of a given length and the expected number of occurrences of the peptide based on the frequency of the single residues constituting the peptide in the dataset StrucSeq. (d) R between the mean percentage occurrence of a peptide of a given length and the expected percentage occurrence of the peptide based on the mean percentage occurrence of the single residues constituting the peptide in the dataset StrucSeq. Bars represent the same respective datasets as in Figure 1g.

molecular weight amino acids occur less compared to lighter amino acids in all four datasets, and, (b) heavier dipeptides and tripeptides have lesser occurrences in StrucSeq compared to IDPs from Figure 6a due to $R < -0.5$, it is clear that the chemical nature of amino acids (reflected by the hydrophobicity values) does not have any role in primary sequences of all four datasets. Here it is important to note earlier reports indicate factors such as metabolic costs of amino acids (Krick et al., 2014) or availability of different number of codons per amino acid in the genetic code (Mittal & Jayaram, 2012) or availability of pools of tRNA (Mittal et al., 2019) may be responsible for relative abundance of amino acids in proteomes. However, considering parameters such as codon bias, habitat constraints, existence of essential and non-essential amino acids etc. in different species, the search for universal factors resulting in the observed stoichiometric margins of life for individual amino acids remains an open question.

Discussion and conclusions

It may appear obvious that any two protein sets may show differences in amino acid compositions. Alternatively, it also may appear obvious that any sets of proteins having same functions may show similarity in amino acid compositions. However, there are substantial examples in literature where both of the above need to be closely inspected and a case-by-case basis set of interpretations emerges. E.g., translational dependent folding observed is indeed observed in proteins with similar (or even identical) primary sequences but some differences in folding lead to different protein functionalities (Komar, 2007; Sharma et al., 2008). In this work, we do not specifically do any case-by-case comparisons in manually segregated protein datasets. Rather, we have compared the complete available population of existing natural protein sequences in order to extract universal indicators and features. Thus, while the actual distinct assessment of where exactly is the "margin of life" located in

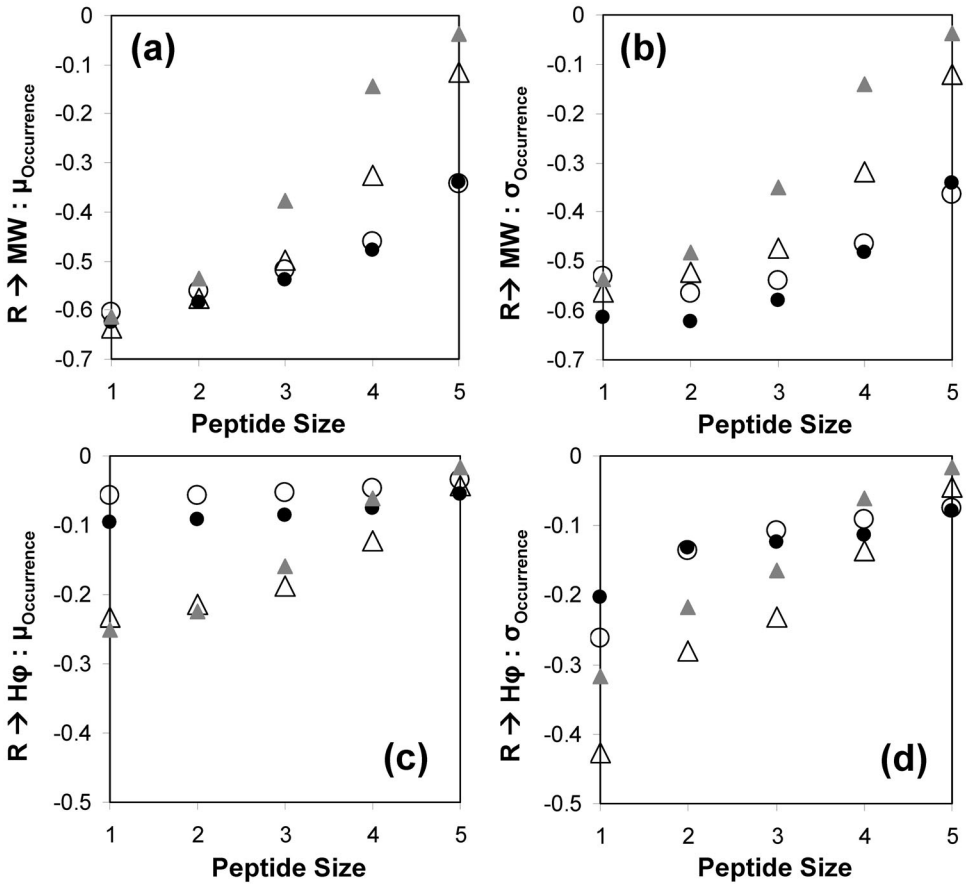

Figure 6. Occurrences of amino acids and peptides are not related to their size and hydrophobicity. (a) Pearson's correlation coefficient (R) between Mean (μ) of percentage occurrence of a single amino acid or a peptide and its molecular weight as a function of peptide size (i.e. single amino acid = 1, dipeptide = 2, tripeptide = 3, tetrapeptide = 4 and pentapeptide = 5). (**b**) R between Standard-deviation (σ) of percentage occurrence of a single amino acid or a peptide and its molecular weight as a function of peptide size. (**c**) R between Mean (μ) of percentage occurrence of a single amino acid or a peptide and its hydropathy index, i.e. sum of hydropathy indeces of its constituting residue(s), as a function of peptide size. (**d**) R between Standard-deviation (σ) of percentage occurrence of a single amino acid or a peptide and its hydropathy index as a function of peptide size. In all the panels, solid circles represent StrucSeq, open circles represent OnlySeq, open triangles represent IDPsSeq and closed gray triangles represent IDPsUnRev.

continuous distributions in the StrucSeq column of Table 1 may be difficult at this time, the closest we have reached so far towards measuring the margins are in Figure 1G and 5 respectively in terms of the (a) ratios of standard deviation to mean for different datasets and, (b) the correlation between actual and expected occurrence of amino acids in different datasets, respectively. How these differences can be extrapolated to individual proteins, rather than whole populations, still remains a challenge – this is similar to analyzing rare or marginal individual events occurring near the tails in statistical distributions. The above said, this work opens up very promising avenues for (a) understanding the complete primary sequence space of natural proteins based on stoichiometric margins of life (Mittal et al., 2019) and, (b) developing somewhat straightforward predictors of intrinsic disorder based on stoichiometric compositions and occurrence of specific peptides in primary sequences which is beyond the scope of the current manuscript (Mittal et al., 2020).

In conclusion, by analyzing the complete data available on primary sequences of proteins, we comprehensively show that stoichiometric margins of life, i.e. a narrow band of relative occurrences of amino acids discovered earlier from structural data, are applicable to all naturally occurring primary sequences from the perspective of synthesis of protein sequences by peptide bonding of individual residues. We also show that deviations beyond these stoichiometric margins are clearly observed in intrinsically disordered proteins and these deviations are substantially amplified with increasing the number of peptide bonds – i.e. longer sequences having occurrences of residues (individual or peptide bonded) beyond the compositional constraints of amino acids in structured proteins are prone to be IDPs. Therefore, our results unambiguously lead to the conclusion, at least at a highly qualitative and/or semi-quantitative level, that structural disorder in natural proteins originates beyond the narrow stoichiometric margins of amino acids that constitute structured proteins. Of course, the physical mechanisms behind these stoichiometric margins need to be explored in future studies.

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 sequence data and wrote the codes for counting the number of occurrences. AM and AMC analyzed the data and prepared the figures. DG, AP and IS recollected the sequence data at different times and wrote independent codes for counting the number of occurrences to confirm whether there were any changes in the results (fortunately negligible changes in raw data and no changes in overall results were observed). AM designed the study, supervised the work and wrote the manuscript.

Disclosure statement

The authors declare no competing interests.

References

Agutter, P. S. (2011). Stoichiometry-driven protein folding: A comment. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 643–644. doi:10.1080/073911011010524974

Bansal, S., & Mittal, A. (2015). A statistical anomaly indicates symbiotic origins of eukaryotic membranes. *Molecular Biology of the Cell*, *26*(7), 1238–1248. doi:10.1091/mbc.E14-06-1078

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. doi:10.1016/j.jmb.2018.04.003

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 issue), D301–303. doi:10.1093/nar/gkl971

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.

Caetano-Anollés, G., Kim, K. M., & Caetano-Anollés, D. (2012). The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *Journal of Molecular Evolution*, *74*(1-2), 1–34. doi:10.1007/s00239-011-9480-1

Caetano-Anollés D. (2013). Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. *PLoS One*. *8*, e72225. doi:10.1371/journal.pone.0072225.

Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, *6*(6), 201–209. doi:10.1007/BF02173653

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

Dill, K. A., Ghosh, K., & Schmit, J. D. (2011). Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences of the United States of America*., *108*(44), 17876–17882. doi:10.1073/pnas.1114477108

Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences ofmembrane proteins. *Annu Rev Biophys Biophys Chem*, *15*, 321–353.

Gaur, R. K. (2014). Amino acid frequency distribution among eukaryotic proteins. *IIOAB Journal*, *5*(2), 6–11.

Ghosh, K., de Graff, A. M. R., Sawle, L., & Dill, K. A. (2016). Role of proteome physical chemistry in cell behavior. *The Journal of Physical Chemistry B*, *120*(36), 9549–9563. doi:10.1021/acs.jpcb.6b04886

Ghosh, K., & Dill, K. A. (2010). Cellular proteomes have broad distributions of protein stability. *Biophysical Journal*., *99*(12), 3996–4002. doi:10.1016/j.bpj.2010.10.036

Komar, A. A. (2007). SNPs, Silent but not invisible. *Science*, *315*(5811), 466–467. doi:10.1126/science.1138239

Krick, T., Verstraete, N., Alonso, L. G., Shub, D. A., Ferreiro, D. U., Shub, M., & Sánchez, I. E. (2014). Amino Acid metabolism conflicts with protein diversity. *Mol Biol Evol*. *31*, 2905–12. doi:10.1093/molbev/msu228.

Li, X. H., & Babu, M. M. (2018). Human diseases from gain-of-function mutations in disordered protein regions. *Cell*, *175*(1), 40–42. doi:10.1016/j.cell.2018.08.059

Lightfield, J., Fram, N. R., & Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*., *6*(3), e17677. doi:10.1371/journal.pone.0017677

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. doi:10.1016/j.cell.2018.08.019

Mezei, M. (2011). Discriminatory Power of Stoichiometry-Driven Protein Folding?. *Journal of Biomolecular Structure and Dynamics*., *28*(4), 625–626. doi:10.1080/073911011010524966

Mezei, M. (2019). On predicting foldability of a protein from its sequence, *Protein*, 88 (2), 355–365. doi:10.1002/prot.25811

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. doi:10.1080/07391102.2019.1682051

Mittal, A., Changani, A. M., & Taparia, S. (2020). Unique and exclusive peptide signatures directly identify intrinsically disordered proteins from sequences without structural information. *Journal of Biomolecular Structure and Dynamics*. 10.1080/07391102.2020.1756410

Mittal, A., & Jayaram, B. (2011a). Backbones of Folded Proteins Reveal Novel Invariant Amino Acid Neighborhoods. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 443–454. doi: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. doi:10.1080/073911011010524984

Mittal, A., & Jayaram, B. (2012). A possible molecular metric for biological evolvability. *Journal of Biosciences*, *37*(3), 573–577. doi:10.1007/s12038-012-9210-x

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. doi:10.1080/07391102.2010.10507349

O. V., Galzitskaya, O. V., Lobanov, M. Y., & Finkelstein, A. V. (2011). Cunning simplicity of a stoichiometry driven protein folding thesis. *Journal of Biomolecular Structure and Dynamics*., *28*(4), 595–598. doi:10.1080/073911011010524958

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.

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

Santoni, D., Felici, G., & Vergni, D. (2016). Natural vs. random protein sequences: Discovering combinatorics properties on amino acid words. *Journal of Theoretical Biology*., *391*, 13–20. doi:10.1016/j.jtbi.2015.11.022

Sarma, R. H. (2011). A conversation on protein folding. *Journal of Biomolecular Structure and Dynamics*, *28*(4), 587–588. doi:10.1080/073911011010524955

Schirò, G., Caronna, C., Natali, F., Koza, M. M., & Cupane, A. (2011). The "protein dynamical transition" Does not require the protein polypeptide chain. *The Journal of Physical Chemistry Letters*, *2*(18), 2275–2279. doi:10.1021/jz200797g

Sharma, M., Hasija, V., Naresh, M., & Mittal, A. (2008). Functional control by codon bias in magnetic bacteria. *Journal of Biomedical Nanotechnology*., *4*, 44–51.

Song, Y., Song, Y., & Chen, X. (2011). The yeast prion case: Could there be a uniform concept underlying complex protein folding?. *Journal of Biomolecular Structure & Dynamics*, *28*(4), 663–666. doi:10.1080/073911011010524982

Sun, F.-J., & Caetano-Anollés, G. (2008). Evolutionary patterns in the sequence and structure of transfer RNA: A window into early translation and the genetic code. *PLoS One.* , *3*(7), e2799. doi:10.1371/journal.pone.0002799

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*. *47*(D1), D506–515.

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. doi:10.1016/j.molcel.2014.05.032

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. doi:10.1021/cr400525m

Watson, J. D., & Crick, F. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. doi:10.1038/171737a0

Zhang, H., Li, J., Wang, R., Zhi, J., Yin, P., & Xu, J. (2019). Comparative analysis of expansin gene codon usage patterns among eight plant species. *Journal of Biomolecular Structure and Dynamics*, *37*(4), 910–917. doi:10.1080/07391102.2018.1442746

Zhou, H. Q., Ning, L. W., Zhang, H. X., & Guo, F. B. (2014). Analysis of the relationship between genomic GC Content and patterns of base usage, codon usage and amino acid usage in prokaryotes: Similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages. *PLoS One*, *9*(9), e107319. doi:10.1371/journal.pone.0107319