Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-functional Interactions

Benjamin Dubreuil, Or Matalon and Emmanuel D. Levy

Department of Structural Biology, *Weizmann Institute of Science, Rehovot 7610001, Israel*

Correspondence to Emmanuel D. Levy: *emmanuel.levy@weizmann.ac.il*
https://doi.org/10.1016/j.jmb.2019.08.008
Edited by Monika Fuxreiter

Abstract

In eukaryotes, disordered regions cover up to 50% of proteomes and mediate fundamental cellular processes. In contrast to globular domains, where about half of the amino acids are buried in the protein interior, disordered regions show higher solvent accessibility, which makes them prone to engage in non-functional interactions. Such interactions are exacerbated by the law of mass action, prompting the question of how they are minimized in abundant proteins. We find that interaction propensity or "stickiness" of disordered regions negatively correlates with their cellular abundance, both in yeast and human. Strikingly, considering yeast proteins where a large fraction of the sequence is disordered, the correlation between stickiness and abundance reaches $R = -0.55$. Beyond this global amino-acid composition bias, we identify three rules by which amino-acid composition of disordered regions adjusts with high abundance. First, lysines are preferred over arginines, consistent with the latter amino acid being stickier than the former. Second, compensatory effects exist, whereby a sticky region can be tolerated if it is compensated by a distal non-sticky region. Third, such compensation requires a lower average stickiness at the same abundance when compared to a scenario where stickiness is homogeneous throughout the sequence. We validate these rules experimentally, employing them as different strategies to rescue an otherwise sticky protein fragment from aggregation. Our results highlight that non-functional interactions represent a significant constraint in cellular systems and reveal simple rules by which protein sequences adapt to that constraint. Data from this work are deposited in Figshare, at https://doi.org/10.6084/m9.figshare.8068937.v3.

Introduction

Intrinsically disordered regions (IDRs) are found in proteins across all three domains of life and predominantly among eukaryotes where they cover 35% to 45% of proteomes [1]. IDRs are generally depleted of hydrophobic amino acids, which make them unable to fold autonomously [2,3]. Consequently, they are characterized by a lack of stable tertiary structure under physiological conditions [4]. The high-conformational variability associated with IDRs grants them unique functional and mechanical properties [4–11]. For example, nucleoporins are size-filtering devices of the nuclear pore complex containing disordered regions rich in phenylalanine–glycine repeats that influence the nuclear pore complex gating behavior [12]. In addition, disordered regions are notorious for their role in mediating and scaffolding protein-protein interactions [13–17]. One recognition mechanism involves domain–peptide interactions, where a globular domain such as WW, PDZ, or SH3 recognizes a linear motif in a disordered region [18–25]. Another class of recognition mechanism has been described as "induced fit" [26,27], where folding occurs concurrently with binding, promoting highly specific and transient interactions [15,22,28–30] often found in cell signaling and regulation processes [19,23,31–33]. Disordered regions thus play a pivotal role in the evolution [34–38] and the wiring of protein-protein interactions networks [34–37,39–42].

Importantly, however, properties of disordered regions that promote functional recognition are also expected to promote dysfunctional recognition [43–45]. Indeed, disordered regions naturally exhibit high solvent exposure, and we know that exposed protein surfaces show a natural tendency to bind one another promiscuously [46,47], suggesting that a delicate balance exists in cells, between functional (selected) and non-functional (non-selected) binding [44,48–60]. Such a delicate balance is reflected in the dosage sensitivity associated with over-expression of proteins that contain disordered regions or that are highly promiscuous [43,44,61,62]. Consistent with this view, proteins containing disordered regions are under tight regulation with shorter half-lives, slower translation, and faster degradation, so as to minimize their presence when they are not needed [48,57]. In addition, such potential for promiscuous binding is exploited by viruses to hijack cell regulation [63–65]. Finally, beyond promiscuous and dysfunctional interactions, disordered regions should exhibit specific properties so as not to be recognized as misfolded proteins by cellular chaperones, thereby threatening cellular homeostasis [66–71].

As for any chemical equilibrium, dysfunctional interactions are expected to be concentration dependent [72,73]. Therefore, we anticipate that sequence properties protecting against dysfunctional interactions [45,57,72,74–77] are enriched in IDRs of abundant proteins. To investigate this hypothesis, we considered the proteomes from *Saccharomyces cerevisiae* and *Homo sapiens* and compared properties of disordered regions present in low *versus* high-abundance proteins. As observed before, we saw that disordered regions are twice as rare in the latter [44,48]. Importantly, however, because abundance spans several orders of magnitude, the small fraction of disordered regions in abundant proteins represents a large mass of disordered residues in the cell, suggesting that these regions must be endowed with properties that render them compatible with high abundance. We show that IDRs' interaction propensity (stickiness) covaries with abundance more than other physicochemical properties do, including hydrophobicity and beta-amyloid formation propensity. Thus, dysfunctional interactions constrain the overall chemical composition of disordered regions. Analyzing the distribution of stickiness among disordered regions revealed that compensatory effects exist, whereby a non-sticky region in one part of a protein can buffer a sticky region hundreds of amino acids away. We verified this prediction experimentally and showed that an aggregation-prone misfolded protein can be rescued by the addition of charged residues delocalized relative to the sticky, aggregation-prone region.

Results and Discussion

Disorder content and protein abundance are inversely related

As discussed above, disordered regions are prone to promiscuous binding [43,78,79]. The fact that binding is a concentration-dependent process prompts us to compare the distribution of disordered regions among proteins of high *versus* low abundance. It has been shown before that cellular systems have evolved to regulate the availability of IDRs, whereby proteins enriched in IDRs are less abundant on average [48]. Here, we first confirm this finding. We examined protein disorder content (i.e., the percentage of disordered amino acids per protein) as a function of protein abundance for yeast and human. We defined 10 classes of abundance, henceforth referred to as "bins of abundance" (Figs. 1a and S1a, Table 1). Each bin contains approximately the same number of proteins. Abundance thresholds associated with these bins (in parts per million, or ppm) are kept identical throughout the work, but the number of proteins contained in each bin may differ in subsequent analyses when we concentrate on proteome subsets (e.g., proteins with at least 30 disordered residues). For all non-membrane proteins, we show the distribution of protein disorder content (Figs. 1b and S1b) for each bin of abundance.

Considering the first four bins, we notice that disorder content increases with abundance both in yeast (Fig. 1b) and human (Fig. S1b), although the potential for promiscuous interactions is theoretically increasing. We suggest two possible origins for this trend: (i) Below the abundance range of the fourth bin, disordered regions are not selected against due to very low concentrations equivalent to 5–30 nM (1–7 ppm, Table 1). Such concentrations are indeed below typical affinity constants of domain-peptide interactions (K_d ~ 0.1–150 μM) [80]. In line with this idea, linear motifs in disordered regions are also depleted in this range of abundance in yeast (Fig. S2). (ii) Disorder prediction methods may under-predict disordered regions [81] among proteins with low abundance due to their enrichment in sticky amino acids that are predominantly hydrophobic, as we will see in the next section. In support of this notion, Pfam [82] and SUPERFAMILY [83] domains become rarer as protein abundance decreases (Fig. 1c), suggesting that disorder should increase rather than decrease.

From the fifth to tenth bins of abundance, however, we observe a continuous decrease of protein disorder content, both in yeast and in human. Protein sequences in the fifth bin contain ~20% disorder on average (~40% in human, Fig. S1), whereas proteins with the highest abundance (tenth bin)

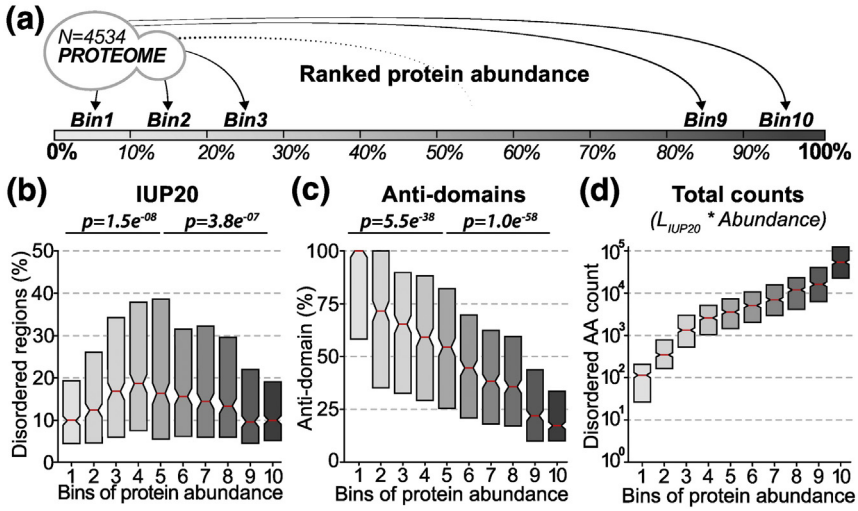

Fig. 1. High-abundance proteins are depleted in disordered regions, yet they dominate the mass fraction of disordered residues in the cell. (a) The full yeast proteome, membrane proteins excluded, is split into 10 bins of protein abundance with equal sizes. Bins 1 and 10 correspond to the 10% proteins with lowest and highest abundance in *S. cerevisiae*. (b) Boxplot distribution of disorder content across the 10 bins of protein abundance. The disorder content (*y*-axis) corresponds to the percentage of disordered residues detected per protein using IUPred [84] (see Methods). *P* values indicate whether the median disorder content significantly differs between bin 5 against bin 1 or 10 (one-sided Wilcoxon signed-rank test). (c) Same as panel b, counting residues not covered by any Pfam [82,85] or SUPERFAMILY [83,86] structural domain ("anti-domain"). *P* values are calculated as in panel b. (d) Boxplot distribution of absolute amino acid counts in disordered regions with increasing levels of protein abundance. The absolute amino-acid counts (*y*-axis) correspond to the number of disordered amino acids multiplied by the cellular abundance of the protein (in ppm). Each box shows 50% of the density distribution where notches represent the 95% confidence interval around the median (red line).

contain only ~10% (~25% in human). This difference is highly significant in both species (yeast: $p = 3.8 \times 10^{-07}$, human: $p = 1.8 \times 10^{-21}$, Wilcoxon one-sided test). The depletion of disordered regions among high-abundance proteins confirms the findings of Gsponer *et al*. [48], where IDR-poor proteins exhibit significantly higher abundances than IDR-rich proteins. These ideas also match results from Vavouri *et al*. [44], who observed that disordered regions and linear motifs were the main determinants of protein toxicity upon over-expression in yeast. Accordingly, we find that disordered regions become depleted of linear motifs in high-abundance proteins (Fig. S2). Overall, these results are consistent with binding promiscuity of disordered regions being a constraint, causing high-abundance proteins to adapt by decreasing their disordered content. Importantly, however, while the depletion is significant, it is not complete. In other words, since protein abundances span several orders of magnitude inside a cell, the ~10% of disordered regions present in high-abundance proteins contribute as much as 60%–75% of all disordered residues in the cell (Fig. 1d). This prompted us to investigate whether these regions bear specific properties that minimize their potential for promiscuous interactions and aggregation.

The stickiness of disordered regions is dependent on protein abundance

We saw in the previous section that disordered regions are expected to exhibit properties minimizing their susceptibility for promiscuous interactions. To evaluate the tendency of disordered regions to engage in promiscuous interactions, we measured their interaction propensity as the average of their amino-acid interaction propensities (Figs. 2a and S3a). Amino-acid interaction propensities were taken from Levy *et al*. [78] and correspond to the log-ratio of their frequency at protein-protein interfaces relative to protein surfaces (Fig. S4a). Thus, in this "stickiness" scale, the higher the score of an amino acid, the more frequent it is at protein–protein interaction interfaces relative to protein surfaces. The stickiness scale shares similarities with amino-acid solubility, hydrophobicity, and aggregation scales as shown in Fig. S4b. Although the stickiness scale was defined on structured domains, we expect it to reflect the binding propensity of disordered regions as well, because similar frequencies of amino acids are seen at the interface cores of both types of complexes [87].

Importantly, disordered regions often cover only a small fraction of the protein, and solvent-exposed

Table 1. Equivalence between bins and range of abundance (A) with estimated cellular concentrations (C) for yeast and human

Bins	Deciles (%)	*S. cerevisiae* (yeast)		*H. sapiens* (human)	
		A (ppm)	C (nM)	A (ppm)	C (nM)
1	0–10	0.003–1.4	0.0–6.5	1e–5–0.1	0.0–0.5
2	10–20	1.4–3.0	6.5–13	0.1–0.3	0.5–1.5
3	20–30	3.0–7.8	13.0–33.9	0.3–0.5	1.5–2.5
4	30–40	7.8–14.1	33.9–65.3	0.5–1.0	2.5–5.0
5	40–50	14.1–23.4	65.3–111.1	1.0–2.1	5.0–10.5
6	50–60	23.4–37.9	111.1–180.9	2.1–4.5	10.5–22.4
7	60–70	37.9–64.9	180.9–311.5	4.5–10.1	22.4–50.8
8	70–80	64.9–119	311.5–583.1	10.1–24.8	50.8–123.6
9	80–90	119–351	583.1–1700	24.8–75	123.6–398.7
10	90–100	351–21,870	1700–109,000	75–30,590	398.7–152,400

amino acids outside of these regions may influence the global protein stickiness. For this reason, we used three data sets of proteins with "standard," "medium," and "high" disorder content (Figs. 2a and S3a, see Methods) and we anticipate that data sets with higher disorder content should better reflect compositional constraints. We show the distributions of protein length, disorder content, and disorder fraction in these three data sets, for yeast and human (Fig. S5).

We compared the average stickiness (S) of IDRs to their abundance (A) in the cell and observed a negative correlation, consistent with a recent report in yeast [51] and with a similar compositional bias seen among structured domains [78]. The S/A correlation appears stronger in yeast ($R = -0.36$, Fig. 2b) than in human ($R = -0.21$, Fig. S3b), and is significant in both species (yeast: $p = 5.2 \times 10^{-68}$ human: $p = 1.7 \times 10^{-75}$, Spearman's rank correlation test). The lower correlation seen in human may reflect the ambiguous nature of protein abundance data in multicellular organisms, where it is averaged over tissues. The S/A correlation increased substantially when considering the "median" and "high" disorder-content data sets, to $R = -0.41$ and $R = -0.55$ in yeast (Fig. 2b), and $R = -0.25$ and

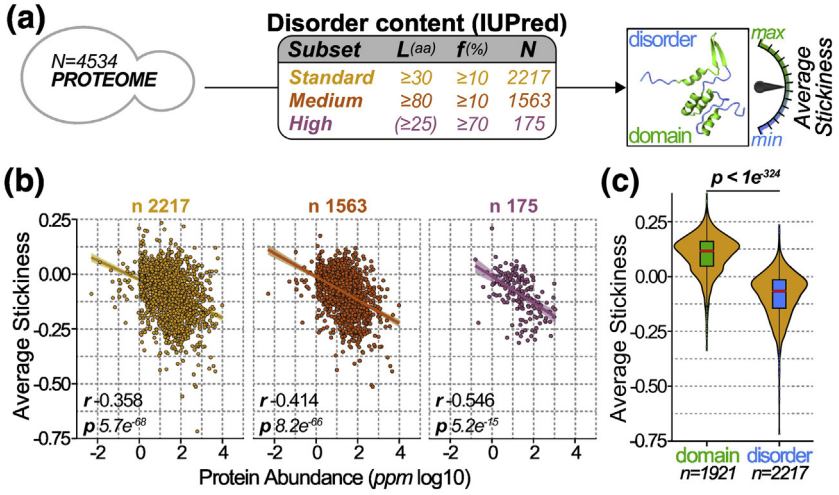

Fig. 2. The stickiness of disordered regions is anti-correlated with protein abundance. (a) We defined three data sets of proteins with increasing disorder content, referred to as "standard," "medium," "high." Proteins were included or excluded from each data set depending on their disorder content (L) and disorder fraction (f), as indicated in the panel table. The average stickiness of a protein was calculated by the mean score of all its amino acids in disordered regions. (b) Average protein stickiness (y-axis) as a function of protein abundance (x-axis) for disordered regions among the three data sets defined in panel a. The Spearman rank correlation coefficients (r) and the p values (p, Spearman's rank correlation test) are indicated. The number of proteins (n) within each data set is given above each scatterplot. (c) Distribution of average protein stickiness (y-axis) calculated based on disordered regions (blue), or, for reference, calculated on domains (green) for the same proteins from the standard data set. Boxes correspond to 50% of the probability density around the median (red line). The p value indicates a significant difference between IDR-stickiness and domains stickiness (one-sided Wilcoxon signed-rank test).

$R = -0.33$ in human (Fig. S3b). Furthermore, if we consider an alternative data set of regions not matched by a known domain (we call this set of regions in protein sequences "anti-domains"), the *S/A* correlation increases further, reaching -0.58 for 1453 yeast proteins (Fig. S6). Finally, we saw a similar correlation ($R = -0.37$, Fig. S7a) when using disorder predictions from the D^2P^2 database [88], which provides a consensus of several methods (PONDR, PrDOS, PV2, IUPred/ANCHOR, Espritz) [2,84,89–92]. We also saw similar correlations when separating proteins in three groups according to their size (Fig. S7).

For all three disorder-content data sets, the decrease in IDR-stickiness with increasing protein abundance is highlighted with regression lines (Figs. 2b and S3b). The regression shows that IDR stickiness for low-abundance proteins (<0.1 ppm) is above 0.0, and it decreases down to -0.2 for proteins with the highest abundance. For reference, we compare this stickiness difference to that seen between structured domains (median, 0.12) and disordered regions (median, -0.07; Fig. 2c), revealing a similar amplitude and stressing that the compositional bias we report is substantial.

Together, these results indicate that disordered regions present in high-abundance proteins exhibit biased compositions, whereby the fraction of their sticky amino acids is minimized. This observation supports the paradigm of a concentration-dependent optimization of amino acid composition in disordered regions so as to minimize dysfunctional interactions.

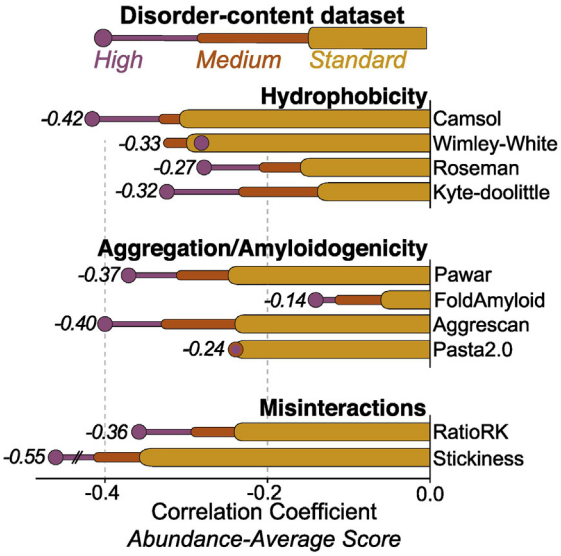

Fig. 3. Impact of abundance on several predicted properties of disordered regions in yeast. We calculated several properties of disordered regions associated with solubility and report their Spearman rank correlation with protein abundance. Colored bars show the correlation coefficients obtained using the same three sets of proteins defined in Fig. 2a: standard (orange), medium (dark orange), and high disorder-content data sets (purple).

Hydrophobicity and β-amyloid propensity exhibit weaker concentration dependence

We explored alternative amino-acid interaction rules to explain the concentration-dependent adaptation in amino-acid composition. We tested whether hydrophobicity was at the origin of the correlations observed. Three distinct hydrophobicity scales (Wimley and White [93], Kyte and Doolittle [94], and Roseman [95]) as well as a sequence-based predictor of protein solubility (CamSol [96]) gave significant *S/A* correlations, albeit of lower magnitude than that observed with the stickiness scale (Figs. 3 and S8). We reached similar conclusions with β-amyloid and aggregation propensity predictors, based on FoldAmyloid [97], Aggrescan [67], Pawar *et al*. [98], and PASTA 2.0 [99].

A notable feature of the stickiness scale compared to other scales is the fact that it discriminates lysine (K) and arginine (R) as amino acids with a large difference in stickiness (Fig. S9). While lysine is the most under-represented amino acid at protein–protein interfaces (and thus the least sticky), arginine is found with nearly equal frequency at interfaces and surfaces [47], making it a significantly stickier amino acid than lysine. In line with this notion, Warwicker *et al*. [100] observed that high lysine and low arginine content correlate with protein solubility. This led us to define a minimal stickiness measure based solely on these two amino acids, the RK-ratio ($R/(R+K)$), which is the number of arginines over the number of both lysines and arginines in disordered regions of a particular protein. The RK-ratio yielded *S/A* correlation coefficients nearly as high or higher than those obtained using other scales both in yeast ($R = -0.24$) and human ($R = -0.12$), for the standard data set. For the high disorder-content data set, these values reached $R = -0.36$ in yeast (Fig. 3) and $R = -0.2$ in human (Fig. S8). The fact that arginine and lysine are considered largely equivalent by other scales (Fig. S9) and that arginine is the most hydrophilic amino acid according to hydrophobicity scales highlights the distinct nature of the stickiness scale and further suggests that promiscuous interactions are selected against among abundant proteins.

Local *versus* global optimization of stickiness

We observed that chemical properties of disordered regions change as a function of their concentration, with interaction-resilient amino acids being enriched in high-abundance proteins. This observation prompted us to ask whether stickiness must be optimized throughout the sequence, or whether compensation

mechanisms can occur between different regions, whereby a non-sticky region in one part of the protein may compensate for a sticky one in a distal part and increase protein solubility. To address this question, we defined the local stickiness ω^k as the average stickiness score in a window of *k* residues sliding through each position *i* of a protein sequence (Fig. 4a). For each protein, we recorded the maximum local stickiness value as $\Omega^k = \max(\omega^k)$ (see Methods for details). Then, we calculated the Spearman rank correlation between protein abundance and Ω^k (Fig. 4b). We reasoned that the window size maximizing the correlation should reflect the distance in the polypeptide chain over which a highly sticky region can be compensated by a distal non-sticky region. We observed that larger windows improve the abundance-stickiness correlation up to a size of ~200 residues for the standard data set and ~800 residues for the high disorder-content data set, beyond which the correlation stabilizes (Figs. 4b and S10). It is difficult to ascribe a meaning to the actual size of these windows, notably because IDRs that are far in sequence may be close in space, for example, due to the protein's structure, or due to conformational sampling [101]. However, the fact that the correlation increases with *k* exceeding several hundred residues does suggest the existence of compensatory effects, whereby a sticky segment in one part of a protein can be tolerated if a non-sticky segment elsewhere in the sequence can compensate for it.

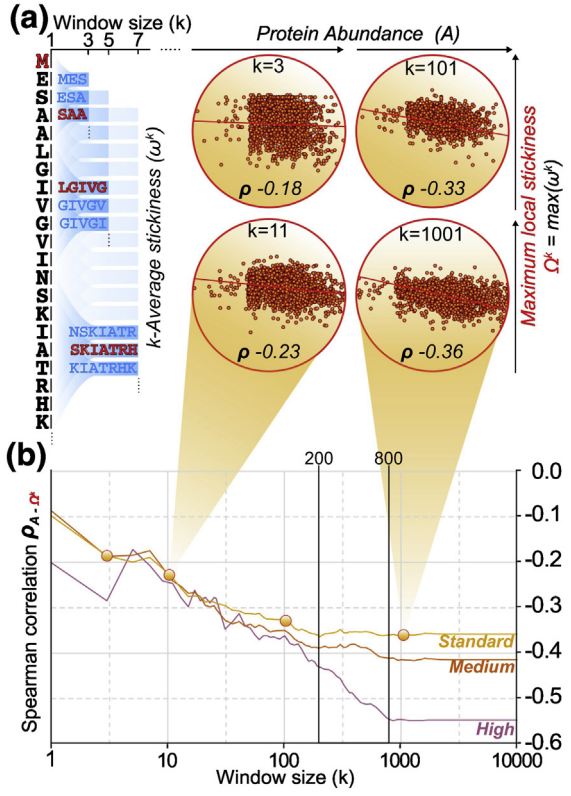

Fig. 4. Window size over which protein stickiness best correlates with protein abundance. (a) The window-averaged, local stickiness (ω^k) is calculated for windows of increasing size (*k*), represented here as funnels, sliding over the sequence of each protein. The local stickiness corresponds to the average of amino acid scores within a window of k residues (see Methods for details). For each window size, the maximum local average stickiness Ω^k is correlated against protein abundance. Scatterplots for selected values of *k* (3, 11, 101, 1001) are shown along with the corresponding Spearman rank correlation coefficient (ρ_k). (b) The Spearman rank correlation coefficient (ρ_k) is reported against increasing window sizes (*k*) considering residues within disordered regions of yeast proteins among the standard, medium, and high disorder-content data sets.

Heterogeneous distribution of stickiness among disordered segments

The long-range stickiness effects suggested by the above analysis led us to analyze stickiness homogeneity *versus* heterogeneity along the sequence. We introduced a measure $\Delta = \Omega^k - \Omega^L$, the difference between the maximal stickiness in a

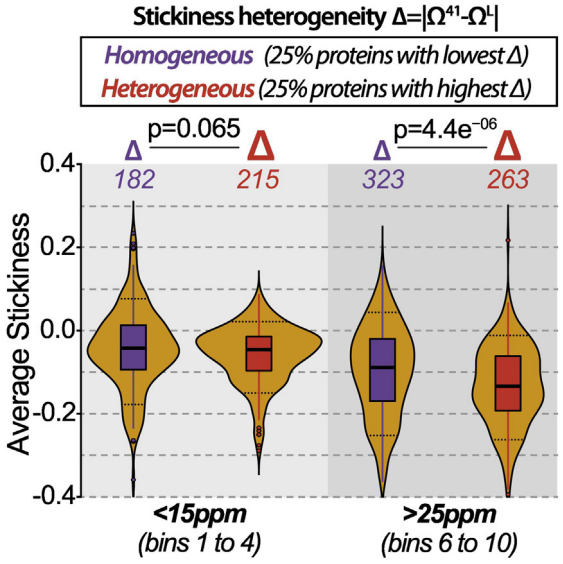

Fig. 5. Impact of stickiness heterogeneity on sequence adaptation to high abundance in yeast. Distributions of average stickiness (*y*-axis) within disordered regions of yeast proteins (standard data set) are displayed as violin shapes. Proteins with low (<15 ppm, bins 1−4) or high abundance (>25 ppm, bins 6−10) are further divided based on their stickiness heterogeneity (Δ). Stickiness heterogeneity is defined as the difference between the maximum local stickiness within a 41-residues window and the average IDR-stickiness of the protein. Homogeneous and heterogeneous proteins are those with the 25% lowest and highest Δ, respectively. *P* values are based on the Wilcoxon test. The number of proteins is given for each class.

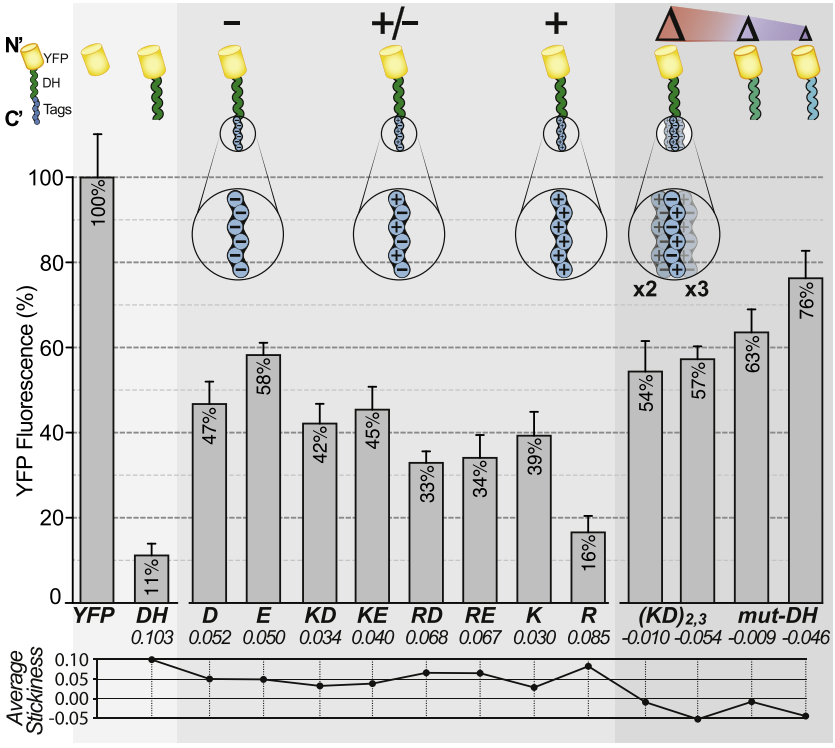

Fig. 6. Increasing the solubility of a protein using different designs of non-sticky fusion tags. Chimeric proteins were expressed in yeast cells, their abundance was measured by microscopy and is given relative to untagged YFP, which shows highest abundance. Cartoons above the bars depict the composition of the constructs, and the average stickiness calculated from each sequence (YFP excluded) is given underneath the bars (*cf*. Table S3 for sequences, strains and fluorescence data). The first construct corresponds to untagged YFP used as a reference for normalizing fluorescence levels. The second construct consists of YFP fused to the N-terminal fragment of mouse DHFR [104]) at its C-terminus. Fusion of YFP to this sticky polypeptide decreased fluorescence levels about 10-fold. Non-sticky tags were fused at the C-terminus of YFP-DH and contain charged amino acids. Fluorescence levels achieved with fusion-tags containing six charges are shown in the first set of bars. Longer fusion tags containing 12 or 18 charges were also used, denoted as KD_2 and KD_3. Finally, in the last two constructs (mut-DH), stickiness was decreased by introducing mutations throughout the DH fragment sequence, resulting in a more homogeneous stickiness profile.

window of size k and the average stickiness of the protein. A large Δ value indicates a heterogeneous stickiness across the protein, and conversely, a low value would be observed when the distribution of stickiness is homogeneous throughout the sequence. For this analysis, we arbitrarily used in both species a value k equal to 41 for the window size, but the results are robust against different values of k (Fig. S11).

We asked whether proteins with high *versus* low values of Δ show different average stickiness (Figs. 5 and S12). Among high-abundance proteins (bins 6–10), proteins with heterogeneous stickiness are less sticky on average when compared to proteins with homogeneous stickiness (yeast: $p = 4.4 \times 10^{-06}$, human: 2.3×10^{-29}, Wilcoxon one-sided test). This result could be influenced by the window size ($k = 41$) as well as by protein length, since long proteins are expected to show larger heterogeneity by chance alone, when compared to shorter proteins. Thus, we controlled for these two parameters (Figs. S11 and S13) and observed a consistently lower stickiness in the heterogeneous class when compared to the homogeneous class. For proteins with lower abundance, the average stickiness of the heterogeneous class did not show a consistent difference relative to the homogeneous class.

These results suggest that delocalized non-sticky regions are not as effective to promote solubility when compared to regions with homogeneous stickiness, among high-abundance proteins. Why different proteins employ different strategies to remain soluble remains to be investigated. We can speculate that functional constraints might require heterogeneous distribution of sticky residues among disordered regions. For example, the highly conserved protein PEX5 contains a long N-terminal IDR harboring several "WxxxF/Y" motifs that mediate recognition and import of proteins from the cytoplasm to the peroxisome [102]. Furthermore, random evolutionary processes are likely to result in heterogeneous stickiness patterns. Indeed, considering a

sticky and solvent-exposed region, there would be many more ways of compensating it with non-sticky residues elsewhere in the sequence than by mutating the region itself.

Experimentally assessing how distal non-sticky regions impact protein solubility

The results of the two previous sections led to three main conclusions. First, stickiness is a global sequence property that can involve compensatory effects in protein sequences. Second, the chemically analogous amino acids R and K exhibit markedly different stickiness properties. Third, homogeneous stickiness throughout the sequence is more effective at promoting solubility than heterogeneous stickiness.

To assess these concepts experimentally, we expressed chimeric fluorescent proteins in yeast cells and measured their solubility under the assumption that higher solubility promotes higher abundance in the cytosol. Maximal cytosolic abundance was observed for the wild-type Venus yellow fluorescent protein (YFP) [103], while minimal cytosolic abundance was observed for YFP fused to a sticky polypeptide derived from the mouse dihydrofolate reductase (DHFR) [104], thereafter referred to as YFP-DH (Fig. 6).

We subsequently aimed to increase the solubility of YFP-DH using different non-sticky fusion tags. First, we employed non-sticky tags composed of different combinations of positively and/or negatively charged amino acids fused at the C-terminus of YFP-DH. Overall, all non-sticky tags containing six charged amino acids increased the cytosolic abundance of YFP-DH, confirming that a non-sticky region can, at least partially, rescue from insolubility caused by a distal sticky region. Consistent with previous findings [105–108], negative charges promoted solubility better than positive charges, with tags based on six aspartate or glutamate yielding cytosolic abundances up to 1.5 times higher than a tag containing six lysines. The increased solubility seen with negatively charged amino acids is at odds with the stickiness scale, where lysine is the least sticky amino acid. This apparent discrepancy may originate in the avoidance of arginine at protein surfaces, leaving lysine as the only non-sticky positively charged amino acid [47]. In line with this idea, the six-arginine tag led to a 2-fold reduction in cytosolic abundance when compared to the six-lysine tag. Tags that were electrostatically neutral gave intermediate cytosolic abundance. Next, we tested the impact of stickiness heterogeneity on solubility. We created two additional non-sticky tags containing 12 and 18 charged residues. These tags decreased the average stickiness and introduced a high heterogeneity. In parallel, we designed two mutant sequences matching the average stickiness

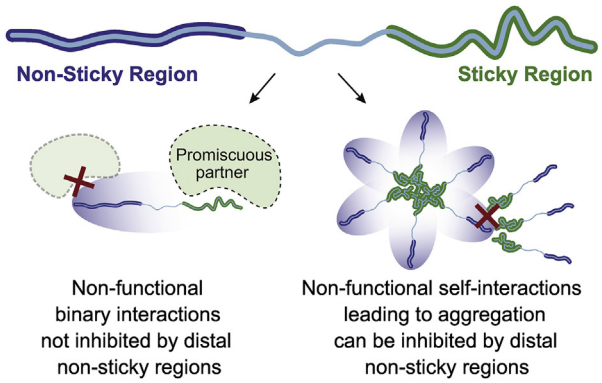

Fig. 7. Hypothetical impact of heterogeneous stickiness on binary interactions *versus* aggregation. A non-sticky region is not expected to shield a distal sticky region from promiscuous binary interactions (left). However, a non-sticky region may inhibit self-interactions leading to aggregation, if this process involves burying the non-sticky region in a desolvated environment (right).

obtained with 12 or 18 charged residues, although the mutations were introduced throughout the DH sequence to create a more homogeneous stickiness profile. The addition of 6 charged residues (from 6 to 12) increased the abundance of YFP-DH, but the further addition of charges (from 12 to 18) had little effect, suggesting that delocalized non-sticky regions cannot fully compensate for a distal sticky region. The strategy consisting of mutating sticky regions directly (e.g., mutations introduced in the DH sequence) showed comparatively higher abundance, consistent with homogeneous stickiness leading to higher solubility.

In our interpretation of these experiments, we hypothesize that fusion of the DH sequence to YFP induces its aggregation and degradation. Alternatively, the fusion of the DH sequence may slow down the translation of YFP-DH relative to YFP, reduce its mRNA stability, or decrease the average YFP fluorescence due to interference with the folding of YFP for example. However, it is unlikely that the diverse sequences of the non-sticky tags could consistently rescue from such effects, as their only common property is to be non-sticky. For this reason, the increase of YFP-DH solubility by the non-sticky tags is the most parsimonious explanation explaining their respective increase in fluorescence.

Conclusion

It has been observed previously that proteins containing disordered regions are constrained in terms of their regulation [48,57]. In addition to regulatory mechanisms, our results show that disordered regions tune their amino-acid composition to their cellular abundance. We observed this trend

globally, across the yeast and human proteomes, thus generalizing previous observations made on fewer proteins of known structures [78] and in a recent study [51]. Uniquely, our approach allowed comparing how amino-acid propensities associated with specific properties (hydrophobicity, amyloid formation, aggregation, and protein interactions) covaried with abundance. Interaction propensity or "stickiness" showed the strongest covariation, implying that avoidance of promiscuous interactions is a key property shaping protein sequences. Non-functional interactions can indeed be deleterious due to their potential to sequester proteins into aggregates of misfolded proteins [43,62,109–111] or agglomerates of folded proteins [58], and they may also compete with functional interactions [55,57,64,65,79,112]. The local optimization of stickiness in the sequence is expected to minimize deleterious effects associated with binary promiscuous interactions. In contrast, the fact that stickiness can also be optimized globally, across different segments separated by 200 or more amino acids, suggests an adaptation mechanism against promiscuous self-interactions leading to aggregation (Fig. 7). We experimentally validated this concept by rescuing an aggregation-prone protein-fragment after adding non-sticky residues delocalized relative to the sticky region.

In addition to the insights into sequence features preventing non-functional interactions, our work has several implications. Practically, our results imply that methods may under- and over-predict disorder in low and high-abundance proteins respectively. For example, IUPred predicts disordered regions as those lacking hydrophobic amino acids to stabilize a folded protein core. Such predictions might be improved further if taking into consideration protein abundance to model the expected hydrophobic and hydrophilic character of globular domains and disordered regions. Our findings may also have important evolutionary implications since protein abundance is known to be the main determinant of sequence conservation across species [113–119]. It will therefore be important to consider stickiness in comparative sequence analyses, for example, to examine patterns of hydrophobic amino acids [120], and more generally, to understand routes and mechanisms of adaptation to high cellular abundance.

Methods

Filtering proteins

Some proteins have a biased amino-acid composition due to their cellular function. In particular, proteins with transmembrane helices are enriched in hydrophobic residues but these are unlikely to become solvent-exposed. Hence, we discarded all membrane proteins from this study. In order to identify membrane proteins, we used predictions from TMHMM [121] and meta-predictions from TOPCONS [122–128]. A total of 1913 (29%) yeast and 7297 (36%) human proteins were discarded for which at least one transmembrane segment was predicted according to the consensus prediction from TOPCONS. In the absence of consensus, we relied on the agreement between predictions from TMHMM and three out of five predictors compiled in TOPCONS.

Protein abundance data

Protein abundances were obtained from Pax-Db (v4.0, May 2015) [129], which provides relative abundances for unicellular and multicellular organisms including tissue-specific data. For both organisms, we used overall abundance in the whole organism, inferred from all available data sets (integrated data set). The unit conversion from relative abundance in ppm to cellular concentration in molar was obtained by the following formula: $C = (k \cdot A) / N_A$ where $k \approx 3 \cdot 10^6$ proteins/fL as an estimate of cellular density of protein molecules per femtoliter (fL) obtained from [130], the Avogadro constant $N_A = 6.02 \times 10^{23}$ molecules/mol, and A is the abundance (e.g., 1 ppm yields $A = 10^{-6}$ and results in $C \approx 5$ nM). Table 1 provide the estimated range of cellular concentrations for the 10 bins of abundance.

Intrinsic disorder and domain predictions

We predicted disordered regions in both yeast and human proteomes using IUPred [84], combining short and long disorder predictions. In yeast, we selected the 20% amino acids with the highest disorder probabilities. The human proteome is known to exhibit higher disorder content [88] when compared to yeast, prompting us to use 40% of amino acids with the highest disorder probabilities.

To predict domains, we aligned profiles from Pfam-A (v27.0, May 2013) and Superfamily (v1.75, March 2013) to sequences of both proteomes and kept predictions with an E-value score below 10^{-3}.

Defining the three sets of disordered proteins used in the analyses

We defined three data sets containing proteins with increasing disorder content. We refer to these data sets as "standard," "medium," and "high." In yeast, the standard data set consists of 2217 proteins (7619 for human) where IDRs represent at least 10% (20% for human) of the sequence and contain at least 30 residues in total (50 for human). In yeast, the medium data set consists of 1563 proteins (3649 for human) where IDRs represent at least 10% (50% for human) of the sequence and contain at least 80 residues in total (50 in human). Finally, the

high disorder data set contains 175 yeast proteins (944 for human), with at least 70% (80% for human) of the sequence being predicted as disordered, and a minimum of 25 disordered residues per sequence (40 residues in human). These criteria are summarized in Fig. 2a (Fig. S3a for human).

Interface propensity (stickiness) scale

Amino-acid interface propensities were obtained from Levy *et al*. [78]. In this scale, amino-acid interface propensities were calculated by the log-ratio of the frequency at the interface-core *versus* at the surface, based on a non-redundant set of *Escherichia coli* proteins of known structure.

$$\text{Interface propensity }(\text{AA}) = log\left(\frac{\text{Freq}_{\text{AA}}^{\text{interface}}}{\text{Freq}_{\text{AA}}^{\text{surface}}}\right)$$

This scale (Fig. S4) measures the propensity for an amino acid to interact with other amino acids and thus can be viewed as a "stickiness" score, which serves to estimate the interaction propensity of disordered regions.

Other propensity scales

We employed three distinct amino-acid hydrophobicity scales: (1) Kyte and Doolittle [94] determined vacuum-to-water transfer free energies of amino acid side chain. (2) Roseman [95,131] estimated the free-energies of completely transferring the amino acids side-chain along with the polypeptide backbone from water to a hydrocarbon phase, correcting for self-solvation effects of polar residues induced by the proximity of flanking non-hydrogen bonded peptide bond. (3) Wimley and White [93] measured the transfer free energy for amino acids from water to the interface of a phosphatidylcholine membrane bilayer.

For CamSol [96], sequence-based predictor of protein solubility, we used opposite values (i.e., scores were multiplied by -1) to maintain a comparable ranking of amino acids with other scales.

We also considered amino-acid potentials derived from experimental data [132] for three methods that assess β-amyloid formation: Aggrescan [67], FoldAmyloid [97], and pH 7 scale from Pawar *et al*. [98].

Finally, we applied PASTA 2.0 [99] on protein sequence and kept the highest energy prediction for evaluating the stability of putative cross-beta pairings among disordered regions.

Calculation of global and local average stickiness of disordered regions

The stickiness of a sequence is calculated by the mean of its amino-acid stickiness scores (Fig. S4a). The global stickiness of IDRs in a protein is calculated using residues predicted as disordered, and amino acids outside of IDRs are ignored from the calculation. The local stickiness ω^k corresponds to the mean stickiness score calculated within a window of k residues sliding through each position *i* of a protein sequence. For each protein, we recorded the maximum local stickiness value as $\Omega^k = \max(\omega^k)$. Disordered regions are scattered along the sequence and can consist of few residues only. Thus, we imposed restrictions to avoid inferring a local stickiness value when the number of disordered residues was too small compared to the size of the window. Specifically, we inferred a local stickiness value only when the number of disordered residues in the window was equal or greater than the lowest value between:

- half the window size $h = (k + 1)/2$
- the size *D* of the largest disordered segment within the protein

Scanning windows of increasing size, ranging from one residue to the protein length *L* enabled us to assess distance-dependent fluctuations of the local maximum stickiness. For example, with a window consisting of a single amino acid ($k = 1$), Ω^1 corresponds to the most sticky amino acid of the protein considered. With a window of size three ($k = 3$), Ω^3 corresponds to the average stickiness of the most sticky stretch of three consecutive amino acids, and so on, and when the window size is the length of a protein *L*, Ω^L is the average stickiness.

Solubility tags and fusion proteins

Fusion proteins were expressed under the GPD promoter in the p413 plasmid [133]. As an unfolded and insoluble polypeptide chain we used the N-terminal fragment of the mouse DHFR enzyme. This protein fragment contains 108 residues out of 188 in the full-length protein and was fused at the C-terminus of the YFP. Various synthetic sequence tags were also fused at the C-terminus of the DHFR fragment. All cloning was carried out using the PIPE cloning method [134]. The different sequence tags consisted of stretches of charged amino acids separated by glycine and serine residues. All sequences can be found in Table S3. The mutations in DH were determined manually to decrease the stickiness of the fragment while matching the stickiness of the medium and long non-sticky tag variants. Synthetic DNA sequences were ordered from IDT and were cloned by PIPE at the C-terminus of YFP, to yield the "mut-DH" variants.

Imaging and fluorescence quantification

Imaging of cells and processing of the data was carried out as described in Ref. [135]. Briefly, cells were inoculated from their glycerol stock in 384-well glass-bottom optical plates (Matrical) with a pintool (FP1 pins, V&P Scientific) operated by a Tecan robot (Tecan Evo200 with MCA384 head). Cells were grown in YPD for a minimum of 10 h before they reached an optical density of at most 1, and were imaged with a confocal spinning disk microscope using a 60 × plan apo oil-immersion objective. Each image set was composed of two brightfield (BF) images (one in focus and one defocused to facilitate cell segmentation, each with 50-ms exposure) as well as one image quantifying fluorescence with 300-ms exposure. Individual cells were segmented from the brightfield images, and fluorescence intensity was estimated from the 50th quantile of pixel intensity within each segmented cell. For each strain, eight technical replicates were imaged, the median fluorescence across cells (typically over 100 per construct) was calculated, and the mean of the eight replicates and associated standard error are reported.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.08.008.

Acknowledgments

We thank Keith Dunker, Hagen Hofmann, Amnon Horovitz, Ben Lehner, Koby Levy, Tzachi Pilpel, Claus Wilke, and Shoshana Wodak for insightful discussions. We thank Shoshana Wodak and Mauricio Macossay-Castillo for helpful comments on the manuscript. We thank Joseph Georgeson for help with operating the microscope, and Harry Greenblatt for help with computer systems. This work was supported by the Israel Science Foundation (1452/18); by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 819318); by a research grant from A.-M. Boucher; and by research grants from the Estelle Funk Foundation, the Estate of Fannie Sherr, the Estate of Albert Delighter, the Merle S. Cahn Foundation, Mrs. Mildred S. Gosden, the Estate of Elizabeth Wachsman, and the Arnold Bortman Family Foundation. E.D.L. is incumbent of the Recanati Career Development Chair of Cancer Research.

Authors Contributions: B.D. and E.D.L. designed the analyses and experiments. B.D. carried out the computational analyses. B.D. and E.D.L. analyzed the data. O.M. created the yeast strains and carried out the microscopy screen. B.D. analyzed the microscopy data. B.D. and E.D.L. wrote the manuscript.

Received 7 May 2019;
Received in revised form 7 August 2019;
Accepted 10 August 2019
Available online 20 August 2019

Keywords:
intrinsic disorder;
disordered regions;
non-functional interactions;
protein abundance;
aggregation

References

[1] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2012) 137–149.
[2] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins. 42 (2001) 38–48.
[3] V.N. Uversky, J.R. Gillespie, A. Fink, Why Are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: Struct. Funct. Bioinf. 41 (2000) 415–427.
[4] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.
[5] N.P. Pavletich, C.O. Pabo, Zinc finger-DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 A, Science 252 (1991) 809–817.
[6] A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva, Z. Obradović, Intrinsic disorder and protein function, Biochemistry. 41 (2002) 6573–6582.
[7] P. Tompa, The interplay between structure and function in intrinsically unstructured proteins, FEBS Lett. 579 (2005) 3346–3354.
[8] R. Van Der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, Others, classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631.
[9] O. Schueler-Furman, S.J. Wodak, Computational approaches to investigating allostery, Curr. Opin. Struct. Biol. 41 (2016) 159–171.
[10] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, Nat. Rev. Mol. Cell Biol. 6 (2005) 197–208.
[11] J.D. Forman-Kay, T. Mittag, From sequence and forces to structure, function, and evolution of intrinsically disordered proteins, Structure. 21 (2013) 1492–1499.
[12] S.S. Patel, B.J. Belmont, J.M. Sante, M.F. Rexach, Natively unfolded nucleoporins gate protein diffusion across the nuclear pore complex, Cell. 129 (2007) 83–96.

[13] M. Domanski, M. Hertzog, J. Coutant, I. Gutsche-Perelroizen, F. Bontems, M.-F. Carlier, E. Guittet, C. van Heijenoort, Coupling of folding and binding of thymosin beta4 upon interaction with monomeric actin monitored by nuclear magnetic resonance, J. Biol. Chem. 279 (2004) 23637–23645.
[14] H.J. Dyson, P.E. Wright, Coupling of folding and binding for unstructured proteins, Curr. Opin. Struct. Biol. 12 (2002) 54–60.
[15] B. Mészáros, P. Tompa, I. Simon, Z. Dosztányi, Molecular principles of the interactions of disordered proteins, J. Mol. Biol. 372 (2007) 549–561.
[16] R. Pancsa, M. Fuxreiter, Interactions via intrinsically disordered regions: what kind of motifs? IUBMB Life 64 (2012) 513–520.
[17] A. Stein, R. Mosca, P. Aloy, Three-dimensional modeling of protein interactions and complexes is going 'omics, Curr. Opin. Struct. Biol. 21 (2011) 200–208.
[18] N.E. Davey, K. Van Roey, R.J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, T.J. Gibson, Attributes of short linear motifs, Mol. BioSyst. 8 (2012) 268–281.
[19] H. Dinkel, S. Michael, R.J. Weatheritt, N.E. Davey, K. Van Roey, B. Altenberg, G. Toedt, B. Uyar, M. Seiler, A. Budd, L. Jödicke, M.A. Dammert, C. Schroeter, M. Hammer, T. Schmidt, P. Jehl, C. McGuigan, M. Dymecka, C. Chica, K. Luck, A. Via, A. Chatr-aryamontri, N. Haslam, G. Grebnev, R.J. Edwards, M.O. Steinmetz, H. Meiselbach, F. Diella, T.J. Gibson, ELM—the database of eukaryotic linear motifs, Nucleic Acids Res. 40 (2012) D242–D251.
[20] H. Hu, J. Columbus, Y. Zhang, D. Wu, L. Lian, S. Yang, J. Goodwin, C. Luczak, M. Carter, L. Chen, M. James, R. Davis, M. Sudol, J. Rodwell, J.J. Herrero, A map of WW domain family interactions, Proteomics. 4 (2004) 643–655.
[21] Y. Ivarsson, Plasticity of PDZ domains in ligand recognition and signaling, FEBS Lett. 586 (2012) 2638–2647.
[22] B. Mészáros, Z. Dosztányi, I. Simon, Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition, PLoS One 7 (2012), e46829.
[23] K. Van Roey, B. Uyar, R.J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T.J. Gibson, N.E. Davey, Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation, Chem. Rev. 114 (2014) 6733–6778.
[24] X. Xin, D. Gfeller, J. Cheng, R. Tonikian, L. Sun, A. Guo, L. Lopez, A. Pavlenco, A. Akintobi, Y. Zhang, J.-F. Rual, B. Currell, S. Seshagiri, T. Hao, X. Yang, Y.A. Shen, K. Salehi-Ashtiani, J. Li, A.T. Cheng, D. Bouamalay, A. Lugari, D.E. Hill, M.L. Grimes, D.G. Drubin, B.D. Grant, M. Vidal, C. Boone, S.S. Sidhu, G.D. Bader, SH3 interactome conserves general function over specific form, Mol. Syst. Biol. 9 (2013) 652.
[25] A. Kelil, E.D. Levy, S.W. Michnick, Evolution of domain–peptide interactions to coadapt specificity and affinity to functional diversity, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) E3862–E3871.
[26] D.E. Koshland, The key–lock theory and the induced fit theory, Angew. Chem. Int. Ed Engl. 33 (1995) 2375–2378.
[27] D.E. Koshland, Application of a theory of enzyme specificity to protein synthesis, Proc. Natl. Acad. Sci. U. S. A. 44 (1958) 98–104.
[28] V.N. Uversky, The multifaceted roles of intrinsic disorder in protein complexes, FEBS Lett. 589 (2015) 2498–2506.
[29] C.J. Oldfield, J. Meng, J.Y. Yang, M.Q. Yang, V.N. Uversky, A.K. Dunker, Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, BMC Genomics 9 (2008) S1.
[30] W.-L. Hsu, C. Oldfield, J. Meng, F. Huang, B. Xue, V.N. Uversky, P. Romero, A.K. Dunker, Intrinsic protein disorder and protein–protein interactions, Pac. Symp. Biocomput. (2012) 116–127.
[31] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N.P. Brown, G. Trave, T.J. Gibson, Understanding eukaryotic linear motifs and their role in cell signaling and regulation, Front. Biosci. 13 (2008) 6580–6603.
[32] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.
[33] R.B. Berlow, H.J. Dyson, P.E. Wright, Hypersensitive termination of the hypoxic response by a disordered protein switch, Nature. 543 (2017) 447–451.
[34] R.P. Bhattacharyya, A. Reményi, B.J. Yeh, W.A. Lim, Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits, Annu. Rev. Biochem. 75 (2006) 655–680.
[35] R. Mosca, R.A. Pache, P. Aloy, The role of structural disorder in the rewiring of protein interactions through evolution, Mol. Cell. Proteomics 11 (2012). M111.014969.
[36] P. Beltrao, L. Serrano, Specificity and evolvability in eukaryotic protein interaction networks, PLoS Comput. Biol. 3 (2007), e25.
[37] M. Buljan, G. Chalancon, S. Eustermann, G.P. Wagner, M. Fuxreiter, A. Bateman, M.M. Babu, Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks, Mol. Cell 46 (2012) 871–883.
[38] S. Pechmann, J. Frydman, Interplay between chaperones and protein disorder promotes the evolution of protein networks, PLoS Comput. Biol. 10 (2014), e1003674.
[39] D. Ekman, S. Light, A.K. Björklund, A. Elofsson, What properties characterize the hub proteins of the protein–protein interaction network of Saccharomyces cerevisiae? Genome Biol. 7 (2006) R45.
[40] K. Van Roey, T.J. Gibson, N.E. Davey, Motif switches: decision-making in cell regulation, Curr. Opin. Struct. Biol. 22 (2012) 378–385.
[41] V.N. Uversky, Intrinsic disorder, protein–protein interactions, and disease, in: Advances in Protein Chemistry and Structural Biology, Elsevier, 2018, pp. 85–121.
[42] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions, Trends Biochem. Sci. 33 (2008) 2–8.
[43] E.M. Marcotte, M. Tsechansky, Disorder, promiscuity, and toxic partnerships, Cell. 138 (2009) 16–18.
[44] T. Vavouri, J.I. Semple, R. Garcia-Verdugo, B. Lehner, Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity, Cell. 138 (2009) 198–208.
[45] S. Pechmann, E.D. Levy, G.G. Tartaglia, M. Vendruscolo, Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 10159–10164.
[46] H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E.D. Levy, Proteins evolve on the edge of supramolecular self-assembly, Nature. 548 (2017) 244–247.
[47] E.D. Levy, A simple definition of structural regions in proteins and its use in analyzing interface evolution, J. Mol. Biol. 403 (2010) 660–670.

[48] J. Gsponer, M.E. Futschik, S.A. Teichmann, M.M. Babu, Tight regulation of unstructured proteins: from transcript synthesis to protein degradation, Science. 322 (2008) 1365–1368.
[49] E.D. Levy, J. Kowarzyk, S.W. Michnick, High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation, Cell Rep. 7 (2014) 1333–1340.
[50] C.R. Landry, E.D. Levy, D. Abd Rabbo, K. Tarassov, S.W. Michnick, Extracting insight from noisy cellular networks, Cell. 155 (2013) 983–989.
[51] M. Macossay-Castillo, G. Marvelli, M. Guharoy, A. Jain, D. Kihara, P. Tompa, S.J. Wodak, The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity, J. Mol. Biol. (2019).
[52] J. Zhang, S. Maslov, E.I. Shakhnovich, Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size, Mol. Syst. Biol. 4 (2008) 210.
[53] M. Heo, S. Maslov, E. Shakhnovich, Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 4258–4263.
[54] M.E. Johnson, G. Hummer, Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 603–608.
[55] E.D. Levy, C.R. Landry, S.W. Michnick, How perfect can protein interactomes be? Sci. Signal. 2 (2009), e11.
[56] J.-R. Yang, J.-R. Yang, B.-Y. Liao, S.-M. Zhuang, J. Zhang, Protein misinteraction avoidance causes highly expressed proteins to evolve slowly, Proc. Natl. Acad. Sci. 109 (2012) E831–E840.
[57] J. Gsponer, M.M. Babu, Cellular strategies for regulating functional and nonfunctional protein aggregation, Cell Rep. 2 (2012) 1425–1437.
[58] H. Garcia-Seisdedos, J.A. Villegas, E.D. Levy, Infinite assembly of folded proteins in evolution, disease, and engineering, Angew. Chem. Int. Ed Engl. (2018).
[59] G. Schreiber, A.E. Keating, Protein binding specificity versus promiscuity, Curr. Opin. Struct. Biol. 21 (2011) 50–61.
[60] H. Schweke, M.H. Mucchielli, S. Sacquin-Mora, W. Bei, Protein interaction energy landscapes are shaped by functional and also non-functional partners, bioRxiv (2018) [Preprint].
[61] K. Tomala, R. Korona, Evaluating the fitness cost of protein expression in Saccharomyces cerevisiae, Genome Biol. Evol. 5 (2013) 2051–2060.
[62] B. Bolognesi, N. Lorenzo Gotor, R. Dhar, D. Cirillo, M. Baldrighi, G.G. Tartaglia, B. Lehner, A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression, Cell Rep. 16 (2016) 222–231.
[63] T. Hagai, A. Azia, M.M. Babu, R. Andino, Use of host-like peptide motifs in viral proteins is a prevalent strategy in host–virus interactions, Cell Rep. 7 (2014) 1729–1739.
[64] M.M. Babu, R. van der Lee, N.S. de Groot, J. Gsponer, Intrinsically disordered proteins: regulation and disease, Curr. Opin. Struct. Biol. 21 (2011) 432–440.
[65] N.E. Davey, G. Travé, T.J. Gibson, How viruses hijack cell regulation, Trends Biochem. Sci. 36 (2011) 159–169.
[66] W.E. Balch, R.I. Morimoto, A. Dillin, J.W. Kelly, Adapting proteostasis for disease intervention, Science. 319 (2008) 916–919.
[67] N. Sanchez de Groot, M. Torrent, A. Villar-Piqué, B. Lang, S. Ventura, J. Gsponer, M.M. Babu, Evolutionary selection for protein aggregation, Biochem. Soc. Trans. 40 (2012) 1032–1037.
[68] C.M. Dobson, Protein folding and misfolding, Nature. 426 (2003) 884–890.
[69] K.A. Geiler-Samerotte, M.F. Dion, B.A. Budnik, S.M. Wang, D.L. Hartl, D.A. Drummond, Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 680–685.
[70] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C.M. Dobson, M. Stefani, Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases, Nature. 416 (2002) 507–511.
[71] P. Sormanni, M. Vendruscolo, Protein solubility predictions using the CamSol method in the study of protein homeostasis, Cold Spring Harb. Perspect. Biol (2019).
[72] P. Ciryam, G.G. Tartaglia, R.I. Morimoto, C.M. Dobson, M. Vendruscolo, Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins, Cell Rep. 5 (2013) 781–790.
[73] E.D. Levy, S.W. Michnick, C.R. Landry, Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information, Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 367 (2012) 2594–2606.
[74] R.F. Albu, G.T. Chan, M. Zhu, E.T.C. Wong, F. Taghizadeh, X. Hu, A.E. Mehran, J.D. Johnson, J. Gsponer, T. Mayor, A feature analysis of lower solubility proteins in three eukaryotic systems, J. Proteome 118 (2015) 21–38.
[75] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, C.M. Dobson, Rationalization of the effects of mutations on peptide and protein aggregation rates, Nature. 424 (2003) 805–808.
[76] G.G. Tartaglia, S. Pechmann, C.M. Dobson, M. Vendruscolo, A relationship between mRNA expression levels and protein solubility in E. coli, J. Mol. Biol. 388 (2009) 381–389.
[77] S. Ventura, Sequence determinants of protein aggregation: tools to increase protein solubility, Microb. Cell Factories 4 (2005) 11.
[78] E.D. Levy, S. De, S.A. Teichmann, Cellular crowding imposes global constraints on the chemistry and evolution of proteomes, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 20461–20466.
[79] K. Meyer, M. Kirchner, B. Uyar, J.-Y. Cheng, G. Russo, L.R. Hernandez-Miranda, A. Szymborska, H. Zauber, I.-M. Rudolph, T.E. Willnow, A. Akalin, V. Haucke, H. Gerhardt, C. Birchmeier, R. Kühn, M. Krauss, S. Diecke, J.M. Pascual, M. Selbach, Mutations in disordered regions can cause disease by creating dileucine motifs, Cell 175 (2018) 239–253.e17.
[80] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T.J. Gibson, J. Lewis, L. Serrano, R.B. Russell, Systematic discovery of new recognition peptides mediating protein interaction networks, PLoS Biol. 3 (2005), e405.
[81] J.T. Nielsen, F.A.A. Mulder, Quality and bias of protein disorder predictors, Sci. Rep 9 (2019).
[82] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn, The Pfam protein families database in 2019, Nucleic Acids Res. 47 (2019) D427–D432.
[83] A.P. Pandurangan, J. Stahlhacke, M.E. Oates, B. Smithers, J. Gough, The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver, Nucleic Acids Res 47 (2019) D490–D494.

[84] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, J. Mol. Biol. 347 (2005) 827–839.
[85] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L.L. Sonnhammer, The Pfam protein families database, Nucleic Acids Res. 30 (2002) 276–280.
[86] J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure1, J. Mol. Biol. 313 (2001) 903–919.
[87] E.T.C. Wong, D. Na, J. Gsponer, On the importance of polar interactions for complexes containing intrinsically disordered proteins, PLoS Comput. Biol. 9 (2013), e1003192.
[88] M.E. Oates, P. Romero, T. Ishida, M. Ghalwash, M.J. Mizianty, B. Xue, Z. Dosztányi, V.N. Uversky, Z. Obradovic, L. Kurgan, A.K. Dunker, J. Gough, D2P2: database of disordered protein predictions, Nucleic Acids Res. 41 (2013) D508–D516.
[89] X. Li, P. Romero, M. Rani, A.K. Dunker, Z. Obradovic, Predicting protein disorder for N-, C-, and internal regions, Genome Inform. Ser. Workshop Genome Inform. 10 (1999) 30–40.
[90] T. Ishida, K. Kinoshita, PrDOS: prediction of disordered protein regions from amino acid sequence, Nucleic Acids Res. 35 (2007) W460–W464.
[91] M.F. Ghalwash, A.K. Dunker, Z. Obradović, Uncertainty analysis in protein disorder prediction, Mol. BioSyst. 8 (2012) 381–391.
[92] I. Walsh, A.J.M. Martin, T. Di Domenico, S.C.E. Tosatto, ESpritz: accurate and fast prediction of protein disorder, Bioinformatics. 28 (2012) 503–509.
[93] W.C. Wimley, S.H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, Nat. Struct. Biol. 3 (1996) 842–848.
[94] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[95] M.A. Roseman, Hydrophobicity of the peptide C=O…H–N hydrogen-bonded group, J. Mol. Biol. 201 (1988) 621–623.
[96] P. Sormanni, F.A. Aprile, M. Vendruscolo, The CamSol method of rational design of protein mutants with enhanced solubility, J. Mol. Biol. 427 (2015) 478–490.
[97] S.O. Garbuzynskiy, M.Y. Lobanov, O.V. Galzitskaya, FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence, Bioinformatics. 26 (2010) 326–332.
[98] A.P. Pawar, K.F. DuBay, J. Zurdo, F. Chiti, M. Vendruscolo, C.M. Dobson, Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases, J. Mol. Biol. 350 (2005) 379–392.
[99] I. Walsh, F. Seno, S.C.E. Tosatto, A. Trovato, PASTA 2.0: an improved server for protein aggregation prediction, Nucleic Acids Res 42 (2014) W301–W307.
[100] J. Warwicker, S. Charonis, R.A. Curtis, Lysine and arginine content of proteins: computational analysis suggests a new tool for solubility design, Mol. Pharm. 11 (2014) 294–303.
[101] R.K. Das, R.V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 13392–13397.
[102] A. Barros-Barbosa, T.A. Rodrigues, M.J. Ferreira, A.G. Pedrosa, N.R. Teixeira, T. Francisco, J.E. Azevedo, The intrinsically disordered nature of the peroxisomal protein translocation machinery, FEBS J. 286 (2019) 24–38.
[103] T. Nagai, K. Ibata, E.S. Park, M. Kubota, K. Mikoshiba, A. Miyawaki, A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications, Nat. Biotechnol. 20 (2002) 87–90.
[104] I. Remy, S.W. Michnick, Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 5394–5399.
[105] A.M. Ruschak, J.D. Rose, M.P. Coughlin, T.L. Religa, Engineered solubility tag for solution NMR of proteins, Protein Sci. 22 (2013) 1646–1654.
[106] R.M. Kramer, V.R. Shende, N. Motl, C.N. Pace, J.M. Scholtz, Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility, Biophys. J. 102 (2012) 1907–1915.
[107] S.R. Trevino, J.M. Scholtz, C.N. Pace, Measuring and increasing protein solubility, J. Pharm. Sci. 97 (2008) 4155–4166.
[108] S.R. Trevino, J.M. Scholtz, C.N. Pace, Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa, J. Mol. Biol. 366 (2007) 449–460.
[109] A. Mogk, B. Bukau, H.H. Kampinga, Cellular handling of protein aggregates by disaggregation machines, Mol. Cell 69 (2018) 214–226.
[110] A. Cumberworth, G. Lamour, M.M. Babu, J. Gsponer, Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes, Biochem. J. 454 (2013) 361–369.
[111] R. Gemayel, S. Chavali, K. Pougach, M. Legendre, B. Zhu, S. Boeynaems, E. van der Zande, K. Gevaert, F. Rousseau, J. Schymkowitz, M.M. Babu, K.J. Verstrepen, Variable glutamine-rich repeats modulate transcription factor activity, Mol. Cell 59 (2015) 615–627.
[112] H. Olzscha, S.M. Schermann, A.C. Woerner, S. Pinkert, M.H. Hecht, G.G. Tartaglia, M. Vendruscolo, M. Hayer-Hartl, F.U. Hartl, R.M. Vabulas, Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions, Cell. 144 (2011) 67–78.
[113] D.A. Drummond, J.D. Bloom, C. Adami, C.O. Wilke, F.H. Arnold, Why highly expressed proteins evolve slowly, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 14338–14343.
[114] C. Pal, B. Papp, L.D. Hurst, Highly expressed genes in yeast evolve slowly, Genetics. 158 (2001) 927–931.
[115] E.P. Rocha, A. Danchin, An analysis of determinants of amino acids substitution rates in bacterial proteins, Mol. Biol. Evol. 21 (2004) 108–116.
[116] D.A. Drummond, A. Raval, C.O. Wilke, A single determinant dominates the rate of yeast protein evolution, Mol. Biol. Evol. 23 (2006) 327–337.
[117] M. Kimura, Evolutionary rate at the molecular level, Nature. 217 (1968) 624–626.
[118] D.M. Krylov, Y.I. Wolf, I.B. Rogozin, E.V. Koonin, Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution, Genome Res. 13 (2003) 2229–2235.

[119] J. Zhang, J.R. Yang, Determinants of the rate of protein sequence evolution, Nat. Rev. Genet. 16 (2015) 409–420.
[120] S.G. Foy, B.A. Wilson, J. Bertram, M.H.J. Cordes, J. Masel, A shift in aggregation avoidance strategy Marks a long-term direction to protein evolution, Genetics. 211 (2019) 1345–1355.
[121] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[122] A. Bernsel, H. Viklund, A. Hennerdal, A. Elofsson, TOPCONS: consensus prediction of membrane protein topology, Nucleic Acids Res. 37 (2009) W465–W468.
[123] H. Viklund, A. Elofsson, OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, Bioinformatics. 24 (2008) 1662–1668.
[124] L. Käll, A. Krogh, E.L.L. Sonnhammer, An HMM posterior decoder for sequence feature prediction that includes homology information, Bioinformatics. 21 (Suppl. 1) (2005) i251–i257.
[125] S.M. Reynolds, L. Käll, M.E. Riffle, J.A. Bilmes, W.S. Noble, Transmembrane topology and signal peptide prediction using dynamic bayesian networks, PLoS Comput. Biol. 4 (2008), e1000213.
[126] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, A. Elofsson, Prediction of membrane-protein topology from first principles, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 7177–7181.
[127] H. Viklund, A. Bernsel, M. Skwark, A. Elofsson, SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology, Bioinformatics. 24 (2008) 2928–2929.
[128] T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, Nature. 450 (2007) 1026–1030.
[129] M. Wang, C.J. Herrmann, M. Simonovic, D. Szklarczyk, C. von Mering, Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines, Proteomics (2015).
[130] R. Milo, What is the total number of protein molecules per cell volume? A call to rethink some published values, Bioessays. 35 (2013) 1050–1055.
[131] M.A. Roseman, Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds, J. Mol. Biol. 200 (1988) 513–522.
[132] A.B. Ahmed, A.V. Kajava, Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence, FEBS Lett 587 (2013) 1089–1095.
[133] D. Mumberg, R. Muller, M. Funk, Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds, Gene. 156 (1995) 119–122.
[134] H.E. Klock, E.J. Koesema, M.W. Knuth, S.A. Lesley, Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts, Proteins. 71 (2008) 982–994.
[135] O. Matalon, A. Steinberg, E. Sass, J. Hausser, E.D. Levy, Reprogramming protein abundance fluctuations in single cells by degradation, bioRxiv (2018) [Preprint].