Cellular and Molecular Life Sciences
https://doi.org/10.1007/s00018-019-03166-6

ORIGINAL ARTICLE

Codon selection reduces GC content bias in nucleic acids encoding for intrinsically disordered proteins

Christopher J. Oldfield[1] · Zhenling Peng[2] · Vladimir N. Uversky[3,4] · Lukasz Kurgan[1]

Received: 3 January 2019 / Revised: 14 May 2019 / Accepted: 28 May 2019

Abstract

Protein-coding nucleic acids exhibit composition and codon biases between sequences coding for intrinsically disordered regions (IDRs) and those coding for structured regions. IDRs are regions of proteins that are folding self-insufficient and which function without the prerequisite of folded structure. Several authors have investigated composition bias or codon selection in regions encoding for IDRs, primarily in Eukaryota, and concluded that elevated GC content is the result of the biased amino acid composition of IDRs. We substantively extend previous work by examining GC content in regions encoding IDRs, from 44 species in Eukaryota, Archaea, and Bacteria, spanning a wide range of GC content. We confirm that regions coding for IDRs show a significantly elevated GC content, even across all domains of life. Although this is largely attributable to the amino acid composition bias of IDRs, we show that this bias is independent of the overall GC content and, most importantly, we are the first to observe that GC content bias in IDRs is significantly different than expected from IDR amino acid composition alone. We empirically find compensatory codon selection that reduces the observed GC content bias in IDRs. This selection is dependent on the overall GC content of the organism. The codon selection bias manifests as use of infrequent, AT-rich codons in encoding IDRs. Further, we find these relationships to be independent of the intrinsic disorder prediction method used, and independent of estimated translation efficiency. These observations are consistent with the previous work, and we speculate on whether the observed biases are causal or symptomatic of other driving forces.

Keywords Intrinsically disordered proteins · Amino acid composition · GC content · Codon selection

Introduction

Intrinsically disordered regions (IDRs) are protein regions that are folding self-insufficient, having conformations that vary over time and over populations [1–3]. Despite this lack of stable structure, IDRs have been found to perform many important molecular functions across a wide range of biological processes [4–7]. Furthermore, IDRs are predicted to be quite frequent in nature [8–12]; 25–40% of proteins in eukaryotic organisms contain long intrinsically disordered region [8]. Although less frequent in Archaea and bacterial proteins, IDRs still comprise a significant portion of proteins in those domains of life [7, 9]. These estimates of IDR abundance have been made using algorithms that predict per-residues intrinsic disorder from amino acid sequence, e.g., [13–18]. In addition to distinct amino acid sequence signatures [15], IDRs have been found to have several genetic signatures in protein-coding nucleotide sequences. In particular, IDRs are closely associated with alternative splicing [19]—where splice sites are preferentially located

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00018-019-03166-6) contains supplementary material, which is available to authorized users.

✉ Christopher J. Oldfield
cjoldfield@vcu.edu

✉ Lukasz Kurgan
lkurgan@vcu.edu

[1] Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

[2] Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

[3] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

[4] Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

Published online: 07 June 2019

in IDRs, likely due to their structural permissiveness—and codon selection [20, 21]—where suboptimal codon usage slows translation, allowing proper folding and function of neighboring ordered regions [21]. Several studies have also examined coding guanine and cytosine (GC) content in relation to IDRs in eukaryotes [22–24] and prokaryotes [25].

Local GC content of DNA has many biological implications. For example, GC content is closely linked with mutability [26] and gene age [27]. For higher Eukaryota, genes are often found with GC-rich regions of DNA [26], and introns and exons display differences in GC content, in which relative GC enrichment of exons plays a role in regulation of alternative splicing (AS) [28]. Also among Eukaryota, studies have found that gene regions coding for IDRs have a high GC content, relative to gene regions coding for ordered regions [22, 23]. The relatively high GC content of IDRs is suggested to be due to the high GC content of codons of some disorder-promoting amino acids, and low GC content of some codons of order-promoting amino acids [22, 23]. It is particularly interesting that consistent relationships have been observed between AS, IDRs, and GC content: AS and IDRs [19], AS and GC content [28], and IDRs and GC content [22–24]. Additionally, IDRs have been linked to elevated recombination rates, likely due to the increased GC content of IDR coding regions [24]. The relationship between IDRs and GC content has been attributed to amino acid usage bias between structured and disordered regions [23]. In general, amino acids are biased between IDRs and ordered regions; disordered regions are depleted in hydrophobic residues and enriched in polar, charged residues, and proline [29]. In terms of the first two codon positions, the codons of several hydrophobic residues are depleted in GC, and charged residues and proline are enriched in GC. In this way, IDRs can influence the GC of its source coding sequence.

Several studies examined the relationship between IDRs and GC content of their coding regions within particular domains of life, only Eukaryota [22, 23] or only prokaryotes [25]. These studies have not attempted to explain this relationship beyond the simple influence of amino acid composition bias. For instance, the role of the wobble position in GC content in IDRs has been ignored. To explore the deeper relationship between GC and disorder content, we examine the observed GC content in IDRs relative to the GC expected from unbiased codon usage between IDRs and structured regions. We find that often the GC content of IDRs is significantly reduced relative to the GC content expected from amino acid composition alone. This phenomenon is dependent on the overall coding GC content of an organism. Moreover, this relationship is explored over a diverse and balanced set of organisms than previously reported. All domains of life—Eukaryota, Archaea, and Bacteria—are represented by the 44 selected species. We are also the first to ensure that these organisms provide a balanced sampling of the full range of coding GC content; an equal number of organisms with low, intermediate, and high coding GC content were used for each domain of life. Further, the relationship between GC and disorder content was examined at several levels: organism, protein, and residue. Finally, we examine the effects of estimated translation efficiency, as measured by biased usage of synonymous codons [30], which has the potential to effect GC content. To our knowledge, this is the broadest examination of the relationship between coding sequences and IDRs.

Materials and methods

Protein and coding sequences for 44 diverse species from all domains of life were collected, including: 15 Eukaryota, 14 Archaea, and 15 Bacteria. High-quality annotations of protein-coding regions in the human and mouse genomes were taken from the collaborative consensus coding sequence (CCDS) project [31]. Complete proteomes sequences for the other 42 species were collected from UniProt release 2013_11 [32]. The corresponding transcripts for each protein were retrieved from EMBL [33]. We aimed to include popular model organisms and to obtain a balanced sample of coding GC content, with an equal number of organisms with high, intermediate, and low values for each domain of life (Table 1). More specifically, for each domain of life, we include five organisms in the low, intermediate, and high GC groups that are characterized by the median per protein GC content between 25% and 40%, 45% and 55%, and 57% and 75%, respectively. We select at least 4 or 5 organisms for each group and domain of life to ensure that sample size is sufficient to run statistical tests.

Characterization of intrinsic disorder in protein sequences

Intrinsic disorder predictions were made using a consensus of five predictions generated by two popular methods: IUPred [13] and ESpritz [14]. We select them based on their favorable predictive quality [34, 35], runtime that is sufficiently fast to process the 44 genomes, and complementary designs. To the latter point, the consensus includes two versions of IUPred that specialize in prediction of long disordered regions (30 or more consecutive residues) and short disordered segments (typically present in structured, globular proteins) and three versions of ESpritz that were designed for three types of annotations of disordered residues: using crystal structures, nuclear magnetic resonance structures, and annotations from the DisProt database [36]. We implemented the consensus using the majority vote where three or more out of five

Table 1 Dataset summary

Domain	Species	NCBI taxonomy ID	Number of proteins	Median length	GC group	Per protein proportion coding GC (%)		Per protein proportion disordered residues (%)	
						Median	Quartiles [25,75]	Median	Quartiles [25,75]
Eukaryota	*A. anophagefferens*	44056	7382	452	High	71.7	[67.7,74.8]	9.3	[3.7,20.4]
	C. reinhardtii	3055	9419	341		66.8	[63.8,69.5]	15.3	[6.4,30.8]
	L. infantum	5671	7908	474		61.9	[59.8,63.8]	15.2	[6.1,31.2]
	P. sojae	1094619	20,235	323		59.2	[55.6,62.6]	10.8	[4.2,27.5]
	N. caninum	572307	6992	553		58.4	[54.9,61.3]	31.5	[13.7,50.2]
	H. sapiens	9606	29,063	434	Intermediate	52.5	[45.6,59.4]	13.8	[4.7,32.4]
	M. musculus	10090	23,088	415		52.3	[47.2,56.9]	12.0	[3.5,31.1]
	B. hominis	12968	5795	286		49.0	[44.6,54.3]	5.7	[2.0,16.9]
	G. intestinalis	5741	9234	420.5		48.1	[46.0,51.0]	6.8	[2.4,17.1]
	A. lyrata	81972	30,478	308		44.3	[42.3,46.6]	8.5	[3.1,23.5]
	T. adhaerens	10228	9627	360	Low	38.0	[36.3,39.6]	5.0	[1.9,15.1]
	N. gruberi	5762	14,768	402		34.8	[33.0,36.4]	6.7	[2.3,19.6]
	N. bombycis	578461	4049	195		30.0	[27.5,32.7]	4.2	[1.3,11.9]
	E. dispar	370354	7869	315		28.3	[25.8,30.7]	3.5	[1.2,10.4]
	I. multifiliis	857967	6787	324		24.7	[21.8,27.6]	3.4	[1.2,8.8]
Archaea	*H. mukohataei*	485914	3341	250	High	67.2	[64.5,69.3]	11.5	[5.3,22.1]
	M. kandleri	190192	1672	256		61.0	[58.9,63.3]	4.4	[2.0,8.8]
	T. archaeon	1054217	1527	247		59.9	[57.6,61.4]	3.9	[1.8,8.3]
	C. symbiosum	414004	2011	213		58.1	[54.8,60.6]	6.8	[3.2,16.3]
	Ca. M. alvus	1236689	1642	255.5		57.3	[54.6,59.5]	3.7	[1.9,7.8]
	M. thermautotrophicus	187420	1783	241	Intermediate	50.8	[47.9,52.8]	3.0	[1.2,6.5]
	N. gargensis	1237085	3522	163		49.8	[45.0,53.0]	6.1	[2.6,14.0]
	K. cryptofilum	374847	1600	262		49.8	[47.8,51.5]	2.3	[1.0,4.7]
	A. fulgidus	224325	2350	242		49.7	[47.3,51.5]	1.9	[0.7,4.5]
	M. hungatei	323259	3078	263		46.7	[43.3,49.6]	3.0	[1.4,6.4]
	A. boonei	439481	1538	253	Low	39.5	[37.1,41.8]	1.8	[0.7,4.4]
	Ca. N. limnia	886738	2035	194		32.9	[30.6,35.2]	3.5	[1.4,8.5]
	M. FS406-22	644281	1813	244		32.3	[29.8,34.8]	1.4	[0.0,3.6]
	N. equitans	228908	529	230		30.8	[29.1,33.0]	1.3	[0.0,3.3]

Table 1 (continued)

Domain	Species	NCBI taxonomy ID	Number of proteins	Median length	GC group	Per protein proportion coding GC (%)		Per protein proportion disordered residues (%)	
						Median	Quartiles [25,75]	Median	Quartiles [25,75]
Bacteria	*C. woesei*	469383	5911	296	High	73.0	[70.5,75.2]	6.6	[3.6,12.5]
	P. mikurensis	1142394	3266	305		73.0	[70.3,76.0]	9.6	[5.0,18.5]
	S. thermophilus	479434	3470	292.5		68.4	[66.5,70.2]	5.8	[3.0,11.5]
	A. paucivorans	584708	2390	299		68.3	[65.4,70.8]	4.7	[2.4,9.7]
	T. scotoductus	743525	2446	260		65.4	[63.5,66.9]	3.3	[1.6,7.0]
	T. primitia	545694	3511	295	Intermediate	52.3	[46.9,56.2]	3.2	[1.5,7.7]
	P. marinus	59922	2968	201		52.0	[47.4,55.1]	6.9	[3.0,17.7]
	E. coli	83334	6243	248		51.9	[48.1,54.3]	3.8	[1.7,9.2]
	S. linguale	504472	6866	293		51.1	[48.0,53.6]	3.1	[1.3,6.7]
	Ca. S. RAAC3	1394711	919	215		50.5	[47.9,52.2]	4.7	[2.0,11.9]
	S. aureus	450394	2724	243.5	Low	33.2	[30.8,35.3]	3.0	[1.2,7.6]
	T. africanus	484019	1916	279		30.8	[28.3,33.1]	1.3	[0.2,3.1]
	L. buccalis	523794	2217	259		30.6	[27.4,33.4]	2.0	[0.8,4.8]
	Ca. P. australiense	980422	976	184		28.0	[25.2,30.4]	4.1	[1.4,14.9]
	Ca. B. massiliensis	673862	978	282		27.9	[25.6,30.5]	1.2	[0.2,3.8]

methods must predict the disorder for a given residue to be predicted as disordered. This is motivated by an observation that consensuses secure better predictive quality when compared to the use of individual predictors [37, 38]. The same consensus was used in a number of other studies [7, 39–43], including the recent study that investigated relationship between GC and disorder content in Eukaryota [22]. Our approach is also similar to consensus-based putative annotations of disorder available from the MobiDB [44, 45] and D^2P^2 [46] databases. We verified the robustness of our results using an independent prediction method, VLXT [47]. The usual VLXT threshold of 0.5 was adjusted separately for each domain set, so that the median organism median fraction disorder residues was the same as for the consensus prediction method. This gave VLXT thresholds of 0.82, 0.85, and 0.80 for Bacteria, Archaea, and Eukaryota, respectively. The other work that looked into this relationship in Eukaryota has applied six disorder predictions [23], while the older study that analyzed Prokaryotes relied on a single prediction [25]. Neither of these previous approaches tested the robustness of results with independent disorder predictions. The putative disorder is annotated at the amino acid level allowing us to identify disordered regions and to quantify the amount of disorder per protein and per species.

The sequences for the 44 organisms together with the disorder predictions are available at http://biomine.cs.vcu.edu/datasets/IDPGC/.

Characterization of coding sequences

GC content was calculated as the fraction of guanine and cytosine in the coding sequences associated with each protein. Organisms were selected to have a large range of GC content in protein-coding genetic regions. The median proportion of GC overall genes ranged from 24.7 to 73%, with a similar distribution across each of the three domains (Table 1). In contrast to GC content, median fraction of predicted intrinsic disorder per protein is not similarly distributed among domains, with Eukaryotic organisms generally having a higher content of disorder than Archaea or Bacteria. This domain bias in intrinsic disorder agrees with many previous observations, e.g., [7, 9].

Similarly, GC content in disordered and ordered regions annotated in the protein sequences is defined as a fraction calculated over the coding regions associated with predicted disordered and ordered residues. For proteins with a minimal amount of both ordered and disordered residues, the relative GC content is calculated as the relative difference between disordered and ordered GC contents: (disorder_GC − order_GC)/order_GC. Expected relative GC content was calculated in the same way, except the average GC content of codons for each amino acid type was used to calculate the GC content of ordered and disordered regions, rather than the actual coding sequence. This gives the expected relative GC given compositional differences between ordered and disordered regions assuming that codon usage is the same

in both types of regions, accounting for differences in amino acid composition.

For classification of codons as frequent or infrequent, observed codon frequencies were compared to the uniform distribution, which is dependent on the number of codons for each amino acid type. In other words, ignoring bias in codon usage, frequent codons occur more than expected and infrequent codons occur less than expected. For example, for an amino acid with four codons, the uniform distribution has a fraction of 0.25 for each codon. Codons with an observed fraction greater than uniform are classified as frequent, and other codons are classified as infrequent. This approach is similar to the established codon bias index (CBI), which assigns numerical values proportional to the most frequent codon for each amino acid [48], but allows for the binary classification of codons.

Characterization of translational efficiency

We used the tRNA pairing index (TPI) [30] as an estimate of the translation efficiency of protein-coding genes. The TPI measures the degree to which isoaccepting codons are used to code for subsequent amino acids of the same type in mRNA sequences. TPI ranges from a value of 1, when subsequent isoaccepting codons are used much more than expected, to − 1, when subsequent isoaccepting codons are used much less than expected. TPI has been found to be correlated with translation efficiency [30]. TPI calculations were implemented as described [49], using version 2 of the calculation. Isoaccepting codon groups were defined as in previous work [30, 50].

For segregating efficiently translated coding sequences from other coding sequences, we selected a threshold TPI value of 0.9. This selection was based on the derivation of TPI, which is calculated from the probability that observed codon pairs occur at the observed frequency or less, relative to a null model where no pair bias is observed: $\mathrm{TPI} = 1 - 2p$. A probability of 0.05 corresponds to a TPI value of 0.9. The median fraction of coding sequences per organism with a TPI value greater than 0.9 is 0.4.

Results

Relationship between GC content and intrinsic disorder

GC and disorder content values were examined at several levels: organism, protein, and residue. The organism level provides a broad view of the typical protein in each organism. This does not account for systematic biases in compositions between organisms, but measured differences are robust because all proteins are included. At the protein level, ordered and disordered regions within the same protein are compared directly. This does correct for differences between organism compositions, since all qualities are relative within the same protein, but is less robust since only a portion of proteins contain both ordered and disordered regions. The residue level gives insight into the relationship of overall properties to protein composition and is examined in the next section.

At the organism level, the median protein values of GC and disorder content were used to characterize each organism. For comparison of GC and disorder content, organisms from each domain were grouped into three groups according to their median fraction GC (Table 1): low, intermediate, and high. Among the selected organisms, a higher median fraction of GC in protein-coding genes is associated with a higher median fraction of intrinsic disorder (Fig. 1). Across all domains, the high GC group contains significantly more

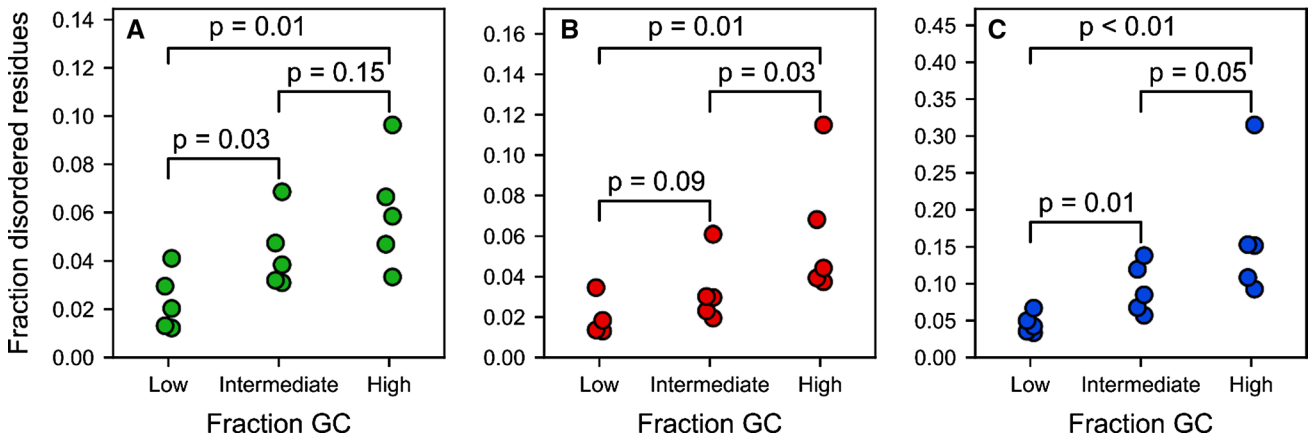

Fig. 1 Relationship between fraction of GC in protein-coding regions and fraction of disordered residues for selected **a** Bacteria, **b** Archaea, and **c** Eukaryota species. Medians of the fraction of residues in IDRs are plotted for each organism, which are grouped by relative proportion of GC (Table 1). The *p* values are obtained from the Mann–Whitney test

disordered proteins than the low GC group (p value < 0.05). For Eukaryota and Bacteria, the intermediate GC group contains a significantly greater amount of disorder than the low GC group (p value < 0.05). Moreover, the intermediate and high GC groups are not or marginally significantly different (p value = 0.05 for Eukaryota and 0.15 for Bacteria), although the abundance of disorder is visibly higher for the high GC group. Archaea shows similar relationships, with the high GC group significantly greater than the intermediate GC group (p value < 0.05), and low and intermediate GC groups that are not significantly different (p value = 0.09) but where the disorder content is visibly higher for the intermediate GC group. The relationship between intrinsic disorder and GC content was found to be insensitive to the prediction method used to estimate intrinsic disorder; comparison of intrinsic disorder predicted by the VLXT method (ref) for GC groups yielded nearly identical results (Fig S1), where all group differences are significant but one. This indicates that results are robust to the disorder prediction method used. Direct comparison of median GC and intrinsic disorder shows moderate positive correlations between median disorder content and GC content within each domain (Fig. S2): 0.74, 0.70, and 0.61 for Bacteria, Archaea, and Eukaryota, respectively.

At the protein level, the GC content of ordered and disordered regions within the same proteins was compared. The relative GC content between disordered and ordered regions was calculated by the fraction of GC in disordered regions relative to the fraction of GC in ordered regions within the same protein, and summarized over each organism. Proteins with a minimal number of both ordered residues and disordered residues were selected to obtain sufficient sampling of each structure type. Based on sampling of minimum residue thresholds between 1 and 100 amino acids (Fig. S3), a threshold of at least 20 ordered and disordered residues was selected because it is consistent with higher thresholds, while retaining a larger number of proteins. These results show that the residues of IDRs are consistently encoded by higher GC content sequences (Fig. 2); nearly all individual organisms have a positive relative median fraction GC. On average, Eukaryota show the largest bias in GC for encoding disordered regions, although the difference in domain medians is not significantly different (two-tailed student t test, not shown).

Among each domain, one organism does not show a significant bias in GC between disordered and ordered regions (Fig. 2): the bacterium *Leptotrichia buccalis*, the archaeon *Methanocaldococcus* sp. *FS406-22*, and the eukaryote *Ichthyophthirius multifiliis*. The compositions of IDRs in these organisms are extreme, relative to other selected organisms in their domains (Fig. S4A, B, and C). In particular, the IDRs of these organisms have the lowest

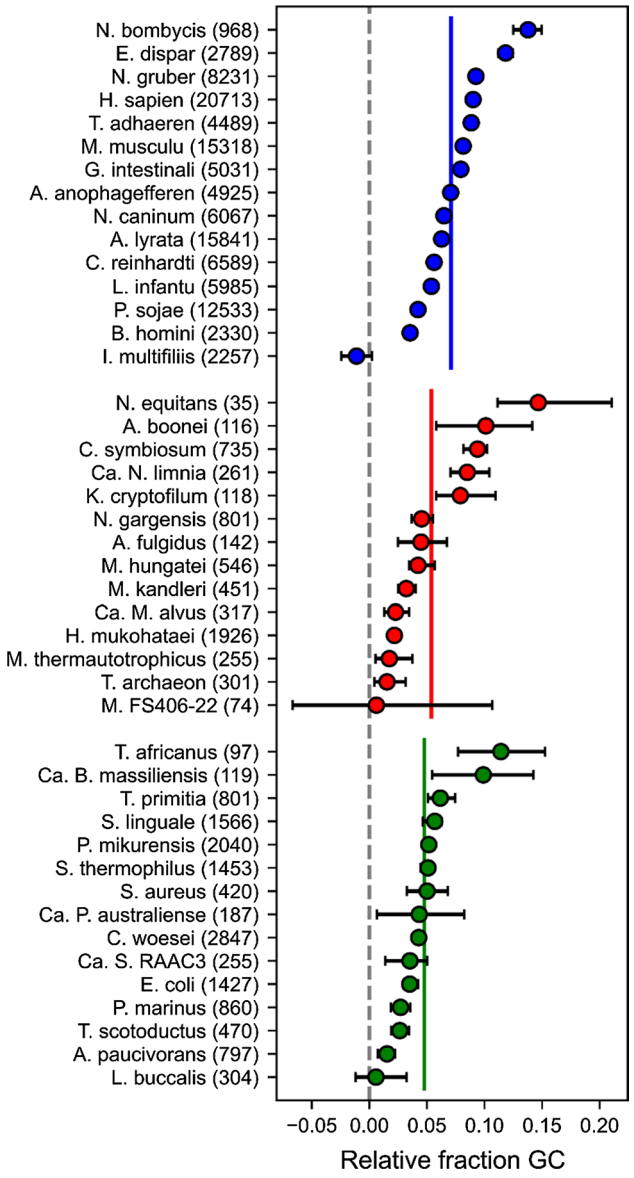

Fig. 2 Relative fraction of GC in disordered regions with respect to fraction GC in ordered regions. Points give the median value among proteins from each species with at least 20 disordered residues and 20 ordered residues (number of proteins). Error bars indicate the 95% confidence interval on the median. Horizontal lines indicate the mean of each of the three domains, which are (top to bottom): Eukaryota, Archaea, and Bacteria

proline contents among all species in their domains, where proline is one of the disorder-promoting amino acids most strongly associated with high GC content. Additionally, these organisms have the greatest, or near the greatest, content of lysine, asparagine, and isoleucine, all of which have among the least GC content. These data show that GC content bias in IDRs is not universal, but dependent on amino acid composition.

Role of amino acid composition in relative GC bias

To examine the residue-level relationship between GC and disordered content, the GC content of each amino acid type—relative to the overall GC content—was compared to its composition in disordered regions—relative to its composition in ordered regions (Fig. S5). Several amino acids consistently contribute to both disorder and GC content across all organisms studied. Proline and arginine—generally known to be disorder-promoting residues [15]—are positively associated with both disorder and GC content. Consistent with this, isoleucine, tyrosine, and phenylalanine—strongly order-promoting residues [15]—are negatively associated with both disorder and GC content. These five amino acids are the most consistent drivers in the GC and disorder content relationship. Of the remaining residues, a few violate the general trend, but most have codon GC content close to the overall species values, or—in the case of asparagine, glycine, and alanine—show an inconsistent content in ordered and disordered regions.

There are some notable exceptions to the residue-level relationship between disorder promotion and GC content. Lysine, which is generally strongly disorder promoting [15], has some of the most GC-poor codons. Tryptophan is a consistently order-promoting residue [15], with a codon that is more GC rich than most organism GC content, but is a rare amino acid and contributes little to overall GC. It should be noted that methionine shows a consistent bias toward IDRs in both Bacteria and Archaea (Fig. S5A and S5B). However, this is a result of the location of methionine at the amino terminus in proteins in these domains, where termini are frequently predicted to be disordered. Methionine is more common in eukaryotic proteins, and the values observed in this domain are inconsistent in disorder bias (Fig. S5C).

The general dependence of relative GC content on the amino acid bias of IDRs was verified by comparing the observed relative GC content to the expected value given the amino acid composition and codon distribution in each organism. This expected value differs from the observed value only in the GC value of each codon. For the expected value, the average GC content of codons for each amino acid is used to calculate GC content for ordered and disordered residues, whereas the observed value used the GC content of the coding sequence directly. Differences between observed and expected values are attributable only to differences in codon usage between ordered and disordered regions. In general, expected values (Fig. S6) show the same bias toward greater GC in disordered regions as observed values (Fig. 2). Direct comparison of observed and expected values of relative GC (Fig. 3) reveals two important features of the relationship between GC and disorder content. First, as expected, observed relative GC is generally proportional to the values expected from composition alone, which demonstrates the dependence of relative GC content on composition. Second, observed and expected values show a systematic bias, indicating the presence of codon selection bias between ordered and disordered region coding sequences.

Codon usage bias effects observed GC bias

Although observed relative GC is proportional to the value expected from compositions of predicted intrinsically disordered and ordered regions, most organisms show significantly less GC bias than expected (Fig. 4). The difference between observed and expected values is significant in all three domains (p value < 0.01). Similar to the observed results, the expected relative GC values of each domain are not significantly different from each other. The

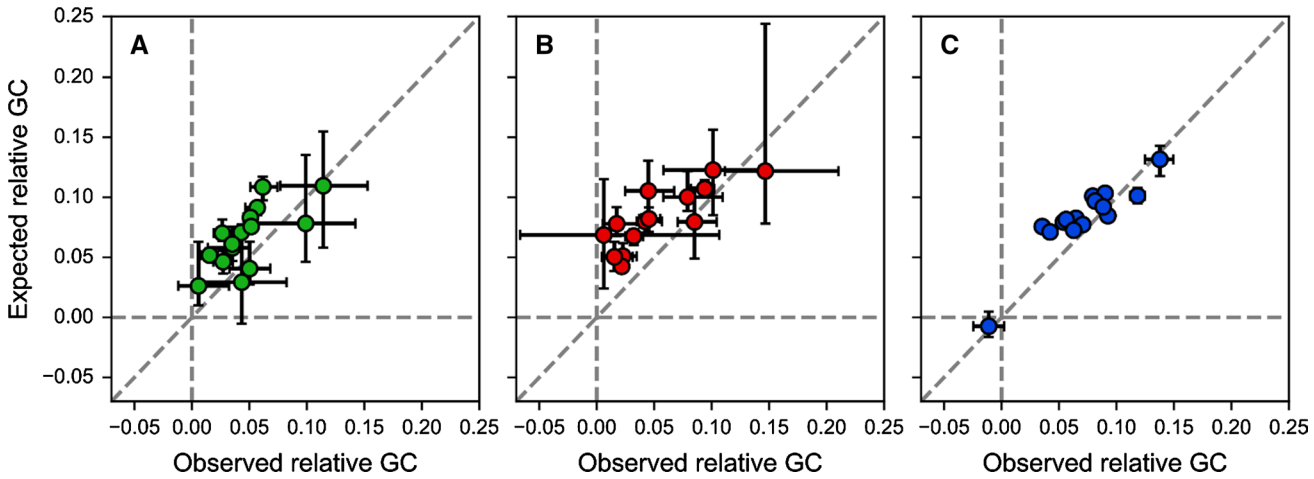

Fig. 3 Comparison of observed relative GC with expected relative GC in each species in **a** Bacteria, **b** Archaea, and **c** Eukaryota species. Medians of relative GC are plotted for each organism, where error bars indicate the 95% confidence interval on the median

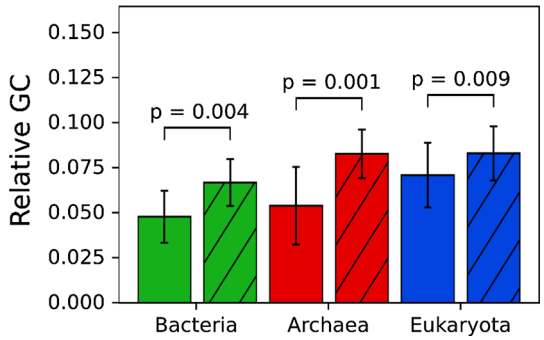

Fig. 4 Comparison of mean observed relative GC (solid bars) with mean expected relative GC (thatched bars). Means correspond to horizontal lines in Fig. 2 and Fig. S6, and error bars are the 95% confidence interval on the mean. The *p* values are results of a paired two-tailed *t* test

lower-than-expected observed relative GC indicates that while amino acid composition promotes a higher GC content in IDRs, codon selection counteracts this to some extent. In other words, selection among codons coding for the same amino acid is the only free parameter between the observed and expected values; so, this difference demonstrates that, in many organisms, GC-poor codons are preferentially selected for encoding the amino acids in the IDRs of many of the organisms studied here.

The relative difference between observed and expected relative GC compared to overall coding GC (Fig. 5) suggests that codon selection is dependent on the overall GC content of protein-coding genes. Organisms with low overall GC content show little to no difference between observed and expected relative GC content in IDRs, whereas organisms with moderate to high GC content show a decrease in observed relative GC due to codon selection. One apparent exception to this trend is the eukaryote *A. anophagefferen* (Fig. 5c, right most point), which has the greatest GC composition among selected Eukaryota, but little difference between observed and expected relative GC. The composition of the IDRs of *A. anophagefferen* was found to have the highest content of proline, arginine, and alanine among all selected Eukaryota IDRs (Fig. S4D), which are some of the amino acids most strongly associated with high GC content (Fig. 5c). While this contributes to the high GC content of this organism, relative codon usage between ordered and disordered regions is responsible for the lack of difference between observed and expected relative GC.

The relationship of GC content to codon usage in each organism was examined by calculating the fraction of infrequent codons that contain G or C in the third position. Here, infrequent codons are defined as those codons used less frequently than expected at random (see "Materials and methods"). It is seen that organisms with a high GC content have infrequent codons that are GC poor, and organisms with a low GC content have infrequence codons that are GC rich (Fig. 5), which is the expected relationship. In other words, the proportion of frequent codons that are GC rich is directly related to the overall GC content of the organism. This, combined with the relationship between overall GC content and the difference between observed and expected relative GC, suggests that GC-poor, infrequent codons may be used preferentially in IDRs of organisms with moderate to high GC contents.

Previous studies have found that IDR coding regions show significant bias in codon usage [20, 21]. One explanation for the difference between observed and expected relative GC is that GC-poor codon variants are relatively infrequent and used preferentially in coding for IDRs. The relative use of codons in ordered and disordered regions was compared

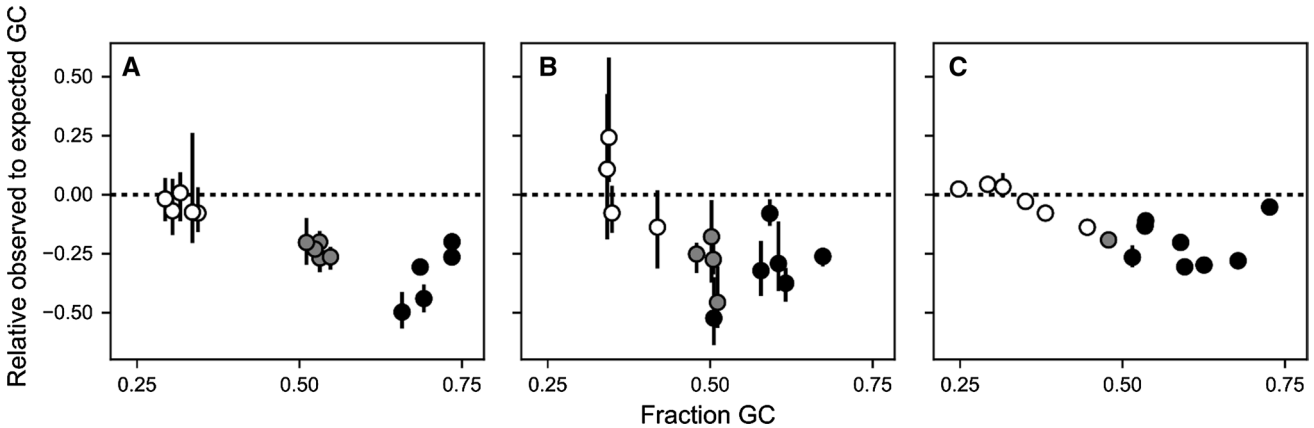

Fig. 5 Comparison of overall fraction GC with the median relative difference between observed and expected relative disorder GC for each organism in the three domains: **a** Bacteria, **b** Archaea, and **c** Eukaryota. Error bars indicate the 95% confidence interval on the median. Points are color coded by proportion of G or C in the third position of infrequent codons: 0–33% (black), 33–66% (gray), and 66–100% (white)

against the overall abundance of codons, noting the third position GC content (Fig. S7). These data show that, for many organisms, infrequent codon usage is biased toward coding for IDRs, and infrequent codons are often the third position A or T codon variants for a particular amino acid. These data were summarized by the median relative fraction of disorder for infrequent codons, noting the proportion of infrequent codons with third position G or C, and compared to the difference between observed and expected relative GC content (Fig. 6). This comparison shows that organisms with a reduced relative GC have IDRs enriched in GC-poor, infrequent codons. Further, this relationship is proportional, with moderate relative GC-depleted organisms having less bias in codon usage with fewer GC-poor infrequent codons, and organisms with close to the expected relative GC values having little codon bias and GC-rich infrequent codons. The extreme exception to the overall trend is *A. anophagefferen*, which shows the typical A and T third position bias for infrequent codons, but a bias for these codons to occur slightly more frequently in ordered regions, rather than IDRs (Fig. S7).

Translation efficiency and GC bias

Another view of codon usage bias has been proposed in terms of subsequent codons for the same amino acid [49]. The idea is that nature has optimized codon usage, so subsequent codons for the same amino acid are recognized by the same tRNA, which enhances translation efficiency. These isoaccepting codon pairs have been found to be greatly overrepresented in both eukaryotic [30] and prokaryotic [50] coding sequences, and the degree of isoaccepting codon pair bias is correlated with translation efficiency [30]. This suggests that efficiently translated IDRs, in the sense of optimized isoaccepting codon pairs, may show a stronger GC-bias effect than other IDRs.

We investigated this possibility by calculating the TPI value of each coding sequence in our dataset and dividing coding sequences for each organism into low TPI (lower translation efficiency) and high TPI (higher translation efficiency) sets and repeating the analysis on both sets separately. We find that the relationship between organism-level GC content and intrinsic disorder is generally insensitive to this division of coding sequences (Fig S8); the pattern of significant differences in intrinsic disorder content between GC content groups is nearly identical when comparing the whole set (Fig. 1), the low TPI set (Fig S8A, B, C) and the high TPI set (Fig S8D, E, F).

Further, we re-examined the relationship between observed and expected GC bias in intrinsically disordered regions as a function of organism GC content. The increase in GC bias in intrinsically disordered regions as organism GC increases observed for the whole set (Fig. 5) is also observed for both the low TPI set (Fig S9A, B, C) and the high TPI set (Fig S9D, E, F). The magnitude of the effect is also qualitatively maintained, except for intermediate GC content organisms, for which the low TPI set shows a somewhat larger effect for intermediate GC content organisms, with the high TPI organisms showing a corresponding slight decrease in effect. Regardless of these effects, consideration of expression efficiency does not explain the observed relationship between the magnitude of GC bias and organism GC content observed for the whole set.

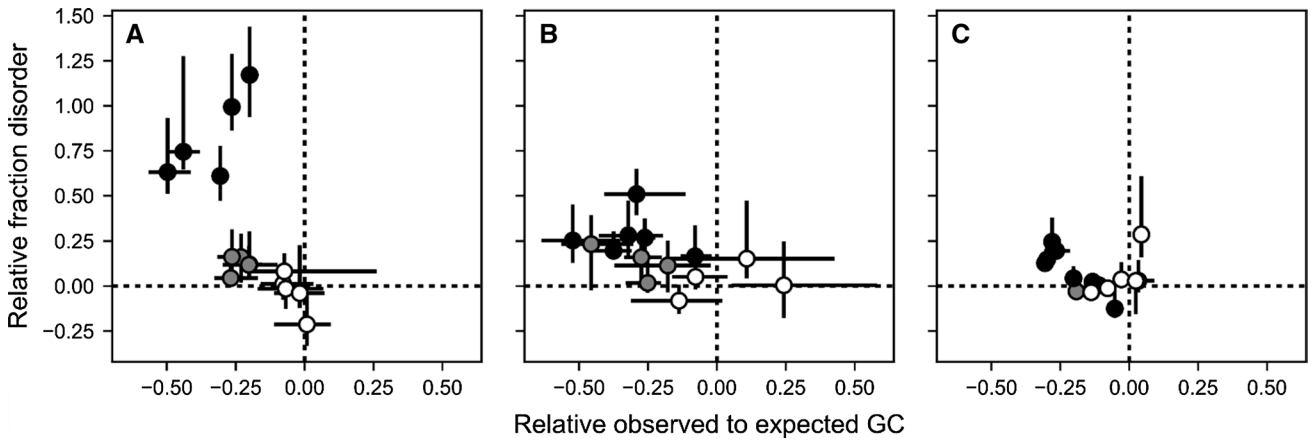

Fig. 6 Bias of infrequent codon usage in coding for disordered regions compared to the relative difference between observed and expected relative GC for **a** Bacteria, **b** Archaea, and **c** Eukaryota species. Disorder bias in infrequent codon usage is calculated as the relative frequency of codon usage in coding for IDRs with respect to its frequency in ordered regions, and is summarized by the median value for each organism. Error bars are the 95% confidence interval on the median. Points are color coded by proportion of G or C in the third position of infrequent codons: 0–33% (black), 33–66% (gray), and 66–100% (white)

Discussion

GC content reduced in IDR coding regions of high GC content organisms

The relationship between GC and disorder content was examined on a collection of protein and coding sequences from a broad set of 44 species spanning all domains of life, and a wide and balanced range of coding GC content. For both overall and for individual proteins containing ordered and disordered regions, we confirm the relationship between GC and disorder content previously observed in eukaryotes [22, 23]. For prokaryotes, the previously reported genome-scale observation [25] is found here to hold at the individual protein level as well. As suggested in this previous work, we find that amino acid biases between ordered and disordered regions play a major role in determining coding GC content. Moreover, we are the first to find that IDRs have a significantly lower GC content than expected, given the observed codon distribution in each organism, particularly for high GC content species. This systematic bias in GC content is observed in all domains of life. This demonstrates a bias in codon selection in IDRs; codons with a wobble position A or T are used preferentially over codons with a wobble position G or C in regions encoding IDRs. Further, this selection bias is manifested by the use of infrequent codons. That is, the G or C wobble codon variants are typically the predominant codon variants, and the use of infrequent variants is biased toward coding for IDR regions.

The current data suggest that codon selection for IDRs is subject to GC-content pressure. That is, organisms with a relatively low GC content show less use of infrequent codons for coding IDRs than organisms with a higher GC content (Fig. 6). These infrequent codons use the A or T codon variants and cumulatively reduce the overall coding GC observed in IDRs for high GC content organisms. Previous studies have examined the coding of IDRs in terms of coding efficiency, which is essentially the rate at which a region can be translated. This is quantified by the relative abundance of tRNAs for each codon. The relative abundance of tRNAs is unknown for most organisms and is commonly approximated using the relative adaption index [51]. The relative adaption index takes the number of tRNA genes for a given codon as an approximation of the codon's efficiency. Several other studies have observed that IDR coding regions are biased toward low-efficiency codons [20, 21]. The reduced translation efficiency of IDRs has been shown to be important for protein structure and biological function, where stalling of translation at IDRs may allow proper folding of structured domains [21]. Additionally, coding efficiency is conserved across homologous proteins, further supporting its biological importance [21]. Results here using codon frequencies and those based on relative adaption are generally comparable; when codon third position irregularities are accounted for, relative adaption is generally related to codon usage frequency [52]. However, the general GC content-based bias observed in this work does not conflict with efficiency-based codon selection for low GC content organisms, which do not show a bias for infrequent codons; the former is observed over all coding regions, but individual genes may have extreme biases. That is, the translation efficiency of individual gene regions may be regulated, even if the codon selection bias is not observed in aggregate.

Further, we do not see an effect on relationship between intrinsic disorder GC bias and organism GC content by coding sequence translation efficiency, as measured by TPI. This seems to contradict agreement between the observation of rare codon usage seen here and the association observed with reduced translation efficiency of IDRs due to suboptimal codon usage observed in previous studies [20, 21]. However, TPI is measured at the gene level, and the pattern of codon usage isolated to IDRs may not have sufficient impact on the calculated TPI. If TPI could be extended to region level calculations, a difference between ordered region TPI and disordered region TPI may be apparent.

Organisms without relative GC bias

There are two observations that do not follow the overall conclusions: lack of relative GC content bias for three lysine-rich organisms and bias of IDRs toward arginine. The overall and per-protein relationship between coding GC content and disorder content can be attributed in large part to amino acid biases in ordered and disordered regions in all domains of life. In reference to the standard translation table and previous examinations of amino acid biases in ordered and disordered regions [15], two disorder-promoting amino acids—proline and arginine—have G and C in the first two codon positions, and several order-promoting amino acids—isoleucine, phenylalanine, and tyrosine—have A and T (U) in the first two positions. These codons are sufficient to enrich regions encoding for IDRs in GC, for nearly all organisms studied here. The few organisms that do not follow this trend were shown to have an extreme lysine content in their IDRs (Fig. S4), where lysine is a disorder-promoting residue but with codons beginning with AA. High lysine content in the IDRs of these few organisms is sufficient to equalize the GC content, on the whole, between ordered and disordered regions. The unusual composition of IDRs in these organisms may be explained by unique requirements of these organisms. One possible explanation is protection from desiccation; lysine-rich IDRs are common in LEA proteins which function to protect other proteins

under desiccating conditions [53]. Both the eukaryotic and bacterial organisms with high lysine content IDRs have life cycles that include exposure to desiccating conditions. *Ichthyophthirius multifiliis* is a parasite that infects fish and has two stages outside the fish body. *Leptotrichia buccalis* is a gram negative, anaerobic, non-motile bacillus found in the oral cavity. A need for desiccation protection is less clear for the archaeon, *Methanocaldococcus* sp. FS406-22, which is an anaerobic, piezophilic, diazotrophic, hyperthermophilic marine archaeon. It is possible this archaeon might need proteins with lysine-rich regions for some specific functions. In fact, lysines are commonly methylated in Archaea, which is rare in Bacteria and eukaryotes [54].

Arginine as a disorder-promoting residue

Our results regarding arginine are somewhat unusual. This amino acid is not always considered as a disorder-promoting amino acid [15], and neither does it have the highest GC codons as judged by the standard translation table. In terms of disorder promotion, it is reasonable that arginine would be associated with disorder, since it is positively charged, and net charge is strong determinant of intrinsic disorder [3]. In fact, arginine-rich IDRs are known to play important roles in RNA binding [55] and in the formation of membraneless organelles [56]. On the other hand, arginine has been found to stabilize folded domains via multiple salt bridges and hydrogen bonds, e.g., [57]. The dual role of arginine likely explains its lack of a strong observed bias in known IDRs, being only marginally enriched in IDRs [15]. The arginine bias observed here is consistent with previous estimates; the association between arginine content and disorder is relatively weak, but consistent over all organisms examined. Further, in terms of GC content, arginine has six codons in the standard translation table, four with C and G in the first two positions and two with A and G. Given that arginine is consistently correlated with high coding GC in these data, it seems that the four C and G codons are generally more common.

Acknowledgements This research was supported in part by the National Science Foundation Grant 1617369 and the Robert J. Mattauch Endowment from Virginia Commonwealth University to L.K.

References

1. Dunker AK, Obradovic Z (2001) The protein trinity-linking function and disorder. Nat Biotechnol 19:805–806
2. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321–331
3. Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41:415–427
4. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208
5. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582
6. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6:1882–1898
7. Peng Z et al (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 72:137–151
8. Peng Z, Mizianty MJ, Kurgan L (2014) Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins 82:145–158
9. Xue B, Dunker AK, Uversky VN (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn 30:137–149
10. Pancsa R, Tompa P (2012) Structural disorder in eukaryotes. PLoS One 7:e34687
11. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645
12. Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37:509–516
13. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839
14. Walsh I, Martin AJ, Di Domenico T, Tosatto SC (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28:503–509
15. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins 42:38–48
16. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinform 7:208
17. Meng F, Uversky VN, Kurgan L (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 74:3069–3090
18. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S (2016) How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. Intrinsically Disord Proteins 4:e1259708
19. Romero PR et al (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci USA 103:8390–8395
20. Homma K, Noguchi T, Fukuchi S (2016) Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. Nucleic Acids Res 44:10051–10061
21. Zhou M, Wang T, Fu J, Xiao G, Liu Y (2015) Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol Microbiol 97:974–987
22. Peng Z, Uversky VN, Kurgan L (2016) Genes encoding intrinsic disorder in Eukaryota have high GC content. Intrinsically Disord Proteins 4:e1262225
23. Basile W, Sachenkova O, Light S, Elofsson A (2017) High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput Biol 13:e1005375
24. Yruela I, Contreras-Moreira B (2013) Genetic recombination is associated with intrinsic disorder in plant proteomes. BMC Genom 14:772

25. Pavlovic-Lazetic GM, Mitic NS, Kovacevic JJ, Obradovic Z, Malkov SN, Beljanski MV (2011) Bioinformatics analysis of disordered proteins in prokaryotes. BMC Bioinform 12:66
26. Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204
27. Yin H, Wang G, Ma L, Yi SV, Zhang Z (2016) What signatures dominantly associate with gene age? Genome Biol Evol 8:3083–3089
28. Amit M et al (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep 1:543–556
29. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins Struct Funct Bioinform 42:38–48
30. Cannarozzi G et al (2010) A role for codon order in translation dynamics. Cell 141:355–367
31. Pruitt KD et al (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19:1316–1323
32. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40:D71–D75
33. Kanz C et al (2005) The EMBL nucleotide sequence database. Nucleic Acids Res 33:D29–D33
34. Peng ZL, Kurgan L (2012) Comprehensive comparative assessment of in silico predictors of disordered regions. Curr Protein Pept Sci 13:6–18
35. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SC (2015) Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 31:201–208
36. Piovesan D et al (2016) DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res D1:D219–D227
37. Peng, Z. and Kurgan, L. (2012). On the complementarity of the consensus-based disorder prediction. In: Pacific symposium on biocomputing, pp 176–187
38. Fan X, Kurgan L (2014) Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn 32:448–464
39. Na I, Meng F, Kurgan L, Uversky VN (2016) Autophagy-related intrinsically disordered proteins in intra-nuclear compartments. Mol BioSyst 12:2798–2817
40. Meng F, Na I, Kurgan L, Uversky VN (2016) Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein–protein interactions in intra-nuclear compartments. Int J Mol Sci 17:24
41. Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, Kurgan L, Uversky VN (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci 71:1477–1504
42. Hu G, Wu Z, Wang K, Uversky VN, Kurgan L (2016) Untapped potential of disordered proteins in current druggable human proteome. Curr Drug Targets 17:1198–1205
43. Wang C, Uversky VN, Kurgan L (2016) Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16:1486–1498
44. Di Domenico T, Walsh I, Martin AJM, Tosatto SCE (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28:2080–2081
45. Potenza E, Di Domenico T, Walsh I, Tosatto SC (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43:D315–D320
46. Oates ME et al (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Res 41:D508–D516
47. Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-, and internal regions. Genome Inform Ser Workshop Genome Inform 10:30–40
48. Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031
49. Friberg MT, Gonnet P, Barral Y, Schraudolph NN, Gonnet GH (2006) Measures of codon bias in yeast, the tRNA pairing index and possible DNA repair mechanisms. In: Bucher P, Moret B (eds) Algorithms in bioinformatics. WABI 2006. Lecture Notes in Computer Science, vol 4175. Springer, Berlin, Heidelberg
50. Guo F-B, Ye Y-N, Zhao H-L, Lin D, Wei W (2012) Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. DNA Res Int J Rapid Publ Rep Genes Genomes 19:477–485
51. Reis MD, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036–5044
52. Novoa EM, Ribas de Pouplana L (2012) Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet 28:574–581
53. Petersen J, Eriksson SK, Harryson P, Pierog S, Colby T, Bartels D, Rohrig H (2012) The lysine-rich motif of intrinsically disordered stress protein CDeT11-24 from *Craterostigma plantagineum* is responsible for phosphatidic acid binding and protection of enzymes from damaging effects caused by desiccation. J Exp Bot 63:4919–4929
54. Botting CH, Talbot P, Paytubi S, White MF (2010) Extensive lysine methylation in hyperthermophilic crenarchaea: potential implications for protein stability and recombinant enzymes. Archaea 2010:106341
55. Varadi M, Zsolyomi F, Guharoy M, Tompa P (2015) Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PLoS One 10:e0139731
56. Uversky VN (2017) Protein intrinsic disorder-based liquid–liquid phase transitions in biological systems: complex coacervates and membrane-less organelles. Adv Colloid Interface Sci 239:97–114
57. Siddiqui KS, Cavicchioli R (2006) Cold-adapted enzymes. Annu Rev Biochem 75:403–433