Uncovering non-random sequence patterns within intrinsically disordered proteins

Megan C. Cohan[1], Min Kyung Shinn[1], Jared M. Lalmansingh[2], and Rohit V. Pappu[1*]

[1]Department of Biomedical Engineering and Center for Science & Engineering of Living Systems (CSELS), Washington University in St. Louis, MO 63130, USA

[2]Department of Physics, Washington University in St. Louis, MO 63130, USA

*pappu@wustl.edu

Abstract

Intrinsically disordered proteins / regions (IDPs / IDRs) pose unique challenges for deriving sequence-function relationships from multiple sequence alignments. These challenges arise from variations in sequence lengths, similarities, and identities across orthologs. Recent computational efforts have demonstrated the utility of comparing large numbers of distinct sequence features as a strategy to identify conserved sequence-function relationships in IDPs / IDRs. Inspired by these efforts, and by biophysical studies that have established the importance of binary patterning features in IDPs / IDRs, we present here a computational method, NARDINI (**N**on-random **A**rrangement of **R**esidues in **D**isordered **R**egions **I**nferred using **N**umerical **I**ntermixing), to uncover truly non-random binary patterns within disordered proteins / regions. Binary patterns refer to the linear clustering or dispersion of specific residues or residue types with respect to all other residues or specific types of residues. Our approach does not use, nor does it require sequence alignments. Instead for each IDR, we generate an ensemble of scrambled sequences and use this to set up expectations from a composition-specific null model for the patterning parameters of interest. We annotate each IDR in terms of pattern-specific z-score matrices by computing how specific patterns deviate from the null model. The z-scores help in identifying the non-random linear sequence patterns within an IDR. We tested the accuracy of NARDINI derived z-scores by assessing the ability to identify sequence patterns that have been identified as determinants of sequence-function relationships in specific IDPs / IDRs.

Introduction

Approximately 30-40% of eukaryotic proteomes contain proteins that are either entirely disordered or include disordered regions [1,2]. Under typical approximations of physiological solution conditions (pH 7.4, 37˚C, and 100-300 mOsm of salts and solutes), IDPs / IDRs adopt heterogeneous ensembles of conformations, although some IDPs / IDRs can be coaxed into differently stable folds in 1:1 or higher-order complexes with their binding partners [3]. Therefore, IDPs / IDRs are distinct from intrinsically foldable proteins / domains, which adopt essentially singular and stable folds or metamorphic proteins / domains that adopt at least two stable folds [4].

Intrinsically foldable proteins can be classified into distinct sequence families based on sequence-structure-function relationships [5,6]. It is often true that folded proteins with similar sequences have similar structural preferences [7], although metamorphic proteins clearly challenge this thinking [4]. However, one can still identify sequence families and construct sequence phylogenies using multiple sequence alignments of intrinsically foldable proteins [5]. In contrast, many IDRs are characterized by significant variations within sequence families [8-10]. While often used to describe genetic regions with high-degrees of variability [11], here we co-opt the term *hypervariability* to describe certain types of IDPs / IDRs. We define hypervariable IDRs as being involved in similar functions across orthologs even though they are characterized by significant differences in sequence lengths, compositional biases, sequence identities, and sequence similarities. As a result, multiple sequence alignments must incorporate numerous gaps to maximize sequence similarity and identity within an alignment [12,13]. These gaps lower the statistical significance of an alignment because the likelihood of realizing an alignment produced by hypervariable sequences can become equivalent to the likelihoods produced by alignments of random sequences [14].

Although sequence identity and similarity tend to be poorly conserved across many IDRs, one can compare amino acid compositions across orthologs [15-18]. This often reveals the conservation of compositional biases that might point to conserved composition-function relationships [17,18]. The relevant parameters include the overall amino acid compositions [15], sequence lengths [19], the net charge per residue [20-22], and the presence of conserved short linear motifs (SLiMs) [23-28]. One example of a family of IDRs with a well-studied composition-to-function relationship is the disordered RGG domain found in proteins that drive phase separation [29-31]. While the actual sequences cannot be aligned without inserting and extending numerous gaps, the overall profile of compositions are fairly similar across this family of sequences [30]. Can sequence-function relationships of IDPs / IDRs be discerned by compositional biases alone? Efforts to date suggest that the answer to this question depends on the functions associated with the IDR in question and on the extent to which compositions vary across IDRs from orthologous systems. For example, the compositions vary considerably for IDRs drawn from bacterial proteomes [10,32,33].

In addition to amino acid compositions, sequence patterns are also important determinants of sequence-ensemble-function relationships of IDPs / IDRs [8,9,14,16,17,34-37]. These include binary patterning parameters that quantify either the segregation or mixing of a specific residue / residue type with respect to another type of residue [38] or all other residues in the sequence [39]. Previous studies have demonstrated the functional importance of binary patterning parameters through mutational studies where sequence compositions were fixed, and the impacts of variations in linear sequence patterns on functions were assessed [12,14,28,35,39,40]. The findings, which highlight the importance of binary patterns, have prompted comparisons of context-dependent features within sets of hypervariable sequences as a route to inferring evolutionarily conserved functions. For example, Buske et al., observed that despite having different compositions, the patterning of

oppositely charged residues, quantified in terms of the parameter (κ), was bounded between 0.15 and 0.4 for the intrinsically disordered C-terminal linkers (CTLs) of FtsZs [41]. This observation led to the conjecture that the bounds on κ might be indicative of functional significance. Whether or not this conjecture is valid depends on whether parameters such as κ are comparable across different amino acid compositions and lengths. Here, we find that the binary patterning parameters introduced to date cannot be deployed for quantitative comparisons across compositions. If the amino acid compositions are different from one another, then a specific κ value does not imply similar degrees of segregation or mixing of oppositely charged residues. To remedy this and enable the extraction of truly non-random binary patterns within IDR sequences, we developed and deployed NARDINI (**N**on-random **A**rrangement of **R**esidues in **D**isordered **R**egions **I**nferred using **N**umerical **I**ntermixing), an algorithm that is available as a plug in of localCIDER [37], a freely available codebase for analyzing physico-chemical properties of IDRs.

In what follows, we first present the problem one encounters when comparing binary patterning parameters across IDRs that have different amino acid compositions and lengths. This analysis is used to motivate the development of NARDINI, which relies on two advances: (i) the development of a suitable null model that allows for proper assessments of the non-randomness of a patterning parameter of interest and (ii) the use of the method of z-scores to annotate each IDR in terms of pattern-specific z-scores. This allows us to uncover non-random binary patterns within IDRs. We show, using a set of examples, that the patterns we extract are concordant with known sequence-function relationships in specific systems.

Results

Parameters that quantify extents of segregation vs. mixing of different types of residues vary with amino acid composition and sequence length

A parameter that has been used to quantify sequence-ensemble-function relationships of IDRs is κ (which we will denote here as κ_{+-}) [38]. This parameter, which was introduced to quantify the overall segregation versus mixing of oppositely charged residues with respect to one another [14,17,38], is defined as follows. For a given IDR, we tally f_+ and f_-, the overall fractions of positive and negatively charged residues, respectively. This is then used to compute the overall charge asymmetry defined as: $\sigma_{+-} = \frac{(f_+ - f_-)^2}{(f_+ + f_-)}$. A sliding window comprising g residues is then defined and within each window j, we compute the local charge symmetry σ_{+-}^{j}. The mean squared deviation between the local and global charge asymmetry is computed as: $\delta_{+-} = \frac{\sum_{j=1}^{n_w}\left(\sigma_{+-}^{j} - \sigma_{+-}\right)^2}{n_w}$, where n_w denotes the number of sliding windows used in the calculation. The mean squared deviation is normalized using a statistical model to estimate the maximal possible deviation δ_{+-}^{max} that is realizable for the composition such that $\kappa_{+-} = \left(\frac{\delta_{+-}}{\delta_{+-}^{max}}\right)$, thereby ensuring that $0 \leq \kappa_{+-} \leq 1$.

Can values of κ_{+-} be compared across sequences of different compositions and lengths? We answered this question by designing a computational experiment to assess if values of κ_{+-} vary or stay fixed with FCR, the fraction of charged residues defined as $(f_+ + f_-)$ [37], and with sequence length for a fixed composition. We considered a simple three-letter alphabet comprising residues A, E, and K. We fixed the net charge per residue (NCPR) to be zero, i.e., $(f_+ - f_-) = 0$. For a given sequence length, set here to be 50 or 100, and fixed composition achieved by fixing the values of f_+, f_-, and f_A such that $(f_+ + f_- + f_A) = 1$, we generated 10^5 randomly shuffled sequence variants. For each sequence variant, we calculated the value of κ_{+-}. For FCR values greater than 0.1, the

distribution of κ_{+-} values could be fit using a gamma distribution (**Figure S1**). The mean (m) of the gamma distribution is the most likely value of κ_{+-} for the specific composition. We asked if the value of m depends on FCR or sequence length. The impact of FCR was assessed by repeating the calculations described above for different values of FCR, operating under the joint constraint of NCPR = 0 and $(f_+ + f_- + f_A) = 1$.

Figure 1 shows that $m(\kappa_{+-})$ varies with FCR for sequences that are 50 and 100 residues long. Contrary to the assumption that is implicit in comparisons of values of κ_{+-} across different compositions [32], we find that the most likely value for κ_{+-} varies with amino acid composition while the variation with sequence length is limited to low FCR values. Although, κ_{+-} is suitable for comparing the patterning of oppositely charged residues across sequence variants that have identical compositions and lengths [14,17,34,35,38], the normalization introduced by Das and Pappu [38] is inadequate for enabling comparisons across sequences with different compositions.

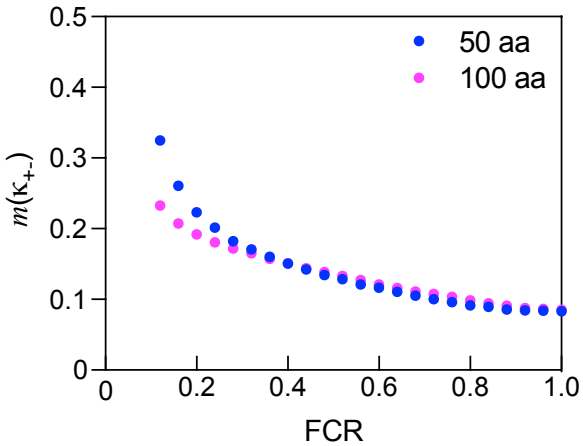

Figure 1. Plot of how the most likely value of κ_{+-}, measured as the mean of the gamma distributions ($m(\kappa_{+-})$), depends on the fraction of charged residues (FCR) for sequences of 50 and 100 residues. The null-scramble expectations, i.e., the mean values of gamma distributed κ_{+-} values are shown for sequences that 50 and 100 residues long as a function of FCR. The mean value of κ_{+-} is dependent on FCR for low values of FCRs (< 0.3) and this dependency is also manifest for different sequence lengths.

Alternative parameters used to quantify the linear mixing vs. segregation of oppositely charged residues include the sequence charge decoration (SCD) parameter of Sawle and Ghosh [42]. The dependence on composition that we observe for κ_{+-} values is also observed for SCD (**Figure 2**), although the dependence is more pronounced. Further, there is a clearer length dependence to the most likely value of SCD. Therefore, just like κ_{+-}, the SCD family of parameters cannot be deployed, as currently prescribed, to compare patterns across sequences of different compositions and lengths. In addition to κ_{+-} and SCD, Martin et al.,[28,39] introduced the Ω_x-family of patterning parameters to quantify the linear mixing versus segregation of a residue x or residues of type x with respect to all other residues. Here, we find that the most likely values of Ω_x show a non-monotonic dependence on amino acid composition (**Figure 3**). Accordingly, we conclude that despite their importance as descriptors of sequence-ensemble-function relationships [14,17,28,34,35,37-39,43], the parameters introduced to quantify mixing vs. segregation of different types of amino acids within IDRs cannot be deployed to compare patterns across sequences of different amino acid compositions and lengths. This poses a challenge for uncovering functionally important patterns across hypervariable IDRs within orthologous systems – a problem we remedy by deploying a new method, NARDINI, as described next.

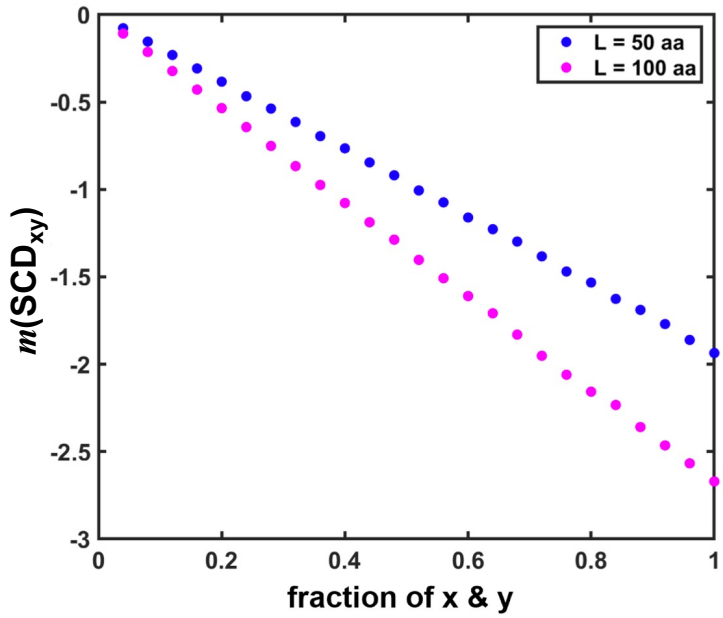

Figure 2. The mean value of SCD, quantified from the gamma distribution of SCD values for 10^5 scrambled sequences, depends on the sequence charge content and length. The null-scramble expectations of SCD are plotted for sequences of 50 and 100 residues as a function of fraction of residue x that is positively charged and y that is negatively charged (SCD_{xy}). The expectation of SCD is dependent on the sequence length.

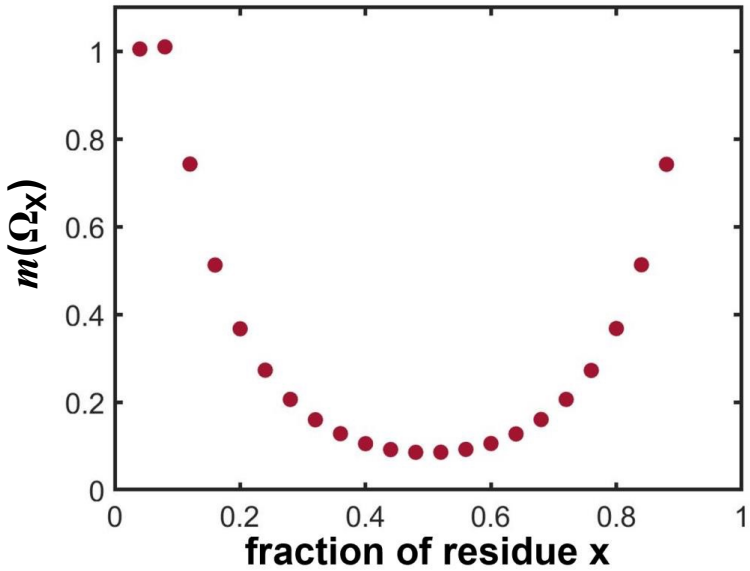

Figure 3. The mean value for the Ω parameter, as extracted from the gamma distribution obtained using 10^5 randomly shuffled sequences, depends on amino acid composition. The plot shows the null-scramble expectations of the most likely value for (Ω_x) for sequences of 100 residues as a function of the fraction of residue x.

Generation of a null model

The first step of the NARDINI method, which is designed as a plug-in for localCIDER [37], is the generation of a suitable null model. The workflow is summarized in **Figure 4**. For an IDR of interest, we generate 10^5 independent scrambled sequences. In this approach, the amino acid composition is fixed, and the linear sequence is randomly shuffled. The patterning parameter of interest is calculated for each of the 10^5 randomly generated sequences. Irrespective of the parameter of interest, the random manner of generating sequence scrambles leads to maximum entropy-based gamma distributions. The mean (m) of the gamma distribution is the null expectation for the patterning parameter of interest. Next, for each sequence i in the set of scrambles, and patterning parameter of interest designated as q, we compute the sequence-specific z-score as: $z\left(q_i\right)=\left(\frac{m-q_i}{s}\right)$. Here, $z(q_i)$ is the z-score of sequence i for the patterning parameter q, whereas m and s are the mean and standard deviation of the distribution $P(q)$. This is a gamma distribution for compositions and lengths that are identical to those of sequence i. For every parameter q, we can use the z-scores to identify patterns that have random versus non-random values. The z-scores enable two types of comparisons: We can quantify the extent to which the value of a specific patterning parameter deviates from random expectation, where $z \approx 0$ corresponds to the truly random scenario. Additionally, we can compare z-scores for the patterning parameter of interest across a set of IDRs from orthologous systems and gain a sense of extent to which the pattern in question is conserved, achieved by comparing z-score values, and whether the pattern in question is random vs. non-random in different orthologs i.e., is $z \approx 0$? Comparisons across different sequences are feasible because the z-score has a rigorous statistical interpretation,

and for each sequence it is calibrated by composition and length-specific distribution of values for the patterning parameter of interest.

Figure 4. Workflow for the calculation of z-score matrices for various binary patterns. Process includes generating the null-scramble model ("null model") and calculating the deviation of the observed value from the null model as z-scores.

Prescribing a set of patterning parameters

There are two broad categories of parameters that quantify binary patterning within IDRs. We refer to these as the δ- and Ω-family of parameters. These are defined as follows: Consider a sequence defined by a specific amino acid composition such that f_x and f_y quantify the fractions of amino acids x and y, respectively. Further, $(f_x + f_y) \leq 1$. We define $\delta_{xy} = \frac{\sum_{j=1}^{n_w}\left(\sigma_{xy}^{j} - \sigma_{xy}\right)^2}{n_w}$, $\sigma_{xy} = \frac{\left(f_x - f_y\right)^2}{\left(f_x + f_y\right)}$, and $\sigma_{xy}^{j} = \frac{\left(f_x^j - f_y^j\right)^2}{\left(f_x^j + f_y^j\right)}$ is the local asymmetry within window *j*. Here, we have replaced + and – from the previous section with x and y. Likewise, for $f_x + f_y = 1$, i.e., for a binary classification where residues either do or do not belong to group x, we can define the Ω-family of parameters using: $\Omega_x = \frac{\sum_{j=1}^{n_w}\left(\psi_x^j - \psi_x\right)^2}{n_w}$, where $\psi_x = \frac{\left(f_x - f_{y\neq x}\right)^2}{\left(f_x + f_{y\neq x}\right)}$ and $\psi_x^j = \frac{\left(f_x^j - f_{y\neq x}^j\right)^2}{\left(f_x^j + f_{y\neq x}^j\right)}$. In both cases, n_w is the number of sliding windows across which the deviation between local and global

asymmetries are calculated. For a sliding window of g-residues, $n_w = N - g+1$, where N is the sequence length. Following previous calibrations [28,38,39], we set $g = 5$.

The δ- and Ω-family of parameters are computed for specific groups of residues / residue types. We define eight such groups and denote these as polar residues μ ≡ {S,T,N,Q,C,H}, hydrophobic residues h ≡ {I,L,M,V}, basic residues + ≡ {R,K}, acidic residues – ≡ {E,D}, aromatic residues π ≡ {F,W,Y}, alanine A≡{A}, proline P ≡ {P}, and glycine G ≡ {G}. These groupings are based on similarities of stereochemistry, charge, aromaticity, size, and hydrophobicity. The groupings we choose lead to the calculation of 36 distinct patterning parameters. These include eight parameters from the Ω-family *viz*., Ω_μ, Ω_h, Ω_+, Ω_-, Ω_π, Ω_A, Ω_P, and Ω_G and 28 parameters from the δ-family, where δ_{xy} represents every unique pairwise combination of {μ,h,+,–,π,A,P, G}. For every IDR of interest, the NARDINI algorithm generates a z-score matrix, which is depicted in **Figure 5**. We use an 8×8 matrix where only the diagonal and upper triangular values are considered. Note that the matrix size changes as we change the number of sets of residues / residue types of interest. The diagonal elements in the matrix are Ω-values that quantify the linear mixing versus segregation of a residue or a residue type with respect to all other residues. The off-diagonal elements in the matrix are δ-values that quantify the linear mixing versus segregation of pairs of residues / residue types. When sequences contain fewer than 10% of a residue or residue type, all z-scores that involve that residue type are set to zero. Note that the z-score matrix is specific to a given sequence and helps identify the non-random binary patterns that are present in and IDP / IDR. We used sequence-specific z-score matrices to assess whether one or more Ω and δ parameter is identifiable as being non-random. For fully random sequences, the entire z-score matrix will have z-score values that are essentially zero.

$z(\Omega_{\mu})$	$z(\delta_{\mu h})$	$z(\delta_{\mu +})$	$z(\delta_{\mu -})$	$z(\delta_{\mu \pi})$	$z(\delta_{\mu A})$	$z(\delta_{\mu P})$	$z(\delta_{\mu G})$
	$z(\Omega_{h})$	$z(\delta_{h+})$	$z(\delta_{h-})$	$z(\delta_{h\pi})$	$z(\delta_{hA})$	$z(\delta_{hP})$	$z(\delta_{hG})$
		$z(\Omega_{+})$	$z(\delta_{+-})$	$z(\delta_{+\pi})$	$z(\delta_{+A})$	$z(\delta_{+P})$	$z(\delta_{+G})$
			$z(\Omega_{-})$	$z(\delta_{-\pi})$	$z(\delta_{-A})$	$z(\delta_{-P})$	$z(\delta_{-G})$
				$z(\Omega_{\pi})$	$z(\delta_{\pi A})$	$z(\delta_{\pi P})$	$z(\delta_{\pi G})$
					$z(\Omega_{A})$	$z(\delta_{AP})$	$z(\delta_{AG})$
						$z(\Omega_{P})$	$z(\delta_{PG})$
							$z(\Omega_{G})$

Figure 5. Elements of a typical, sequence-specific z-score matrix. Each cell quantifies the z-score of a specific patterning parameter. The diagonal elements are z-scores for Ω parameters whereas the off-diagonal elements are z-scores for the δ-parameters. Residues are grouped into eight categories: (μ) (S,T,N,Q,C,H), hydrophobic; (h) (I,L,M,V); positive (+) (R,K); negative (-) (E,D); aromatic (p) (F,W,Y); alanine (A); proline (P); and glycine (G).

Analysis of the prion-like low complexity domain from hnRNPA1

Prion-like low complexity domains (PLCDs) feature prominently among IDRs that are known to be drivers of phase separation and the formation of distinct types of biomolecular condensates [27,44]. The PLCD from the RNA binding protein hnRNPA1 (referred to as the A1-LCD) has been used as an archetypal system to uncover the driving forces for phase separation [45]. A recent study showed that aromatic residues (Tyr / Phe) are the primary cohesive motifs (stickers) in the A1-LCD system [28]. This appears to be a shared feature of other PLCDs [28]. Interestingly, the linear segregation vs. mixing of aromatic residues contributes directly to the driving forces for phase separation and determines whether the condensates are liquid-like or amorphous solids [28].

We used the NARDINI approach to analyze the A1-LCD sequence (**Figure S2**). The results are shown in **Figure 6** as a z-score matrix. There are four distinctly non-random features for the A1-LCD. First, there is a clear preference for linear clustering of polar residues $\mu \equiv \{S,T,N,Q,C,H\}$, with a z-score for Ω_μ of +1.96. This derives from the presence of blocks of Ser residues along the sequence. Second, a similar strong preference for linear clustering is observed for Gly residues with the z-score of +3.03 for Ω_G. Third, the Ser and Gly blocks tend to be segregated from one another, as evidenced by the z-score value of +3.93 for $\delta_{\mu G}$. And fourth, the z-score value for Ω_π is –1.86, and this recapitulates the documented preference for uniform dispersion of aromatic residues along the linear sequence [28]. The four distinctly non-random patterns are concordant with features have been shown to be directly relevant to the phase behavior of A1-LCD [28].

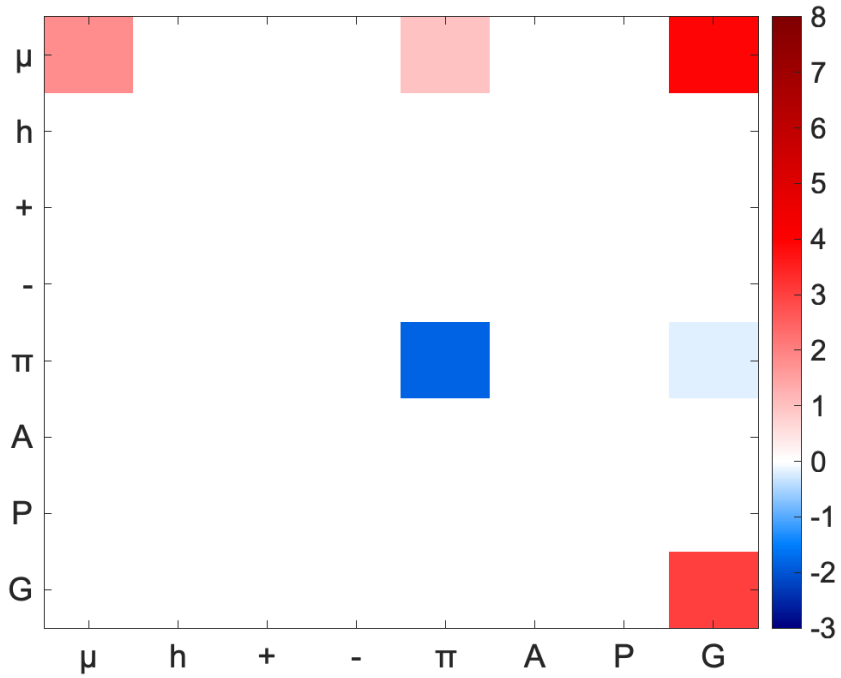

Figure 6. The z-score matrix of Al-LCD. A1-LCD shows non-random segregation of polar and glycine residues as well as from other residues, and non-random linear dispersion of aromatic residues. White squares on the checkerboard plot imply that the associated z-scores are ≈ 0.

Analysis of C-terminal Domains (CTDs) in RNases E

In bacteria, the protein RNase E is a critical driver of the formation of the RNA degradasome [46-48]. The architecture of this protein includes a conserved DEAD-box RNA helicase and a disordered C-terminal domain (CTD). In *C. crescentus,* the RNase E CTD is necessary and sufficient to drive phase separation. *In vivo,* RNase E drives the formation of cytoplasmic foci that colocalize with other exonucleases. This degradation body has been termed Bacterial Ribonucleoprotein body or BR-body [48]. The CTD of the *C. crescentus* RNase E has a blocky patterning of oppositely charged residues [47]. The blocky nature leads to encoded multivalence of opposite charges [49], and this architecture is essential for the formation of BR bodies [48]. In contrast, the *E. coli* RNase E, which lacks the blocky patterning of oppositely charged residues, does not form cytoplasmic condensates; instead, it forms membrane-tethered puncta and does not exhibit liquid-like behavior *in vitro* under the solution conditions that have been investigated to date [46,48]. Sequences of each CTD are shown in Supplementary Information (**Figure S2**).

Are there distinct sequence patterns that distinguish the CTDs of RNases E from *C. crescentus* and *E. coli*? To answer this question, we analyzed the CTDs from *C. crescentus* and *E. coli* using NARDINI **(Figure 7)**. For the CTD of RNase E from *C. crescentus*, we find z-score values being greater than +5 for $\delta_{\mu+}$, $\delta_{\mu h}$, δ_{+-}, and Ω_{+}. In descending order, the most non-random patterns, quantified as patterns with the largest z-scores, are, in descending order, δ_{h+}, δ_{+-}, Ω_{+}, $\delta_{\mu+}$, δ_{h-}, Ω_{-}, Ω_{h}, and $\delta_{\mu-}$, respectively. This highlights a clear preference for linear segregation of positively charged residues, leading to a sequence with a multi-block architecture. Indeed, the CTD of the *C. crescentus* RNase E may be classified as a blocky polyampholyte that also includes clusters of hydrophobic residues.

In contrast, while the preference for segregation of positively charged residues is preserved in the CTD of RNase E from *E. coli*, the z-score values for $\delta_{\mu+}$, δ_{+-}, and Ω_{+} are at least two-fold smaller for this system when compared to those of the CTD of RNase E from *C. crescentus*. This quantitative, rather than qualitative difference in patterning preferences appears to explain the differences in driving forces for phase separation. Specifically, the overall weakening of the linear segregation of charged and hydrophobic residues from one another appears to correlate with the observation of distinct phenotypes and driving forces for CTD mediated phase transitions of RNases E that control the formation of BR bodies.

Our results for the CTD of RNases E suggest that z-score values might be useful as *bona fide* order parameters in theories for phase separation. The use of z-scores, which can be converted to order parameters through suitable normalization would supplant SCD, κ, Ω, and facsimiles thereof that have been used as order parameters for describing sequence-ensemble relationships and phase behaviors of IDPs [50-54]. To use z-scores as generators of multicomponent order parameters, the underlying theories will have to be altered to recast free energy functionals in terms of z-scores rather than the SCD family of parameters [50-54].

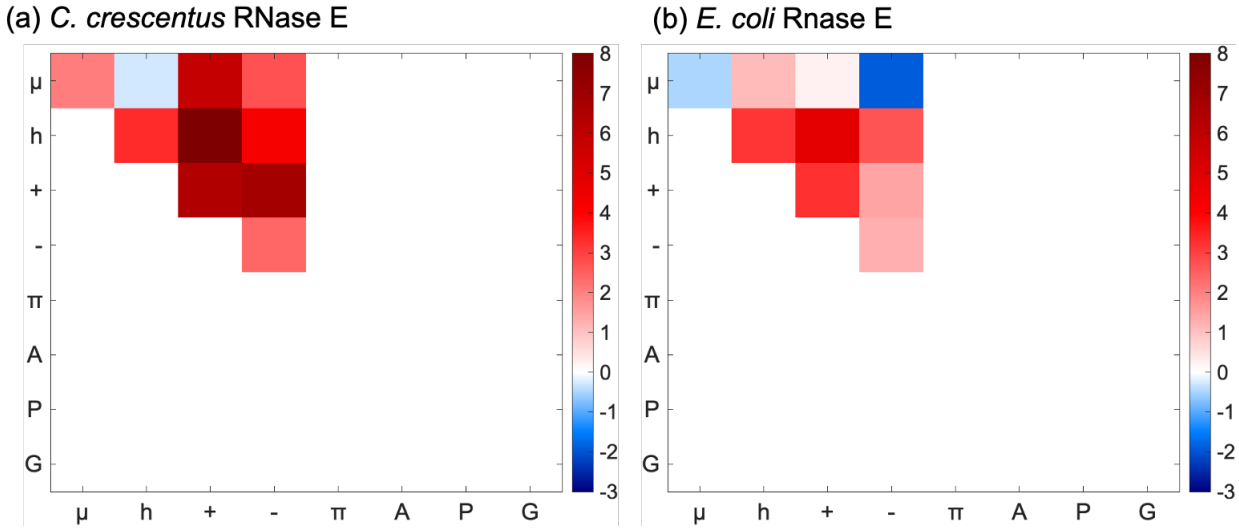

Figure 7. Direct comparison of z-score matrices of RNase E from (a) *C. crescentus* and (b) *E. coli*. Patterns associated with charged residues in *C. crescentus* RNase E (left) are > +2.4 standard deviations away from the null-scramble model in the positive direction. *E. coli* RNase E shows non-random segregation of positive residues and hydrophobic residues as well as from other residues, and hydrophobic residues also contribute to non-random patterns. Unlike the *C. crescentus* RNase E, patterns involving negative residues in *E. coli* RNase E do not significantly deviate from the null-scramble expectation.

Given the strong preferences for segregation of positively charged residues that are shared across CTDs of bacterial RNases E, we set a high threshold for the z-score to distinguish random from non-random features in RNase E ortholog sequences whereby 60% of the CTDs from orthologs would be considered non-random. The resulting cumulative distribution is shown in **Figure S3**, and the threshold z-value was subsequently set to ±1.9. Counting each non-random feature with a z-score that is less than -1.9 or greater than +1.9, we find that over 99% of the deviations are in the positive direction. Given the significant representation of this category, we focused our subsequent analysis on CTD sequences with non-random features that have significant positive deviations from the expected values. Across the complete set of z-score matrices, we tallied the number of positive deviations (z-scores greater than +1.9) observed for each parameter. To compute a positive deviation frequency for each feature, we divided the number of positive deviations by the total number of interrogated sequences. **Figure 8** shows the frequency of non-randomness for each feature. The bar graph on top demonstrates the relative frequency of observing a positive deviation involving each residue / residue type.

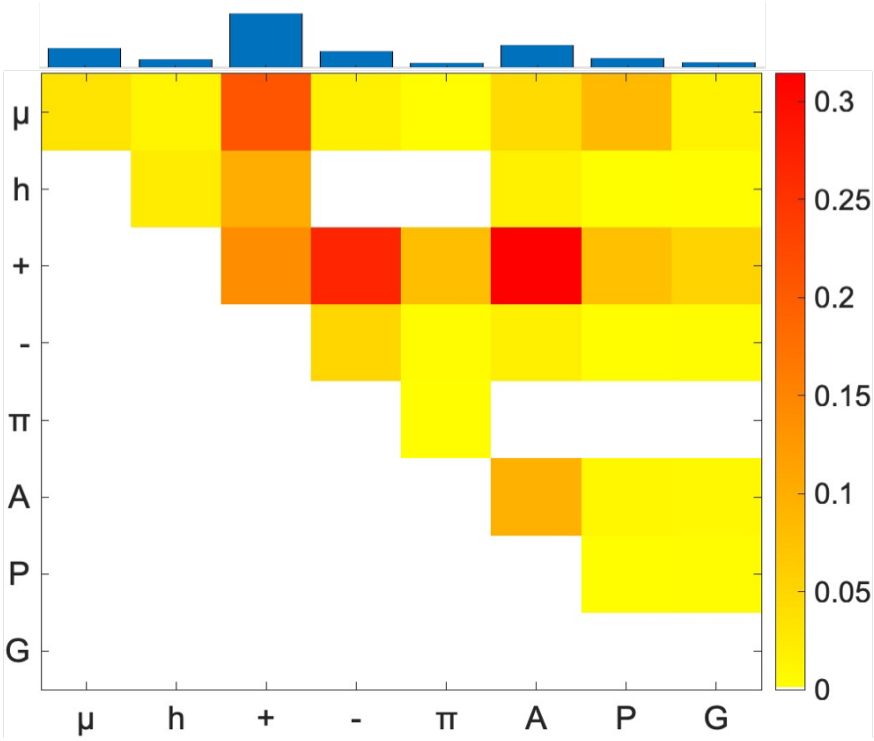

Figure 8. The feature-specific frequency of observing z-scores above 1.9 for the system of RNase E orthologs. 1149 CTD sequences from a system of RNase E were probed to generate the z-score matrices. All z-scores greater than +1.9 were tallied and divided by the total number of sequences to calculate a frequency of observing a positive non-random deviation for each sequence feature. The bar graph on top of the matrix displays the relative frequency of a non-random feature involving each residue / residue type. The z-scores involving positive residues are most frequently above +1.9. The color bar indicates the percentage of orthologs with each feature.

We find that 59.8% of RNase E CTDs contain at least one feature that has a z-score greater than +1.9 with basic residues being most frequently involved in the non-random patterning parameters. Specifically, the preferences for linear segregation of positive charges from negative charges (δ_{+-}) is a prominent patterning feature that is shared across the CTDs from RNases E **(Figure 8)**. However, the distinction between CTDs comes from the quantitative z-scores of δ_{+-} values specific to each sequence, again suggesting that the z-scores are likely to be useful order parameters for comparative modeling phase behaviors of IDPs / IDRs.

Analysis of IDRs in bacterial single-stranded DNA binding proteins (SSBs)

Another example of an essential IDR in bacteria is the intrinsically disordered linker (IDL) in single-stranded DNA binding proteins (SSBs) that plays critical roles in bacterial DNA replication and repair. Their modular architectures include an ordered DNA-binding domain (OB-fold), followed by a hypervariable intrinsically disordered linker (IDL) that is connected to a conserved C-terminal tip [33,55,56]. Recent work has also shown that the SSB from *E. coli* can mediate phase separation with DNA [57]; however, unlike the CTD of RNase E, the sequence requirements for this function have yet to be elucidated. Using NARDINI, we sought to generate testable hypotheses regarding binary patterns that are likely to distinguish functional and non-random IDLs from random ones. We focused on the *E. coli* SSB IDL since it is the most well characterized both in terms of its impact on the cooperativity of ssDNA binding and phase behavior [33,55-63].

We find that for *E. coli* SSB IDL, the distinctly non-random pattern pertains to the linear segregation of Gly from all other residues. Gly residues form a series of short linear clusters, segregated from all other residues in the SSB IDL sequence. The blocks of Gly residues are often interspersed by Pro, giving rise to positive deviations of δ_{PG} from the null model (**Figure 8)**. These features are highlighted in the sequence of the IDL from *E. coli* SSB: TMQML**GG**RQS**GG**APA**GG**NI**GGG**QPQ**GG**W**G**QPQQPQ**GG**NQFS**GG**AQSRPQQSAPAAPSNEPP.

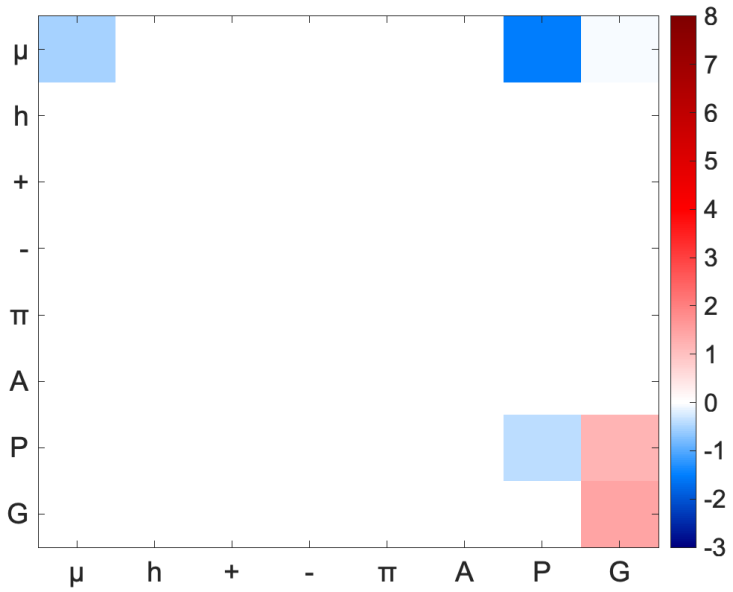

Figure 9. The z-score matrix of *E. coli* SSB IDL. All blank squares on the checkerboard have z-score values that are zero or close to zero. The positioning of glycine, proline, and polar residues along the sequence is significant. The distinctly non-random features include the linear clustering of Gly residues and the punctuation of these clusters by Pro residues giving the IDL its seemingly elastomeric properties – a hypothesis that would be testable based on our predictions.

Cursory analysis suggests that the *E. coli* SSB IDL has many of the features that are reminiscent of elastomeric IDRs that are known to be drivers of responsive phase transitions that control elastic responses of materials such as extracellular matrices [64-67]. Examples of such elastomeric IDRs include Gly-rich regions within resilin, as seen with repeats of PSSSYGAP**GGG**N**GG**R that confer elastic properties. Other examples include stretches such as P**G**Q**G**QQ from Q-rich proteins such as gluten [68], Gly, and Ser rich motifs in silk [69], motifs such as Y**G**H**GG**N/G in cell wall proteins of higher plants [70], and repetitive motifs such as F**GG**M**GGG**K**GG** from abductin, a protein that makes up hinge ligaments that control the swimming behaviors of mollusks [71].

We assessed whether non-random segregation of Gly residues is conserved across a system of 1523 bacterial orthologs of SSB. Using the 60% threshold for identifying sequences with non-

random patterns, we set the z-score threshold value to ±1.5. The cumulative distribution of the magnitude of maximum z-score values is shown in **Figure S4.** Only positive deviations were considered as 99% of the observed deviations were in the positive direction. As with the *E. coli* SSB IDL, we find a recurring theme of non-random linear segregation of Gly from all other residues (**Figure 10)**. This is based on the z-scores we observe for Ω_G. However, unlike *E. coli,* we also find significant segregation of polar residues with respect to all residues – $z(\Omega_\mu)$ is greater than +1.5, compared to $z(\Omega_\mu) \approx 0$ for *E. coli* . The z-score values are also greater than +1.5 for δ_{-G}, δ_{-A}, $\delta_{\mu G}$, and $\delta_{\mu-}$. Accordingly, it appears that the segregation of Gly with respect to acidic and polar residues, as well as the segregation of acidic residues with respect to Ala and polar residues, could be important for the functions of IDLs in bacterial SSBs. These IDLs are known to be different from IDLs in parasitic SSBs [33].

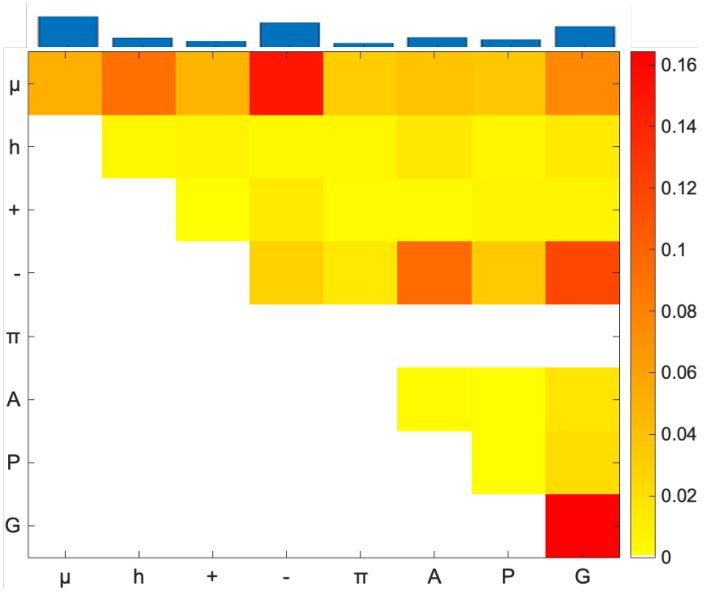

Figure 10. The feature-specific frequency of observing non-random features for the system of SSB orthologs. All z-scores greater than +1.5 were tallied and divided by the total number of sequences to calculate a frequency of observing a positive non-random deviation for each sequence feature. The bar graph displays the relative frequency of a non-random feature involving each residue / residue type. Z-scores involving glycine residues are most frequently above +1.5 and is dominated by three features: glycine, negatively charged residues, and polar residues.

We asked if the linear segregation of Gly and / or polar residues was a phylum- or class-specific feature. Classes that contained more than 5% of the total number of sequences were analyzed separately **(Figure 11)**. Non-random patterning of Gly residues was conserved across three out of the six evaluated classes: actinobacteria, α-proteobacteria, and γ-proteobacteria. The elastomeric features of SSB IDLs are noteworthy because they present testable hypotheses regarding the IDL-mediated phase behavior, which has been reported recently for *E. coli* SSBs [57]. Do SSBs use elastomeric features in the context of heterotypic networks of SSB-interacting proteins [56]? Our analysis presents testable hypotheses for investigations via experiments.

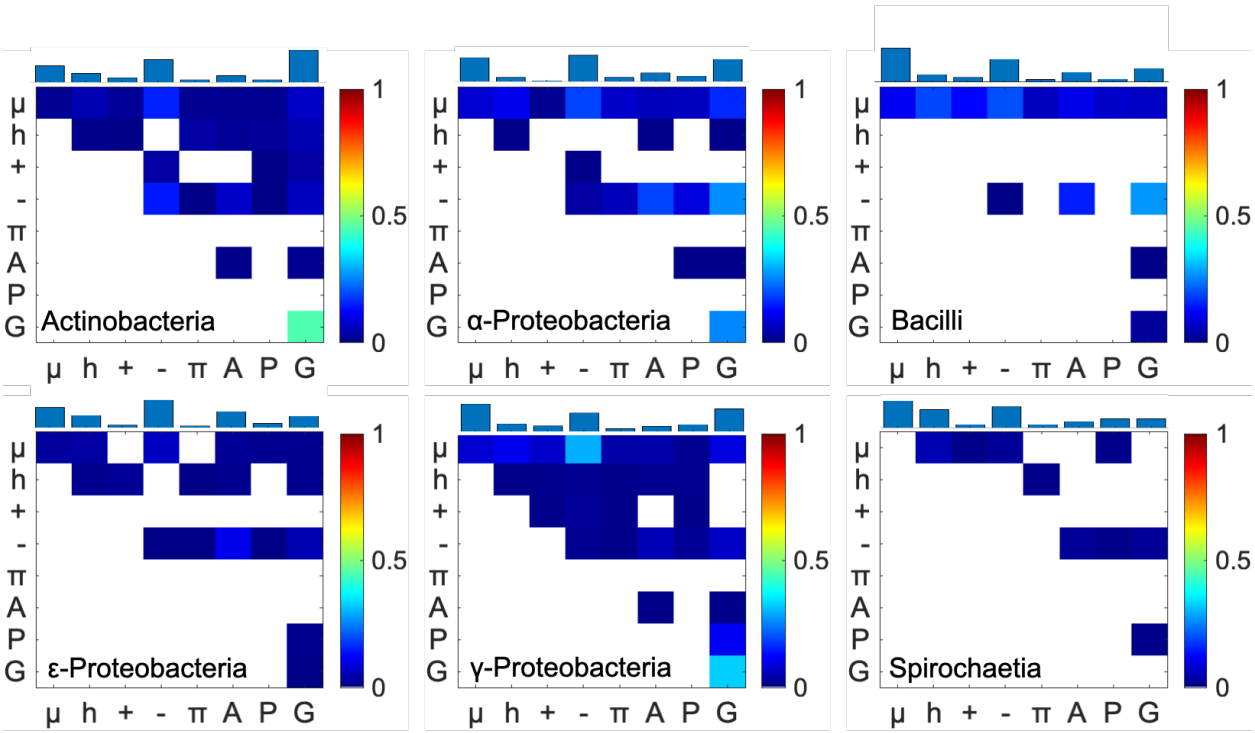

Figure 11. The frequency of observing a non-random pattern for the SSB IDLs. Squares within the checkerboard plot that do not rise about the 10% threshold are shown in white color. In all, we analyzed six phylum classes. These include actinobacteria (n = 190), bacilli (n = 239), γ-proteobacteria (n = 359), α-proteobacteria (n = 122), ε-proteobacteria (n =143), and spirochaetia (n = 101). The bar graph on the top of the matrix represents the relative frequencies of observing a non-random feature (z > +1.5) involving each residue / residue type.

Discussion

We have developed a method to uncover non-random binary patterns within a given IDR. Our work represents a generalization of approaches that have been brought to bear for identifying SLiMs and conserved features in IDRs [21,72-74]. The method, referred to as NARDINI, is available as a plugin for localCIDER [37], and can be deployed across orthologous systems to uncover non-random binary patterns within IDPs / IDRs. NARDINI was designed to address the challenge posed by poor conservation of sequence and composition in hypervariable IDPs / IDRs that makes it difficult to identify patterns likely to be important as determinants of sequence-function relationships [8,9]. This is important because while conserved linear sequence is not required in IDPs / IDRs, specific sequence features, specifically the binary patterns of different categories of residues are important for determining sequence-ensemble [12] and sequence-ensemble-function relationships of IDRs [14]. Such patterns can be masked by poor alignments, and the method we have developed can be deployed independent of sequence alignments, thereby allowing the identification random versus non-random binary patterns in hypervariable IDRs.

Binary patterning parameters measure how a residue, or a residue type is positioned within the sequence with respect to itself and other residues [14,17,34,37,38]. We showed that the original versions of binary patterning parameters labeled κ [38] and Ω [28,39] and adaptations thereof cannot be compared across sequences of different compositions because the value of the patterning parameter that would be expected at random is inherently dependent upon the sequence composition [37]. To overcome this challenge, we use the deviation (z-score) of the observed patterning parameter from a null model where the random scrambling is unbiased. It does not consider the weighted probability of observing such a sequence based upon factors such as codon usage and evolutionary deviation, which recent work aims to consider [75]. Instead, our method is a

zeroth-order approximation of a uniform background that allows for comparisons of hypervariable sequences in a statistically meaningful way.

For comparisons across orthologs, we first set a threshold z-score value to distinguish random and non-random features. This allows us to use the nomenclature of "random" versus "non-random" to emphasize potentially significant patterns within the sequence. The information of interest for identifying significant, non-random binary patterns is contained in the matrix of z-scores, which is sequence-specific. To investigate the presence of similar sequence patterns across orthologs, we set the z-score threshold value such that approximately 60% of the hypervariable sequences have at least one non-random feature. This is not essential, but it is helpful for identifying the minimal number of patterning parameters that mark out members of orthologs. We used the z-score method to identify non-random sequence patterns within the hypervariable C-terminal domains (CTDs) of RNase E from *E. coli* and *C. crescentus*. When swapped with one another, these domains are known to impact function and the *in vivo* phenotype by affecting the nature of the RNA degradasome. In *C. crescentus,* this body is a biomolecular condensate that exhibits liquid-like properties, whereas, in *E. coli,* it is a membrane-bound punctum [48]. This discrepancy has been shown to be dictated by differences in the CTD sequence. In the CTD of *C. crescentus* RNase E, the blocky charge architecture of this sequence is a feature that significantly deviates from the null expectation. While the positively charged residues of the *E.* coli CTD are non-randomly positioned (Ω_{+}), the negative residues are randomly positioned in relation to positively charged residues (δ_{+-}) as well as in relation to other groups of residues (Ω_{-}).The non-random segregation of charged residues (δ_{+-}) is a feature that is consistent with over 27% of RNase E CTDs, and this could be a feature that is relevant for RNA binding within the degradasome. It appears that the RNases E provide an example of adaptation, whereby the charge patterning and

other features within the CTD are the target of evolutionary shuffling to provide control over the formation of BR bodies.

In the IDL of *E. coli* SSB, we observed a conserved non-random patterning of Gly residues (Ω_G) whereby Gly residues are frequently found in short linear clusters along the sequence. We hypothesize that this feature could have implications for the cooperative binding of SSBs to single-stranded DNA and for the driving forces for phase-separation of SSBs [33,55,57-59,61-63]. SSBs form homo-tetramers that generate a tetra-valent system to coordinate interactions with SSB interacting proteins [56,61]. The *E. coli* SSB tetramer binds cooperatively to single-stranded DNA in different binding modes where the binding modes are classified by the number of nucleotides that are occluded by individual tetramers [59,61,62]. Cooperativity of single-stranded DNA binding is governed by sequence features of the IDL [55]. Specifically, cooperativity is enhanced when the IDL has features that are akin to low complexity domains enriched in polar amino acids, primarily Pro, Gln, and Gly [76-79]. Conversely, cooperativity is diminished for long IDLs enriched in charged residues [33,55]. The enrichment of Pro, Gln, and Gly that appear to govern cooperativity and their patterning were observed to be significant non-random patterns in the z-score analysis.

A recent study has shown that in response to DNA damage, membrane associated SSBs form condensates at the sites of DNA damage [57]. These condensates are multicomponent bodies and concentrate other factors that contribute to DNA processing and metabolism. *In vitro* studies showed that the IDL is essential for driving the formation of liquid-like condensates. It could be that the patterning of Gly residues plays an important role in condensate formation – a feature that would be true of elastomeric sequences [67,80].

NARDINI allows the identification of non-random binary patterns within a sequence and comparative analyses of non-random patterns across sequences. The work presented offers a new

methodology to compare hypervariable sequences and drive inferences about conserved sequence parameters to function relationships. We look forward to its usage as a driver of IDP / IDR design for uncovering sequence-ensemble-function relationships and for identifying features that are likely selected for among functional orthologs. It is worth noting that our analysis does not account for the fact that the IDRs invariably integrate the proteins of interest into more extensive protein-protein interaction networks. The selection pressures exerted by interaction networks and the covariation of sequence patterns among IDRs that make up nodes in networks remains unclear. Further analysis will be required to uncover the potential co-evolution of patterning features across orthologous proteins that interact with RNase E and SSB, for example. Finally, although our analysis focused on binary patterns, there may be higher-order correlations that are crucial determinants of sequence-function relationships encoded by hypervariable IDRs. The topic of higher-order correlations in sequence patterns is a focus for continued development of NARDINI.

Acknowledgments

We thank Kiersten Ruff for helpful comments and critical analysis of the codebase as well as results generated using NARDINI. We are grateful to Kiersten Ruff and Mina Farag for critical reading of the manuscript. The source code is available as part of version 3.0 of localCIDER that is distributed through the Pappu lab Github. Version 3.0 of CIDER was written by Alex Keeley, a summer 2020 programmer in the Pappu lab. This work was supported by grants from the US National Science Foundation (MCB 1614766) and the US National Institutes of Health (5R01NS056114). Min Kyung Shinn is supported in part by the Center for Science & Engineering of Living Systems (CSELS) in the McKelvey School of Engineering at Washington University in St. Louis as a CSELS postdoctoral fellow.

References

1. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T., (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol,* **337,** 635-645.

2. Peng, Z., Mizianty, M. J., Kurgan, L., (2014). Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins: Structure, Function, and Bioinformatics,* **82,** 145-158.

3. Wright, P. E., Dyson, H. J., (2009). Linking folding and binding. *Current Opinion in Structural Biology,* **19,** 31-38.

4. Dishman, A. F., Tyler, R. C., Fox, J. C., Kleist, A. B., Prehoda, K. E., Babu, M. M., et al., (2021). Evolution of fold switching in a metamorphic protein. *Science,* **371,** 86.

5. Wiley, E. O., Lieberman, B. S., (2011). Phylogenetics: theory and practice of phylogenetic systematics: John Wiley & Sons.

6. Higgins, D. G., Sharp, P. M., (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene,* **73,** 237-244.

7. Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., Sternberg, M. J., (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *Journal of molecular biology,* **269,** 423-439.

8. Zarin, T., Tsai, C. N., Nguyen Ba, A. N., Moses, A. M., (2017). Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proceedings of the National Academy of Sciences,* **114,** E1450-E1459.

9. Zarin, T., Strome, B., Ba, A. N. N., Alberti, S., Forman-Kay, J. D., Moses, A. M., (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife,* **8,** e46883.

10. Cohan, M. C., Pappu, R. V., (2020). Making the Case for Disordered Proteins and Biomolecular Condensates in Bacteria. *Trends Biochem Sci,* **45,** 668-680.

11. Stoneking, M., (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *The American Journal of Human Genetics,* **67,** 1029-1032.

12. Cohan, M. C., Ruff, K. M., Pappu, R. V., (2019). Information theoretic measures for quantifying sequence-ensemble relationships of intrinsically disordered proteins. *Protein Eng Des Sel,* **32,** 191-202.

13. Lange, J., Wyrwicz, L. S., Vriend, G., (2016). KMAD: knowledge-based multiple sequence alignment for intrinsically disordered proteins. *Bioinformatics,* **32,** 932-936.

14. Sherry, K. P., Das, R. K., Pappu, R. V., Barrick, D., (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proceedings of the National Academy of Sciences,* **114,** E9243-E9252.

15. Moesa, H. A., Wakabayashi, S., Nakai, K., Patil, A., (2012). Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Molecular BioSystems,* **8,** 3262-3273.

16. Beh, L. Y., Colwell, L. J., Francis, N. J., (2012). A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proceedings of the National Academy of Sciences,* **109,** E1063.

17. Das, R. K., Ruff, K. M., Pappu, R. V., (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Current Opinion in Structural Biology,* **32,** 102-112.

18. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., et al., (2021). Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. *bioRxiv,* 2021.2001.2001.425046.

19. Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., Rost, B., (2011). Protein disorder—a breakthrough invention of evolution? *Current opinion in structural biology,* **21,** 412-418.

20. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L., Pappu, R. V., (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences,* **107,** 8183-8188.

21. Zarin, T., Tsai, C. N., Nguyen Ba, A. N., Moses, A. M., (2017). Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc Natl Acad Sci U S A,* **114,** E1450-E1459.

22. Strickfaden, S. C., Winters, M. J., Ben-Ari, G., Lamson, R. E., Tyers, M., Pryciak, P. M., (2007). A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell,* **128,** 519-531.

23. Ba, A. N. N., Yeh, B. J., Van Dyk, D., Davidson, A. R., Andrews, B. J., Weiss, E. L., et al., (2012). Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science signaling,* **5,** rs1-rs1.

24. Davey, N. E., Cyert, M. S., Moses, A. M., (2015). Short linear motifs - ex nihilo evolution of protein regulation. *Cell communication and signaling : CCS,* **13,** 43.

25. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., et al., (2012). Attributes of short linear motifs. *Molecular BioSystems,* **8,** 268-281.

26. Gomes, E., Shorter, J., (2019). The molecular language of membraneless organelles. *Journal of Biological Chemistry* **294,** 7115-7127.

27. Martin, E. W., Mittag, T., (2018). Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry,* **57,** 2478-2487.

28. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., et al., (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science,* **367,** 694-699.

29. Chong, P. A., Vernon, R. M., Forman-Kay, J. D., (2018). RGG/RG motif regions in RNA binding and phase separation. *Journal of molecular biology,* **430,** 4650-4665.

30. Varadi, M., Zsolyomi, F., Guharoy, M., Tompa, P., (2015). Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One,* **10,** e0139731.

31. Boeynaems, S., Bogaert, E., Kovacs, D., Konijnenberg, A., Timmerman, E., Volkov, A., et al., (2017). Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. *Molecular Cell* **65,** 1044-1055 e1045.

32. Buske, P. J., Mittal, A., Pappu, R. V., Levin, P. A., (2015). An intrinsically disordered linker plays a critical role in bacterial cell division. *Seminars in cell & developmental biology,* **37,** 3-10.

33. Kozlov, A. G., Weiland, E., Mittal, A., Waldman, V., Antony, E., Fazio, N., et al., (2015). Intrinsically disordered C-terminal tails of E. coli single-stranded DNA binding protein regulate cooperative binding to single-stranded DNA. *J Mol Biol,* **427,** 763-774.

34. Das, R. K., Huang, Y., Phillips, A. H., Kriwacki, R. W., Pappu, R. V., (2016). Cryptic sequence features within the disordered protein p27^{Kip1} regulate cell cycle signaling. *Proceedings of the National Academy of Sciences,* 201516277.

35. Beveridge, R., Migas, L. G., Das, R. K., Pappu, R. V., Kriwacki, R. W., Barran, P. E., (2019). Ion mobility mass spectrometry uncovers the impact of the patterning of oppositely charged residues on the conformational distributions of intrinsically disordered proteins. *Journal of the American Chemical Society,* **141,** 4908-4918.

36. Banjade, S., Wu, Q., Mittal, A., Peeples, W. B., Pappu, R. V., Rosen, M. K., (2015). Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proceedings of the National Academy of Sciences,* **112,** E6426.

37. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O., Pappu, R. V., (2017). CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical journal,* **112,** 16-21.

38. Das, R. K., Pappu, R. V., (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences,* **110,** 13392-13397.

39. Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V., Mittag, T., (2016). Sequence determinants of the conformational properties of an intrinsically disordered protein prior

to and upon multisite phosphorylation. *Journal of the American Chemical Society,* **138,** 15323-15335.

40. Das, R. K., Ruff, K. M., Pappu, R. V., (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol,* **32,** 102-112.

41. Buske, P. J., Mittal, A., Pappu, R. V., Levin, P. A., (2015). An intrinsically disordered linker plays a critical role in bacterial cell division. *Semin Cell Dev Biol,* **37,** 3-10.

42. Sawle, L., Ghosh, K., (2015). A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of Chemical Physics,* **143,** 085101.

43. Lyle, N., Das, R. K., Pappu, R. V., (2013). A quantitative measure for protein conformational heterogeneity. *The Journal of chemical physics,* **139,** 09B607_601.

44. Kim, H. J., Kim, N. C., Wang, Y.-D., Scarborough, E. A., Moore, J., Diaz, Z., et al., (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature,* **495,** 467-473.

45. Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, Anderson P., Kim, Hong J., et al., (2015). Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell,* **163,** 123-133.

46. Ait-Bara, S., Carpousis, A. J., (2015). RNA degradosomes in bacteria and chloroplasts: classification, distribution and evolution of RNase E homologs. *Molecular Microbiology,* **97,** 1021-1135.

47. Ait-Bara, S., Carpousis, A. J., Quentin, Y., (2015). RNase E in the gamma-Proteobacteria: conservation of intrinsically disordered noncatalytic region and molecular evolution of microdomains. *Molecular genetics and genomics : MGG,* **290,** 847-862.

48. Al-Husini, N., Tomares, D. T., Bitar, O., Childers, W. S., Schrader, J. M., (2018). alpha-Proteobacterial RNA Degradosomes Assemble Liquid-Liquid Phase-Separated RNP Bodies. *Molecular Cell,* **71,** 1027-1039 e1014.

49. Lin, Y.-H., Forman-Kay, J. D., Chan, H. S., (2016). Sequence-Specific Polyampholyte Phase Separation in Membraneless Organelles. *Physical Review Letters,* **117,** 178101.

50. Shea, J.-E., Best, R. B., Mittal, J., (2021). Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Current Opinion in Structural Biology,* **67,** 219-225.

51. Zheng, W., Dignon, G., Brown, M., Kim, Y. C., Mittal, J., (2020). Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *The Journal of Physical Chemistry Letters,* **11,** 3408-3415.

52. Das, S., Amin, A. N., Lin, Y.-H., Chan, H. S., (2018). Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Physical Chemistry Chemical Physics,* **20,** 28558-28574.

53. Lin, Y.-H., Chan, H. S., (2017). Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophysical journal,* **112,** 2043-2046.

54. Rana, U., Brangwynne, C. P., Panagiotopoulos, A. Z., (2021). Phase separation versus aggregation behavior for model disordered proteins. *bioRxiv,* 2021.2006.2016.448686.

55. Kozlov, A. G., Shinn, M. K., Weiland, E. A., Lohman, T. M., (2017). Glutamate promotes SSB protein-protein Interactions via intrinsically disordered regions. *J Mol Biol,* **429,** 2790-2801.

56. Shinn, M. K., Kozlov, A. G., Nguyen, B., Bujalowski, W. M., Lohman, T. M., (2019). Are the intrinsically disordered linkers involved in SSB binding to accessory proteins? *Nucleic Acids Research,* **47,** 8581-8594.

57. Harami, G. M., Kovacs, Z. J., Pancsa, R., Palinkas, J., Barath, V., Tarnok, K., et al., (2020). Phase separation by ssDNA binding protein controlled via protein-protein and protein-DNA interactions. *Proc Natl Acad Sci U S A,* **117,** 26206-26217.

58. Shereda, R. D., Kozlov, A. G., Lohman, T. M., Cox, M. M., Keck, J. L., (2008). SSB as an Organizer/Mobilizer of Genome Maintenance Complexes. *Critical Reviews In Biochemistry And Molecular Biology,* **43,** 289-318.

59. Kozlov, A. G., Galletto, R., Lohman, T. M., (2012). SSB–DNA binding monitored by fluorescence intensity and anisotropy. Single-Stranded DNA Binding Proteins: Springer. p. 55-83.

60. Kozlov, A. G., Shinn, M. K., Weiland, E. A., Lohman, T. M., (2017). Glutamate promotes SSB protein–protein Interactions via intrinsically disordered regions. *Journal of Molecular Biology,* **429,** 2790-2801.

61. Antony, E., Lohman, T. M., (2019). Dynamics of E. coli single stranded DNA binding (SSB) protein-DNA complexes. Seminars in cell & developmental biology: Elsevier. p. 102-111.

62. Dubiel, K., Myers, A. R., Kozlov, A. G., Yang, O., Zhang, J., Ha, T., et al., (2019). Structural mechanisms of cooperative DNA binding by bacterial Single-Stranded DNA-Binding proteins. *Journal of molecular biology,* **431,** 178-195.

63. Kozlov, A. G., Shinn, M. K., Lohman, T. M., (2019). Regulation of Nearest-Neighbor cooperative binding of E. coli SSB protein to DNA. *Biophysical journal,* **117,** 2120-2140.

64. Dzuricky, M., Roberts, S., Chilkoti, A., (2018). Convergence of Artificial Protein Polymers and Intrinsically Disordered Proteins. *Biochemistry,* **57,** 2405-2414.

65. Li, L., Charati, M. B., Kiick, K. L., (2010). Elastomeric polypeptide-based biomaterials. *Polymer Chemistry,* **1,** 1160-1170.

66. Quiroz, F. G., Li, N. K., Roberts, S., Weber, P., Dzuricky, M., Weitzhandler, I., et al., (2019). Intrinsically disordered proteins access a range of hysteretic phase separation behaviors. *Science advances,* **5,** eaax5177.

67. Roberts, S., Dzuricky, M., Chilkoti, A., (2015). Elastin-like polypeptides as models of intrinsically disordered proteins. *FEBS letters,* **589,** 2477-2486.

68. Athamneh, A. I., Barone, J. R., (2009). Enzyme-mediated self-assembly of highly ordered structures from disordered proteins. *Smart materials and structures,* **18,** 104024.

69. Holland, G. P., Creager, M. S., Jenkins, J. E., Lewis, R. V., Yarger, J. L., (2008). Determining secondary structure in spider dragline silk by carbon-carbon correlation solid-state NMR spectroscopy. *J Am Chem Soc,* **130,** 9871-9877.

70. Ringli, C., Keller, B., Ryser, U., (2001). Glycine-rich proteins as structural components of plant cell walls. *Cellular and Molecular Life Sciences CMLS,* **58,** 1430-1441.

71. Cao, Q., Wang, Y., Bayley, H., (1997). Sequence of abductin, the molluscan 'rubber'protein. *Current biology,* **7,** R677-R678.

72. Davey, N. E., Cyert, M. S., Moses, A. M., (2015). Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun Signal,* **13,** 43.

73. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., et al., (2012). Attributes of short linear motifs. *Mol Biosyst,* **8,** 268-281.

74. Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., et al., (2014). Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev,* **114,** 6733-6778.

75. Zarin, T., Strome, B., Nguyen Ba, A. N., Alberti, S., Forman-Kay, J. D., Moses, A. M., (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife,* **8,** e46883.

76. Bujalowski, W., Lohman, T. M., (1986). Escherichia coli single-strand binding protein forms multiple, distinct complexes with single-stranded DNA. *Biochemistry,* **25,** 7799-7802.

77. Lohman, T. M., Overman, L. B., Datta, S., (1986). Salt-dependent changes in the DNA binding co-operativity of Escherichia coli single strand binding protein. *J Mol Biol,* **187,** 603-615.

78. Marceau, A. H., Bahng, S., Massoni, S. C., George, N. P., Sandler, S. J., Marians, K. J., et al., (2011). Structure of the SSB-DNA polymerase III interface and its role in DNA replication. *EMBO J,* **30,** 4236-4247.

79. Bianco, P. R., Pottinger, S., Tan, H. Y., Nguyenduc, T., Rex, K., Varshney, U., (2017). T he IDL of E. coli SSB links ssDNA and protein binding by mediating protein–protein interactions. *Protein Science,* **26,** 227-241.

80. MacEwan, S. R., Chilkoti, A., (2010). Elastin-like polypeptides: Biomedical applications of tunable biopolymers. *Peptide Science,* **94,** 60-77.