Intrinsically Disordered Compositional Bias in Proteins: Sequence Traits, Region Clustering, and Generation of Hypothetical Functional Associations

Bioinformatics and Biology Insights
Volume 18: 1–15

DOI: 10.1177/11779322241287485

Paul M Harrison

Department of Biology, McGill University, Montreal, QC, Canada.

ABSTRACT: Compositionally biased regions (CBRs), ie, tracts that are dominated by a subset of residue types, are common features of eukaryotic proteins. These are often found bounded within or almost coterminous with intrinsically disordered or 'natively unfolded' parts. Here, it is investigated how the function of such intrinsically disordered compositionally biased regions (ID-CBRs) is directly linked to their compositional traits, focusing on the well-characterized yeast (*Saccharomyces cerevisiae*) proteome as a test case. The ID-CBRs that are clustered together using compositional distance are discovered to have clear functional linkages at various levels of diversity. The specific case of the Sup35p and Rnq1p proteins that underlie causally linked prion phenomena ([PSI+] and [RNQ+]) is highlighted. Their prion-forming ID-CBRs are typically clustered very close together indicating some compositional engendering for [RNQ+] seeding of [PSI+] prions. Delving further, ID-CBRs with distinct types of residue patterning such as 'blocking' or relative segregation of residues into homopeptides are found to have significant functional trends. Specific examples of such ID-CBR functional linkages that are discussed are: Q/N-rich ID-CBRs linked to transcriptional coactivation, S-rich to transcription-factor binding, R-rich to DNA-binding, S/E-rich to protein localization, and D-rich linked to chromatin remodelling. These data may be useful in informing experimental hypotheses for proteins containing such regions.

KEYWORDS: Intrinsic disorder, compositional bias, prions, Sup35, Rnq1, protein function, yeast

RECEIVED: May 20, 2024. **ACCEPTED:** August 27, 2024.

TYPE: Research Article

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was performed largely on a Mac Pro computer funded partly by the National Sciences & Engineering Research Council of Canada and by the Canadian Fund of Innovations and on 2 Mac Mini computers funded by the National Sciences & Engineering Research Council of Canada. Computations of blockiness were run on a node cluster of the Digital Alliance of Canada.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Paul M Harrison, Department of Biology, McGill University, Montreal, QC H3A 1B1, Canada. Email: paul.harrison@mcgill.ca

Introduction

Eukaryotic proteins often contain regions that demonstrate a compositional bias for a subset of the 20 amino acids. For example, the tract PSPEPPSESSPSPPSSTSPSTPPP is biased for proline (P) and serine (S). Such compositionally biased regions (CBRs) range from highly biased and repetitive to quite mildly skewed in residue usage.

Intrinsically disordered regions (IDRs) are stretches of protein chains that remain unfolded during part of the functioning.[1,2] The IDRs are implicated in a variety of functional roles including kinase binding, transcription and translation regulation, histone binding, and chromatin remodelling.[2,3] The sequences of IDRs often contain compositional bias, or 'low-complexity', a related concept.[4,5] They are also enriched in short runs of amino acids, sometimes termed homopeptides.[6,7] The CBRs involved in protein interaction networks can be enriched in short linear motifs, particularly rich in serine and proline.[8] Such CBRs can have also have residue patterning that may have functional significance such as residue dispersion,[9] repetitiveness,[10] in addition to homopeptide content.[6]

Here, I formally define intrinsically disordered compositionally biased regions (ID-CBRs) – CBRs that are intrinsically disordered – and cluster them using compositional distance and measures of residue patterning, ie, the degree of bunching or 'blockiness' of residues along the ID-CBR sequences and the amount of homopeptides in them. Functional hypotheses are extracted that are linked to clusters with specific predominant compositional tendencies and types of residue patterning.

Methods

Proteome data

The proteome of budding yeast *Saccharomyces cerevisiae* strain 288c (UP000002311) was downloaded from UniProt in September 2023.[11]

Protein families

The yeast proteome was clustered into protein families using an algorithm described previously that uses the BLASTP program.[12,13] Briefly, similarities that accord to an e-value threshold ($=1 \times 10^{-4}$) and an alignment coverage threshold were tallied and sequences then sorted on decreasing number of similarities. This sorted list was searched to progressively add sequences to families and de-select them from further consideration. The 2 applied alignment coverage thresholds were 50% and 33%, but the former was used primarily (Supplemental File 1). These coverage thresholds were the percentage lengths of both sequences in a pair-wise similarity that need to be included in the respective aligned parts. These protein families were labelled with a numeric family index.

Table 1. Parameter sets examined and totals of hypothetical GO enrichments/depletions for ID-CBRs for each parameter set.

PARAMETER SETS EXAMINED				TOTALS OF HYPOTHETICAL GO ENRICHMENTS/ DEPLETIONS FOR ID-CBRS FOR EACH PARAMETER SET		
FLPS PARAMETER SET (M, M, T)	TARGET LENGTH (LEN=)[A]	ESTIMATED PROTEOME COVERAGE (COV=)[A]	PARAMETER SET LABEL (LEN=X_ COV=Y)	CLUSTERS	GO ENRICHMENTS	GO DEPLETIONS
7, 11, 5.2e−05	10	5%	len=10_cov=5%	183	294	39
12, 16, 5.4e−06	20	5%	len=20_cov=5%	179	283	33
10, 20, 1.8e−05	20	10%	len=20_cov=10%	221	351	51
9, 30, 6.9e−04	20	25%	len=20_cov=25%	412	650	347
21, 25, 6.2e−09	50	5%	len=50_cov=5%	119	182	18
23, 33, 4.1e−07	50	10%	len=50_cov=10%	153	242	28
20, 50, 4.7e−05	50	25%	len=50_cov=25%	233	360	65
32, 36, 7.9e−14	100	5%	len=100_cov=5%	75	135	14
38, 48, 7.3e−10	100	10%	len=100_cov=10%	125	224	27
34, 74, 5.2e−07	100	25%	len=100_cov=25%	161	295	40
68, 78, 4.1e−18	250	10%	len=250_cov=10%	72	142	14
78, 128, 2.1e−11	250	25%	len=250_cov=25%	89	181	29

[a]Source: From Harrison.[16]

Compositional bias

The program fLPS was employed to annotate CBRs in yeast.[14,15] The CBRs are labelled with a bias signature, which is the list of biasing residues sorted in decreasing order of contribution with the 'primary bias' first, and a *P*-value indicating the degree of compositional bias.[14] For any program that annotates CBRs or low-complexity regions, different parameter sets can be chosen that target regions of a certain length or degree of bias.[16] Degree of bias can be conceived as how much of the sequences are 'covered' by a parameter set. For example, parameters can be picked that target regions of, say, length = 15 residues and cover ~10% of the sequences. To detect trends that are independent of parameter choice, and also tendencies that only occur for shorter or for longer regions, a range of 12 parameter sets were applied (Table 1). These parameter sets are labelled according to the length and coverage that they target, eg, len = 50_cov = 10% is for a target length of 50 residues and a target proteome coverage of ~10%.

To limit redundancy, lists of CBRs for any given parameter set were filtered to remove any with the same primary bias whose ends were ⩽10 residues from the ends of a CBR with a smaller *P*-value. Also, only 1 region of the same bias signature was selected for each sequence, to avoid overpopulating cluster trees.

Intrinsic disorder

Intrinsic disorder annotations for yeast were taken from the MobiDB database, namely, *curated-disorder-merge* (experimentally determined) and *prediction-disorder-alphafold* (algorithmic annotations).[17]

Prions and prion-like regions

A list of amyloid-based prion-forming proteins in yeast was formed by updating a list analysed in a previous paper, with more recent examples[18,19] (Supplemental Table 1). A list of intrinsically disordered prion proteins was obtained from the work of Chakrabortee et al[20] (Supplemental Table 1).

Regions with prion-like composition were identified with PLAAC.[21] The PDR score in PLAAC output was inspected (or the LLR score, failing no calculation of a PDR score). Two thresholds were considered: >0.0 or ⩾15.0, as before.[22]

Structural features

Coiled coils were defined using DeepCoil.[23] Assignments across whole protein sequences to atom-record sequences of ASTRALSCOP protein domains (version 2.08) were made using BLASTP and e-value threshold 1×10^{-4}.[12,24]

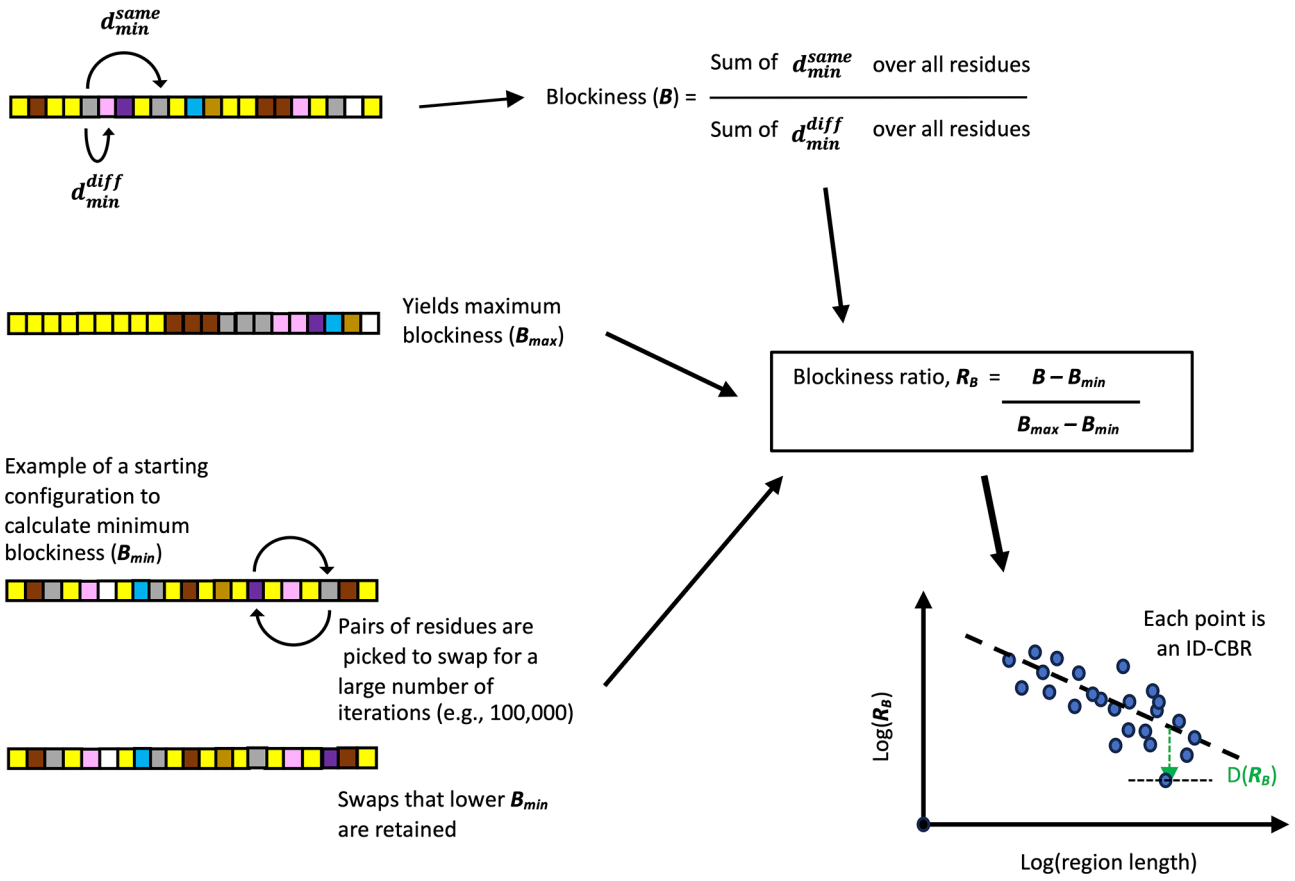

Figure 1. Method for calculating blockiness of ID-CBRs. The blockiness ratio (R_B) is derived from the blockiness (B) of the ID-CBR sequence and from B_{max} and B_{min} as depicted. A plot of log(R_B) vs log(region length) is drawn for a parameter set for an amino-acid primary bias. Here, it would be for the yellow residues. Then, deviations from the linear regression line, D(R_B), are used to characterize the blockiness of the ID-CBR.

Nuclear localization signals

Nuclear localization signals (NLS) sequences for yeast were downloaded from the NLSdb database.[25] Only cases that were labelled 'experimental' or 'by expert' were extracted.

Definition of intrinsically disordered compositionally biased regions

An ID-CBR is a CBR that is likely intrinsically disordered. A CBR was labelled as an ID-CBR if >50% of it was annotated in any specific IDR, or it had an overall disorder propensity >0.0. The former is termed the *overlap criterion*, and the latter the *propensity criterion*. These regional disorder propensities were summed from individual amino-acid residue disorder propensities, which were given by $\log[f_{dis}/f_{struct}]$, where f_{dis} is the amino-acid frequency in a non-redundant set of experimentally determined IDRs taken from the DISPROT database,[26] and f_{struct} is the frequency in the 40% set of protein domain sequences from ASTRAL version 2.08, but using the atom-record sequences, to avoid inclusion of intrinsically disordered protein loops.[24] The DISPROT IDRs were made non-redundant by reducing to a set of cluster representatives arising from application of the protein-family clustering algorithm described above.

Furthermore, as prion-like regions are a subset of IDRs, prion propensities were considered in a similar way, but with a prion propensity scale derived from amino-acid frequencies in known prion-forming domains, compared with the overall frequencies in the yeast proteome (denoted Pr_{prion}). Also, >50% overlap with an annotated PLAAC prion-like region was used as a criterion.

Blockiness of intrinsically disordered compositionally biased regions

The distribution of residues along the expanse of an ID-CBR can vary quite substantially. One aspect is their degree of bunching or 'blockiness' (Figure 1). At one extreme, the most 'blocky', all of the residues are segregated from each other in decreasing (or increasing) order of frequency from one end of the sequence to the other. At intermediate levels of blockiness, there may be smaller 'islands' of different residues; whereas at the other extreme amino acids of a specific type try to be as distant as possible from their fellows (Figure 1). This is quite easy to conceptualize for 2 residue types (they alternate), but

for many residue types, the solution is not trivial. A measure of such blockiness (B) was derived using the following formula:

$$B = \frac{\sum_{i=1}^{L} d_{min}^{diff}}{\sum_{i=1}^{L} d_{min}^{same}}$$

where L is the length of the sequence being considered, and d_{min}^{diff} is the smallest interval from residue i to any residue of a different type, and d_{min}^{same} is the smallest interval to a residue of the same type. The value of B is compared with minimum and maximum values, in the form of a ratio (R_B) given by:

$$R_B = \frac{B - B_{min}}{B_{max} - B_{min}}$$

B_{max} is calculated from the maximally blocky arrangement depicted in Figure 1. To estimate B_{min}, first residues of each type in decreasing order of frequency were filled into positions in a sequence of length L such that they were as far away as possible from any fellow residues of the same type. Then, the sequence was perturbed over 100 000 iterations by switching residues of different types and keeping any changes that lead to a lower B_{min} estimate (Figure 1). R_B is examined on log-log scatter plots vs region length. Linear regression lines are fitted and where they have significant correlation ($P < .01$), the deviation D(R_B) from the linear regression line is used as an indicator of blockiness (negative for less blocky or not blocky, positive for more blocky). These deviations are used to divide data into tertiles with high (H), intermediate (I), or low (L) blockiness.

Homopeptide content

As before, homopeptides were defined as runs of amino acids with a minimum count of 3 residues.[6] The proportion of residues in homopeptides (denoted *hpep*) was calculated for each ID-CBR. Specific data sets are separated into tertiles with high, intermediate, or low homopeptide content.

Clustering of intrinsically disordered compositionally biased regions using compositional distance

Compositional distance (D_{comp}) was used to characterize the differences in amino-acid usage across ID-CBR populations. D_{comp} is given by:

$$D_{comp} = \sqrt{\sum_{i} \left(f_1^i - f_2^i \right)^2}$$

The summation is over all 20 amino-acid types i, for comparing ID-CBRs *1* and *2*. All-to-all comparison of each ID-CBR population arising from an fLPS parameter set was performed. This yielded for each population a distance matrix. This distance matrix was fed into the *neighbour* program from the PHYLIP package (version 3.695),[27] to make a neighbour-joining (NJ) tree in Newick format,[28] out of which an exhaustive list of clusters and cluster members was extracted iteratively. This was achieved by at first searching for clustered pairs, and then expanding outwards from these pairs to define further larger clusters. This NJ analysis is not to infer evolutionary descent, but simply to extract clusters of ID-CBRs that are compositionally similar. The number of ID-CBRs in these trees ranges from 932 to 25 125.

The ID-CBR clusters were labelled with the most common bias occurring among the cluster members, and this was termed a *consensus bias signature*. This was set equal to the most common residue type at each position in the set of bias signatures, if the frequency of this is ⩾50% (if this frequency is <50% for the first position, the consensus is simply 'X'). Contributions to the consensus from beyond the first position are included progressively if similarly they are included in ⩾50% of sequences.

Intrinsically disordered compositionally biased region cluster drawing

The illustrative example tree was drawn in Evolview.[29] ID-CBRs are named: *UniProtAccession_Start_End_Bias Signature*. They were also labelled with numeric protein-family indices, where appropriate.

Gene Ontology data

Gene Ontology (GO) term annotations for the yeast proteome were downloaded from geneontology.org in December 2023.[30] Also obtained was the complete GO digraph (file *go-basic.obo*). This was parsed recursively into a list of over terms for each lower term, an *over-term* being any term above a lower term in the digraph and connected to it.

Generation of hypothetical functional associations

Significant enrichments and depletions of GO terms were calculated for each ID-CBR cluster using hypergeometric probability and a Bonferroni correction, with the P-value threshold for significance being divided by the total number of GO terms in the yeast annotation file for protein-family representatives. Duplicated term annotations for the same protein sequence were not counted. Where the count of terms for an ID-CBR cluster was zero, a normal approximation to the hypergeometric distribution was applied, with an equivalent z-score threshold. Statistics were calculated for the proteome clustered into protein families using the protein-family clustering algorithm described above and the 50% coverage criterion.

The ID-CBR clusters overlap and have much common membership. So, to derive a list of distinct significant hypothetical enrichments/depletions (abbreviated ED), the principle of parsimony was applied, ie, the most concise explanation

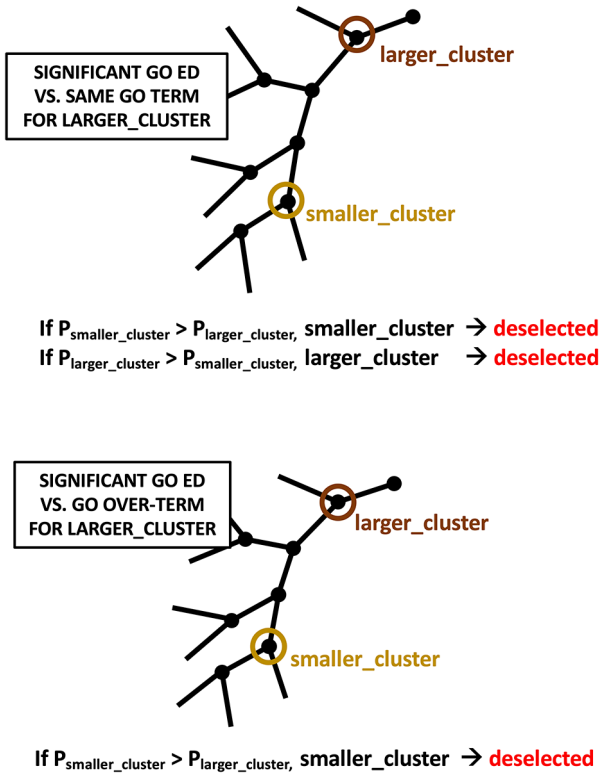

Figure 2. Parsimonious reduction to a list of distinct GO enrichments and depletions. The list of significant EDs is reduced parsimoniously using the criteria illustrated. The upper panel indicates the case where, if there are 2 clusters, 1 smaller and 1 larger that encompass it, and they both have an ED for the same GO term, either is de-selected for the ED if the other has a smaller *P*-value. The lower panel shows the scenario where a smaller cluster has a significant GO ED, but a larger cluster has a significant ED for a GO over-term relative to it. In this situation, the smaller cluster is de-selected for this GO ED.

for the landscape of GO EDs was sought. In doing this, enrichments are only compared with other enrichments, and depletions to other depletions.

First, for each lower ID-CBR cluster, lists of over-clusters were generated, where an *over-cluster* is any cluster containing a smaller cluster. These over-cluster lists were then progressively searched for larger clusters that yield a smaller significant *P*-value for a specific GO ED, and on finding such a cluster, the GO ED for the smaller cluster was de-selected, or vice versa if the smaller cluster *P*-value was smaller (this is illustrated schematically in Figure 2).

Second, for any specific GO ED, lists of GO over-terms in the same cluster or an over-cluster that have a smaller significant *P*-value were searched for, and on finding them, the initial GO ED under examination was also de-selected (Figure 2).

The GO associations were derived for *hpep* and D(R_B) tertiles in the same manner, except they were only processed with the latter criterion considering GO over-terms.

(A)

Parameter sets	% of ID-CBRs found by propensity criterion also found by overlap criterion	% of ID-CBRs found by overlap criterion also found by propensity criterion
len=20_cov=10% (10,20,1.8e-05)	84% (7223/8600)	86% (7223/8439)
len=50_cov=10% (23,33,4.1e-07)	90% (4695/5223)	91% (4695/5187)
len=100_cov=10% (38,48,7.3e-10)	91% (2656/2908)	94% (2656/2828)
len=250_cov=10% (68,78,4.1e-18)	92% (944/1029)	95% (944/991)

(B)

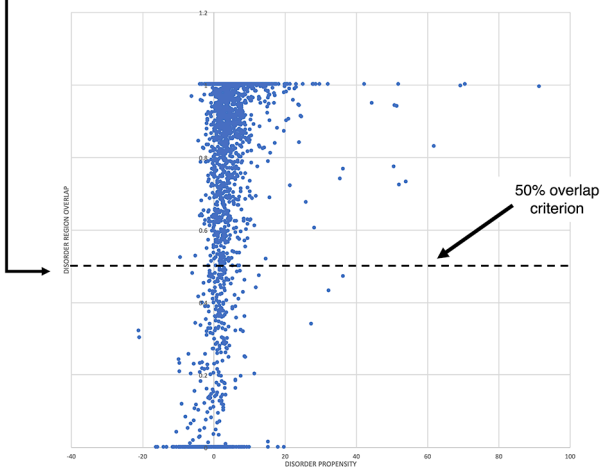

Figure 3. Convergence of the propensity and overlap criteria for ID-CBRs. (A) A tabulation of the percentages of ID-CBRs found by 1 criterion that are found by the other, for the parameter sets that yield 10% proteome coverage. (B) An example of a plot of ID-CBR disorder propensity vs disorder overlap for the len = 100_cov = 10% parameter set.

Results and Discussion

Definition of intrinsically disordered compositionally biased regions

Compositionally biased regions were labelled in the proteome of baker's yeast *S cerevisiae*, using the program fLPS.[15] Different parameters annotate regions of shorter or longer CBR target length or cover more of less of the yeast proteome (ie, the regions can be less or more biased). A diverse panel of parameter sets that target varying region lengths and proteome coverage were applied (Table 1). For example, for the fLPS parameters m = 12, M = 16, t = 5.4e−06, the median target length of CBRs is 20 and the estimated proteome coverage is ~5% (denoted 'len = 20_cov = 5%').

Intrinsically disordered compositionally biased regions were defined as regions that are compositionally biased and demonstrate a propensity for intrinsic disorder. For all the produced data sets, ID-CBRs were identified using 2 criteria, the overlap and propensity criteria as described in section 'Methods'. There is substantial convergence between these criteria for ID-CBRs. Figure 3 presents an analysis of this for the parameter sets for 10% coverage. Here, we can see that typically >90% of cases found by 1 criterion are also found by the other (Figure 3A). An example of a plot of percentage IDR overlap vs disorder for

Table 2. The percentage occurrence of ASTRAL protein domain matches and coiled-coil annotations in the ID-CBRs.

PARAMETER SET	ASTRAL PROTEIN DOMAIN MATCHES (%)		COILED-COIL ANNOTATIONS (%)	
	OF REGIONS	OF RESIDUES	OF REGIONS	OF RESIDUES
len=10_cov=5%	0.1	0.2	0.3	0.2
len=20_cov=5%	0.5	0.5	0.3	0.2
len=20_cov=10%	0.8	0.7	0.5	0.3
len=20_cov=25%	1.5	1.4	0.9	0.6
len=50_cov=5%	0.7	0.4	0.4	0.1
len=50_cov=10%	1.0	0.6	0.7	0.3
len=50_cov=25%	2.2	1.4	1.3	0.7
len=100_cov=5%	1.0	0.5	0.2	<0.1
len=100_cov=10%	1.4	0.7	1.2	0.5
len=100_cov=25%	3.4	1.7	2.3	0.8
len=250_cov=10%	3.0	0.7	2.0	0.4
len=250_cov=25%	7.6	2.2	4.4	0.8

Coiled-coil and ASTRALSCOP domain annotations as described in section 'Methods'.

the len=100_cov=10% parameter set illustrates this convergence (Figure 3B).

In general, there is little contribution to ID-CBRs from structured protein domains or coiled coils (Table 2). Only 0.2% to 2.2% of ID-CBR residues are covered with a structured protein domain match, and only 0.1% to 0.8% are labelled with a coiled coil by the DEEPCOIL algorithm.[23]

Intrinsically disordered compositionally biased regions prevalences

Some ID-CBR types are prevalent regardless of the parameter sets used (Figure 4). In particular, {S}-rich ID-CBRs are always ranked first, with the related signatures {SN}, {ST}, {SP}, and {SK} in the top 20 (Figure 4). {N}-rich and {Q}-rich are the next most prevalent. Such biases are linked to the prion phenomenon in budding yeast, and a large population of N-rich ID-CBRs accumulated during the evolution of the *Saccharomycetes* class.[31] These prevalences tally with the general trends across *Saccharomycetes* for homopeptide frequency.[6] However, some regions are only prevalent for low target length or high proteome coverage, eg, {K}-rich regions, and some only gain prominence at higher proteome coverage (ie, they are more mildly biased), such as the {Y}-rich, {F}-rich, and {G}-rich regions that move into view (Figure 4B). Also, multiple-residue biases become more numerous when longer ID-CBRs are probed (Figure 4C).

Blockiness and homopeptide content

The patterning of residues within an ID-CBR may also have functional importance. One type of patterning is blockiness, where residues of the same type clump together more along the sequence. A measure of blockiness R_B was derived as described in section 'Methods' (Figure 1). It was found to be most informative to plot R_B vs region length on a log-log plot and extract the deviation from the regression line D(R_B) as a relative measure of blockiness for a population of ID-CBRs with the same primary bias (Figure 1). Triads of sequences with roughly the same region length and low, intermediate, or high blockiness have been picked out for 2 example populations (Figure 5). In Figure 5A, one can see that sequences with lower relative blockiness can still have a lot of shorter homopeptides in them. In Figure 5B, the examples are the special case of regions that have bias for both E and K. For the lowest blockiness example, the Es and Ks are dispersed quite evenly along the sequence. Another measure of residue patterning studied here is the proportion of homopeptides (denoted *hpep*), where homopeptides are runs of $\geqslant 3$ residues. In general, D(R_B) only has a shallow correlation with *hpep* (R^2 values <0.05)

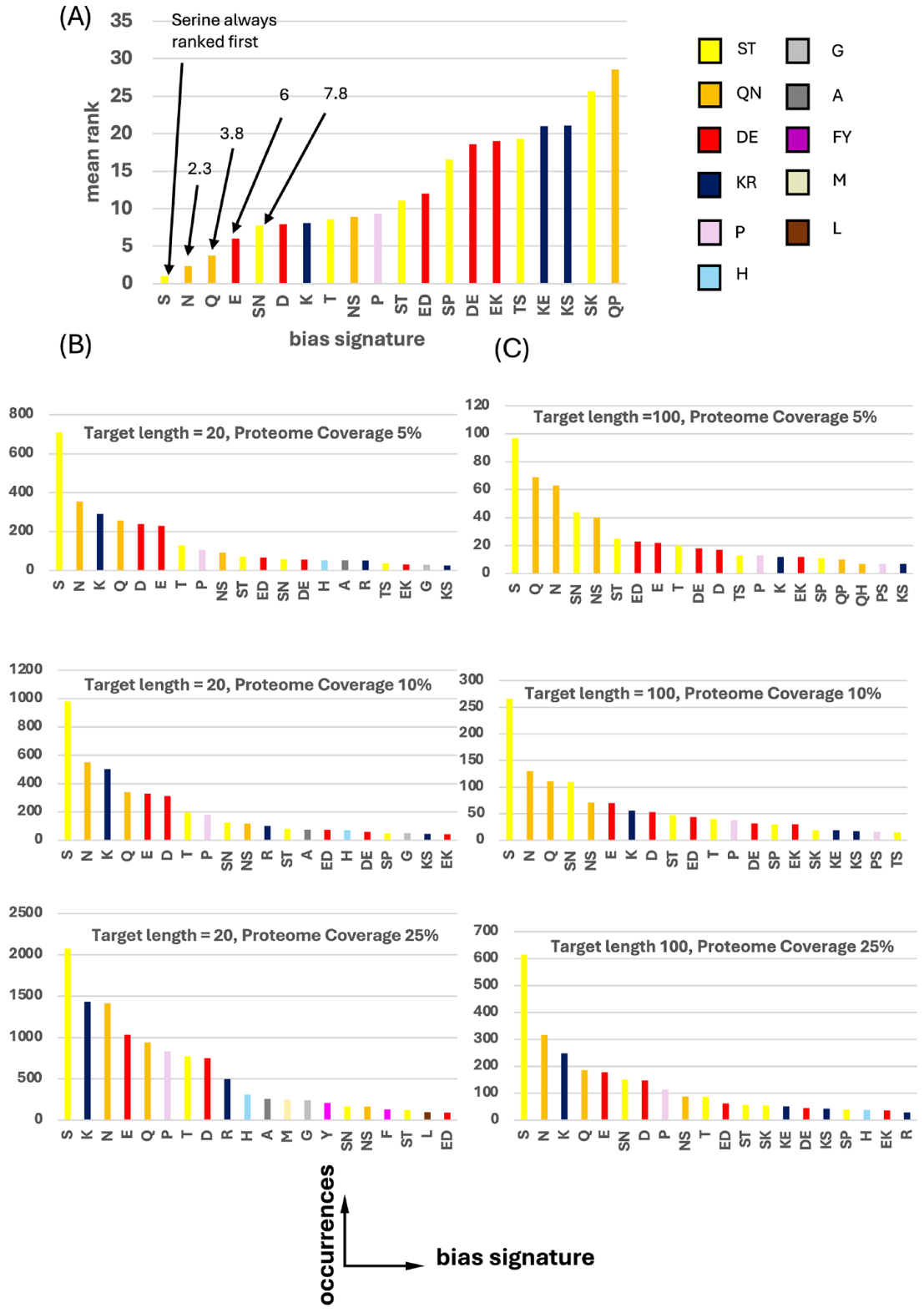

Figure 4. Prevalences of different bias signatures for ID-CBRs. (A) The prevalence of ID-CBR bias signatures across all 12 parameter sets was calculated using a mean rank, where the most prevalent is given rank 1, and so on. The colour key for this whole figure is to the right, and it is the same as used in the example tree figure (Supplemental Figure 2). (B) The parameter sets for a low target length = 20 residues, but with increasing proteome coverage down the page. (C) Similarly to (B), but for a higher target length of 100 residues.

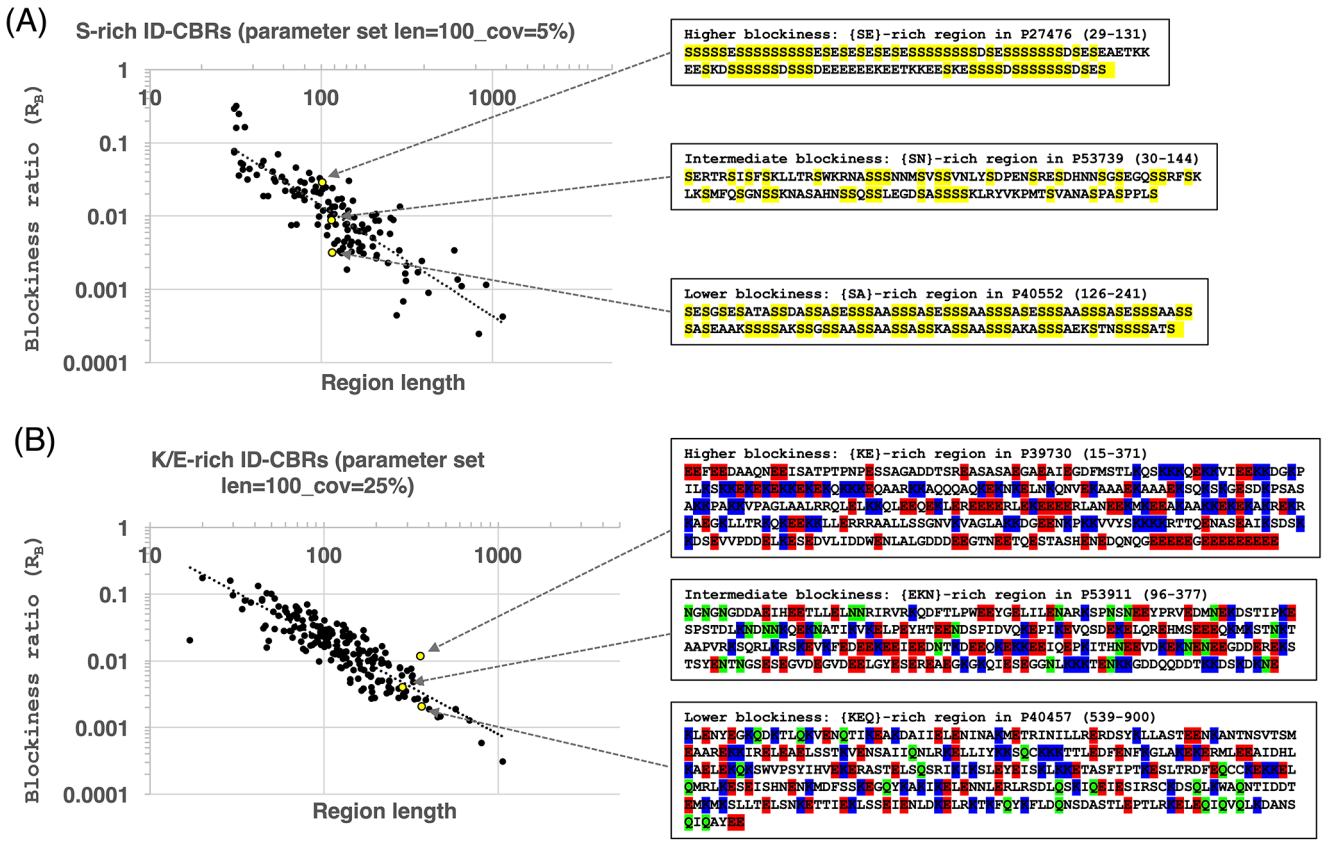

Figure 5. Examples of the blockiness calculation. (A) Log10(R_B) *vs* Log10(*region length*) for ID-CBRs for the len = 100_cov = 5% parameter set and a primary bias of S (serine). Examples of higher, intermediate, and lower blockiness are picked out. (B) As in (A), except for the special category of K/E-rich (ID-CBRs that are rich in both lysine and glutamate), and the len = 100_cov = 25% parameter set (ie, for detection of longer ID-CBRs with more mild bias).

(examples for serine-rich ID-CBRs are displayed in Supplemental Figure 1).

Intrinsically disordered compositionally biased region clustering

The ID-CBR data sets arising for any one parameter set were clustered using compositional-distance matrices fed into the NJ algorithm.[28] An example of this clustering is displayed in Supplemental Figure 2, for annotations with target length 100 residues and coverage 5% (ie, highly biased). Here, ID-CBRs coloured according to their primary biasing residues can be observed coalescing into larger clusters. Of particular note are the following: a large cluster of primarily Q-rich or N-rich regions, almost all of which are also labelled as prion-like in composition, or prion-forming; populations of SN-rich and ST-rich regions, and a charged group that bifurcates between mostly E-rich and most K-rich.

To representatively probe the detail of these clusterings, the positioning of the Sup35 protein was examined (UniProt accession P05453, Figure 6). The Sup35 protein functions in translation termination and stop codon recognition and contains a {QYNG}-rich prion-forming domain that underlies the [PSI+] prion phenomenon.[32,33] It also contains a {KE}-rich central M domain that mediates pH sensing during reversible condensate formation in response to stress.[34,35] Rnq1 (UniProt P25367) is a prion protein that underlies the [RNQ+] prion phenomenon, which is required in yeast cells for the natural occurrence of the [PSI+] prion.[36-38] In Figure 6 (lower left), one can see that a region containing its prion-forming domain clusters pairwise with part of the Sup35 prion-forming domain, indicating that its [PSI+] seeding function may be in part compositionally engendered. Two other prion-forming domains cluster close by, one in Ngr1, a negative-growth regulatory protein, and the other in Pgd1, a mediator of RNA pol II transcription subunit 3. Indeed, for parameter sets targeting 5% coverage (ie, for detecting highly biased regions), this close clustering of Rnq1 and Sup35 is the general result (median cluster size = 3, Table 3). Across the upper half of the figure, we

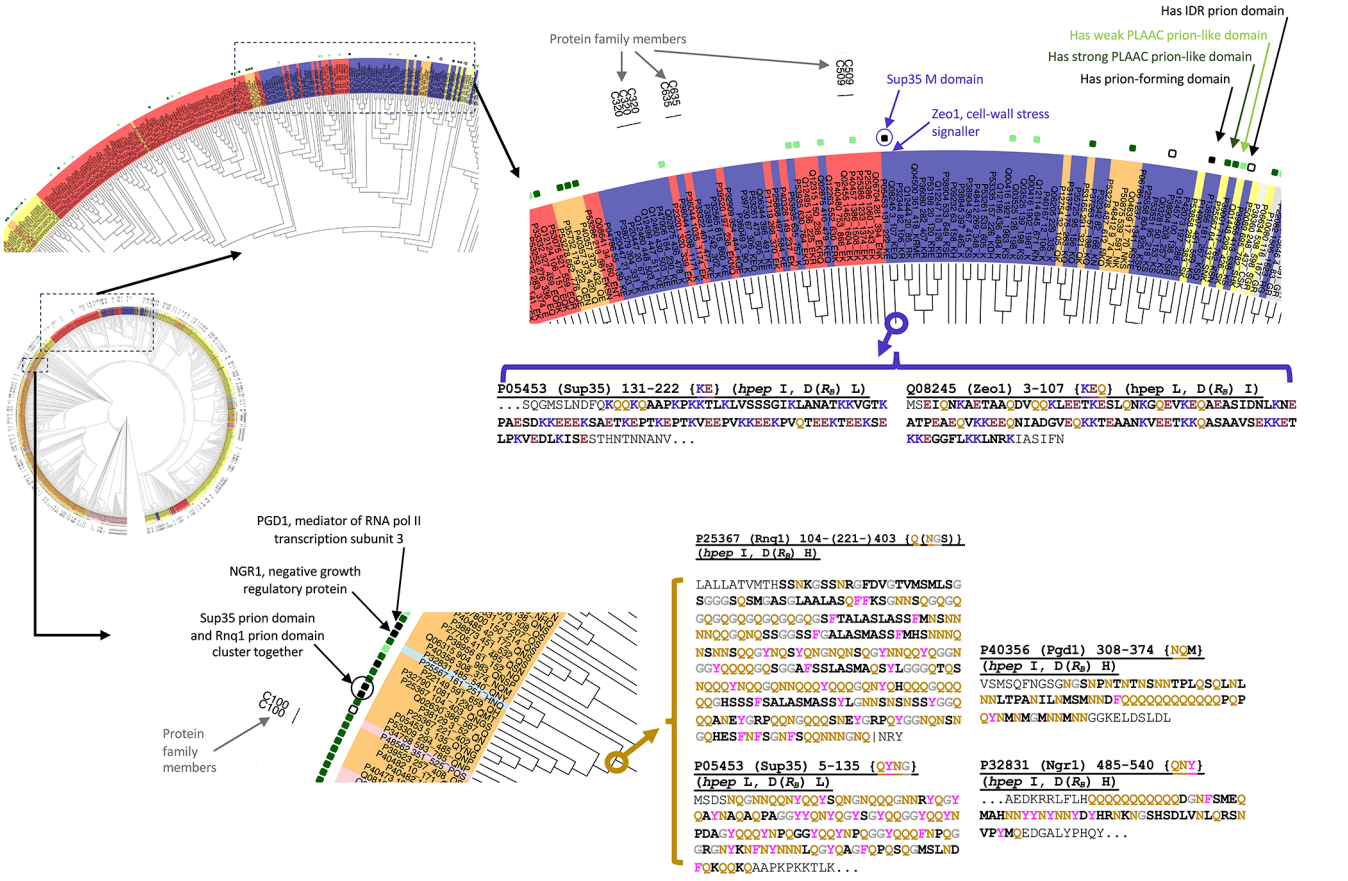

Figure 6. ID-CBR clusters involving the Sup35 prion protein. The illustrative clustering supplied in Supplemental Figure 2 is examined with zoom-ins to the locales of the Sup35 {KE}-rich M domain and the {QYNG}-rich prion-forming domain. The colour-coding is described in the Supplemental Figure legend. The tree is made with parameters for target length = 100 residues and proteome coverage of ~5%. ID-CBRs are named: *UniProtAccession_Start_End_BiasSignature*. A large cluster of predominantly E-biased or K-biased regions is zoomed into (upper left), then a further zoom-in reveals the locale of the Sup35 M domain. The ID-CBR sequences are displayed below this, coloured in concert. Also, a zoom-in to a cluster containing a Sup35 prion-forming ID-CBR is presented along with the sequences of the Sup35 ID-CBR and 3 other prion-forming ID-CBRs within the cluster. Labels that indicate the presence of prion or prion-like domains in the respective protein sequences are also arrayed in an outer rim (see Supplemental Figure legend for details). Numeric protein-family indices are displayed where relevant. The depicted sequences are also labelled with their bias signatures and their D(R_B) and *hpep* tertiles.

Table 3. Smallest clusters in cluster trees containing both Sup35 and Rnq1.

PARAMETER SET	SMALLEST CLUSTER SIZE
len = 10_cov = 5%	**2**
len = 20_cov = 5%	**12**
len = 20_cov = 10%	3
len = 20_cov = 25%	357
len = 50_cov = 5%	**4**
len = 50_cov = 10%	1525
len = 50_cov = 25%	14
len = 100_cov = 5%	**2**
len = 100_cov = 10%	36
len = 100_cov = 25%	1361
len = 250_cov = 10%	147
len = 250_cov = 25%	4

(Continued)

Table 3. (Continued)

The results for 5% coverage (ie, highly biased) are in bold.

progressively zoom into the locale of the Sup35 M domain in the clustering tree, where we find it together with other {KE}/{EK}-rich ID-CBRs, and paired most closely with a {KE}-rich region in Zeo1, which is an antagonist in signalling cell-wall stress to the PKC1-MPK1 cell integrity pathway (Figure 6).

Table 4. Most prominent hypothetical GO enrichments for compositional-distance clusters containing the Sup35 M domain or prion-forming domain ID-CBRs.

CONTAINING SUP35 M DOMAIN ID-CBRS[A]						
CONSENSUS BIAS	NO. OF PARAMETER SETS	CLUSTER SIZE RANGE	CASES WITH GO TERM	*P*-VALUE RANGE	GO TERM	DESCRIPTION
E / K	11	135-1785	10-32	1e–07-1e–17	GO:0030687	Preribosome, large subunit precursor
E / K	10	107-1782	10-32	2e–06-4e–12	GO:0042273	Ribosomal large subunit biogenesis
E / K	10	157-1247	29-84	1e–11-9e–26	GO:0042254	Ribosome biogenesis
E / K	10	166-1900	15-84	4e–06-7e–21	GO:0016887	ATP hydrolysis activity
E / K	10	162-1802	43-162	9e–20-2e–38	GO:0005730	Nucleolus
. . . . .						
E / K	**4***	**586-1095**	**82-143**	**8e–06-1e–08**	**GO:0000166**	**Nucleotide binding**
CONTAINING SUP35 PRION-FORMING DOMAIN ID-CBRS[B]						
CONSENSUS BIAS	NO. OF PARAMETER SETS	CLUSTER SIZE RANGE	CASES WITH GO TERM	*P*-VALUE RANGE	GO TERM	DESCRIPTION
Q / N	**12**	**108-984**	**66-447**	**4e–06-4e–12**	**GO:0005737**	**Cytoplasm**
Q / N	11	120-1713	33-125	3e–15-6e–32	GO:0045944	Positive regulation of RNA pol II transcn.
Q / N	**9**	**11-528**	**5-29**	**9e–06-6e–12**	**GO:0010494**	**Cytoplasmic stress granule**
Q / N	9	113-1467	79-697	5e–11-4e–30	GO:0005634	Nucleus
Q / N	8	234-1177	24-59	4e–11-2e–18	GO:0043565	Sequence-specific DNA binding
. . . . .						
N / Q	**4****	**207-1145**	**32-76**	**6e-08-9e-24**	**GO:0003729**	**mRNA binding**

The terms associated with Sup35 are in bold.
[a]Further categories in the top 20 associated with nucleolar function and compartments include (no. parameter sets in brackets): GO:0006364 rRNA processing (9); GO:0000466 maturation of 5.8S rRNA from tricistronic rRNA transcript (8); GO:0000463 maturation of LSU-rRNA from tricistronic rRNA transcript (8); GO:0030686 90S preribosome (7); GO:0000462 maturation of SSU-rRNA from tricistronic rRNA transcript (5); GO:0000480 endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (4); GO:0000472 endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA (4).
[b]Further categories in the top 20 associated with transcription include (# parameter sets in brackets): GO:0006357 regulation of transcription by RNA polymerase II (8); GO:0003677 DNA binding (7); GO:0001228 DNA-binding transcription activator activity, RNA polymerase II-specific (7); GO:0000981 DNA-binding transcription-factor activity, RNA polymerase II-specific (7); and GO:0000122 negative regulation of transcription by RNA polymerase II (6).
*Parameter sets (len=20_cov=25%, len=20_cov=10%, len=50_cov=10%, len=50_cov=25%).
**Parameter sets (len=10_cov=5%, len=100_cov=10%, len=20_cov=10%, len=20_cov=25%).

Hypothetical Gene Ontology functional cluster associations

The informativeness of this compositional-distance clustering was further probed by generating sets of functional hypotheses using GO.[30] Significant GO associations were filtered for redundancy and for protein-family sequence homology as described in section 'Methods'. Hundreds of GO EDs are detected, with depletions being about ~20% the amounts of enrichments (Table 1). Each calculated GO ED is a hypothesis about the functional importance of clusters. The number of parameter sets that a GO ED occurs in is used as an indicator of parameter independence. Also, more regions that are more distant compositionally are pulled in the larger the cluster is; the more parameter sets that a cluster is found by, more, mildly biased regions are pulled into the hypothesis. Furthermore, significant ED may occur for just, say, low proteome coverage, or long target length parameter sets.

As an illustrative example, the significant GO EDs that occur in most parameter sets for the Sup35 protein have been arrayed in Table 4, for both its {KE}-rich M domain and its {QYNG}-rich prion-forming domain. The most striking trend for the E/K-rich clusters is the predominant association with

Figure 7. Three examples of ID-CBRs clustered by compositional distance. Along the top is the key for the 1-row tables at the top of each figure panel. The *consensus bias* is the predominant bias for the cluster for any parameter set. Also tabulated are the number of parameter sets that designate this cluster, the cluster size range across parameter sets, and the predominant tertiles for *hpep* and D(R_B) for the cluster, ie, H for high, I for intermediate, and L for low (with the proportion in these tertiles in brackets). Gene Ontology associations significant at < Bonferroni-corrected P-value threshold of 1.1e−05 are listed, with the number of cases for each association in brackets. The identity of these cases is colour-coded with dots. The longest ID-CBR sequences for each case are shown.

the nucleolus, and with other categories linked to nucleolar functions and compartments, although these are not specifically annotated for Sup35. Itself, it is part of a more general association of E/K-rich ID-CBRs with nucleotide binding. The Sup35 Q-rich ID-CBR is part of a clear association with cytoplasmic stress granules along with up to 28 other Q/N-rich cases, an association which does not occur for the E/K-rich regions, which is interesting considering the role of the M domain in phase separation.[34,35]

Three specific examples of compositional-distance clusters with clear hypothetical functional associations were picked to examine in further detail (Figure 7). An {AQ}-rich region with low blockiness is linked to transcription coactivator activity and to a general category of transcriptional regulation (Figure 7A), thus providing functional hypotheses for other clustered but un-annotated cases. In Cyc8, length variation of this region has been shown to cause concerted upregulation or downregulation of expression in >150 genes, and transcription-factor binding site analysis further suggests that this is due to Cyc8's role as a coactivator.[39] The {AQ}-rich ID-CBR in GTS1 was shown to be part of a tract that suppresses polyglutamine toxicity of other proteins,[40] implying a dual function for such tracts. The second example (Figure 7B) is an {AK}/{KA}-rich region that overlaps IDRs in proteins associated with tRNA aminoacylation. Arc1p is a protein that binds tRNA and forms the AME complex with methionyl-tRNA and glutamyl-tRNA synthetases and functions in tRNA delivery.[41,42] A long {KA}-rich region in Arc1p is part of a central IDR region whose character (ie, lysine content) indicates it may feature in RNA binding.[43] The third example is a set of {SE}-rich ID-CBRs, 4 of which are linked by GO to regulation of protein localization (Figure 7C), such as SPA2, which is a polarisome subunit linked to actin cytoskeletal organization, establishment of cell polarity, apical bud growth, and regulation of mating projection growth initiation and termination.[44,45] The function of this region has yet to be discerned. Of the other cases, BNI1 is also

a polarisome component, and ZDS1 and ZDS2 are involved in establishment of cell polarity (Figure 7C).

Larger clusters of regions containing hundreds of cases may also be functionally informative. A rather large cluster of 138 R-rich regions significantly associated with nuclear localization was noted at 25% proteome coverage (ie, they are often mildly biased) and shorter target lengths (Supplemental File 2). To check whether these regions arise from possible NLSs (which can have some R bias) or from DNA-binding (as arginine features in DNA-binding mechanisms[46]), they were cross-referenced with the 104 known NLSs in the NLSdb database.[25] Only 11 of the 104 NLSs associate with an R-rich tract, with a further 30 associated with a K-rich tract and 26 being embedded in regions with a different compositional bias (eg, the short NLSs in large ribosomal subunit protein uL15 P02406 are encompassed by an {HG}-rich ID-CBR). Furthermore, when the larger number of R-rich tracts that are also associated with DNA/nucleic-acid binding is accounted for, the nuclear association for these R-rich tracts is no longer significant, indicating that it likely comes primarily from features for such binding. Similarly, for a much larger cluster of >900 cases of short K-rich regions found at 25% coverage, the significant association with nuclear localization disappears when the larger number of cases associated with DNA/nucleic-acid binding is accounted for.

Tertiles of blockiness and homopeptide content and their functional hypotheses

The ID-CBRs with the same primary bias were portioned into tertiles for both *hpep* and D(R_B) values (*H*, high; *I*, intermediate; *L*, low). Functional hypotheses for these tertiles were then examined. First for these tertiles, I continued the theme of examining E/K and Q/N biases that were first picked out in relation to the Sup35p example (Table 5). Notable E/K-rich examples in this table are high *hpep* and D(R_B) ID-CBRs associated with the *nucleolus* (examples are depicted in Supplemental Figure 3C), and low D(R_B) ones linked to the *spindle pole body*. In a recent review, such nucleolar regions are specifically described as associated with the fibrillar centre and dense fibrillar component within nucleoli.[47] There are 2 distinct types of Q/N-rich ID-CBRs; those associated with *transcription coactivator activity* have high *hpep* and high D(R_B), whereas those linked to *kinase activity* tend towards intermediate *hpep* and low D(R_B) (Table 5, examples are depicted in Supplemental Figure 3A and B). Indeed, experimental work on modifying the length of Q/N homopeptide tracts in transcription factors and coactivators has indicated that they can tune or modulate their proteins' regulatory roles.[39]

The most prominent functional hypotheses across all tertiles and parameter sets are arrayed in Supplemental Table 2. One can see that in general, low blockiness and high *hpep* content are favoured among these GO EDs, although these lists overlap rarely (just 3 cases), and 26 cases show both a *hpep* trend and a blockiness trend. For example, a population of up to 36S-rich ID-CBRs with low *hpep* and intermediate blockiness is linked to *DNA-binding transcription-factor activity*, including a tract of SFL1, an activator involved in control of flocculation, stress response and pseudohyphal growth,[48] and VHL1, which is required for induction of vitamin H and biotin-intermediate transporters[49] (Figure 8A). A group of up to 18 cases of D-rich ID-CBRs feature a *chromatin remodelling* linkage (Figure 8B), including the transcription-factor Autonomously-Replicating Sequence binding factor 1, and RAD26, which is involved in transcription-coupled repair of nucleotide excisions. In general, D-rich regions may facilitate and accelerate DNA binding, especially at cytosine-rich sites.[50,51]

Data

In addition to the raw data for ID-CBRs used to define and analyse them in this article, further characteristics have been added to the Supplementary Files, such as the normalized Kyte-Doolittle Hydropathy and 'middleness' scales that were generated for a previous study,[13,52] and proportions of groups of residues based on the Taylor Venn diagram of amino acids (eg, 'tiny polar, 'charged', 'aromatic')[53] (Supplemental File 3). This information can be cross-referenced with the lists of ID-CBR cluster membership (Supplemental File 4) and of GO E/Ds (Supplemental File 2).

Conclusions

Here, intrinsically disordered compositionally biased regions (ID-CBRs) were formally defined, using 2 different criteria that were largely congruent. It was discovered that ID-CBRs can be formed into clusters using compositional-distance and residue patterning (ie, 'blockiness' and homopeptide content), and these clusters can have significant functional associations. In doing so, redundancies were taken account of, such as from the sequence homology of protein-family members, and from similar GO functional terms. The function of CBRs has been generally un-appreciated or under-appreciated in cell biological experiments, constructs, and hypotheses, so these data may be helpful in that regard. The chief advantage of the approach described here is that it mitigates against any parameter dependence in annotating and analysing compositional biases. However, this also means that in many cases, the relevant region boundaries can be ambiguous, and there are multiple solutions to defining the CBRs, so that several protein constructs may be necessary to investigate the relative functional importance of CBR features in cell biology experiments. The functional annotations are also restricted by the detail and structure of the GO digraph created by the GO curators. Nevertheless, they can be used as hypotheses for other proteins in the same cluster or tertile; also, for a given association, the CB-IDRs may perform a more specific functional role that has yet to be elucidated.

Table 5. Gene Ontology enrichments for tertiles of blockiness and homopeptide content for E/K-rich regions and Q-or-N-rich regions.

E/K-RICH								
H			I			L		
TERTILE COUNT[A]	REGION COUNTS[B]	CATEGORY[C]	TERTILE COUNT[A]	REGION COUNTS[B]	CATEGORY[C]	TERTILE COUNT[A]	REGION COUNT[B]	CATEGORY[C]
Blockiness (R_B)								
4(3H,1I)	9-15	***GO:0005730 nucleolus***	-	-	-	3	3	*GO:0008186 ATP-dept. activity, acting on RNA*
Homopeptide content (*hpep*)								
6(4H,2I)	7-21	***GO:0005730 nucleolus***	4	10-14	*GO:0006364 rRNA processing*	2	33-60	*GO:0005816 spindle pole body*
2	3	*GO:0036437 Isw1b complex*	-	-	-	-	-	-
2	11-14	*GO:0016887 ATP hydrolysis activity*	-	-	-	-	-	-
2	28-50	*GO:0005634 nucleus*	-	-	-	-	-	-
Q-OR-N-RICH REGIONS								
H			I			L		
TERTILE COUNT	REGION COUNTS	CATEGORY	TERTILE COUNT	REGION COUNTS	CATEGORY	TERTILE COUNT	REGION COUNTS	CATEGORY
Blockiness (R_B)								
4	17-29	*GO:0003723 RNA binding*	2	8	*GO:0018105 peptidyl-serine phosphorylation*	5(4L,1I)	28-101	*GO:0005886 plasma membrane*
3	7-19	***GO:0003713 transcription coactivator activity***	-	-	-	4	15-33	*GO:0005935 cellular bud neck*
3(2H,1I)	9-17	*GO:0004674 protein Ser/Thr kinase activity*	-	-	-	3	5-6	*GO:0030276 clathrin binding*
-	-	-	-	-	-	3	7-10	*GO:0017148 negative regulation of translation*
-	-	-	-	-	-	3(2L,1I)	25-60	*GO:0016310 phosphorylation*
-	-	-	-	-	-	3(2L,1I)	25-60	***GO:0016301 kinase activity***
Homopeptide content (*hpep*)								
7(6H,1I)	11-28	*GO:0010494 cytoplasmic granule*	5(3I,2L)	8-20	*GO:0003700 DNA-binding TF activity*	2	8-10	*GO:0031929 TOR signalling*
6(5H,1I)	10-24	***GO:0003713 transcription coactivator activity***	5(3I,2L)	4-7	*GO:0044613 nuclear pore central transport channel*	-	-	-
5	7-23	*GO:0061629 RNA pol II specific DNA-binding TF binding*	4(3I,1L)	25-53	***GO:0016301 kinase activity***	-	-	-
4	5-7	*GO:0000288 nuclear-transcribed mRNA catabolism, deadenylation-dept. decay*	4(3I,1L)	53-85	*GO:0008270 zinc-ion binding*	-	-	-
4	84-448	*GO:0005737 cytoplasm*	-	-	-	-	-	-

Enrichments that occur for multiple parameter sets are listed up to a maximum of the top 5, including ties.
[a]The count of tertiles, ie, high (H), intermediate (I), or low (L) for *hpep* and for D(R_B).
[b]The range of region counts for different parameter sets. The upper bounds arise when more mildly biased regions are included (through application of parameter sets with higher proteome coverage).
[c]Enrichments that occur for both a *hpep* and D(R_B) tertiles are in bold.

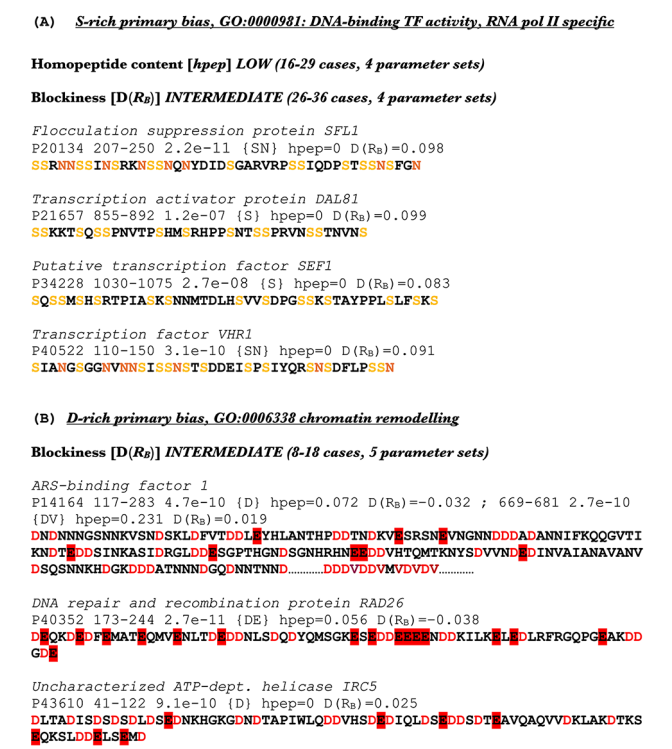

Figure 8. Examples of significant Gene Ontology associations for tertiles of homopeptide content (*hpep*) and blockiness [D(R_B)]. Residues are colour-coded as in other figures. Gene Ontology associations significant at < Bonferroni-corrected *P*-value threshold of 1.1e−05 are listed, with the number of cases for each association in brackets. Each ID-CBR is colour-coded with the coding applied also in previous figures. Listed for each ID-CBR are: the UniProt accession, the range (start to end), the binomial bias *P*-value, the bias signature in curly brackets, and the *hpep* and D(R_B) values. The longest regions for each protein for each tendency are picked to display. (A) RNA polymerase II-specific transcription-factor (TF) activity is a significant association for serine-rich ID-CBRs with *low* homopeptide content and *intermediate* blockiness. (B) Chromatin remodelling is a significant functional hypothesis for aspartate-rich ID-CBRs with *intermediate* blockiness. Glutamate residues are also pointed out with red highlighting.

Acknowledgements

The author thanks Wan-Chun Su for running the DeepCoil program.

Author Contributions

PMH conceived the project, performed the data analysis and wrote the paper.

Data Availability Statement

The source data for the work is available from public sequence databases as described in section 'Methods'. Generated cluster and annotation data are available in several Supplementary Files.

ORCID iD

Paul M. Harrison https://orcid.org/0000-0002-7477-1014

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 1999;293:321-331. doi:10.1006/jmbi.1999.3110
2. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114:6589-6631. doi:10.1021/cr400525m
3. Narasumani M, Harrison PM. Discerning evolutionary trends in post-translational modification and the effect of intrinsic disorder: analysis of methylation, acetylation and ubiquitination sites in human proteins. *PLoS Comput Biol*. 2018;14:e1006349. doi:10.1371/journal.pcbi.1006349
4. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins*. 2001;42:38-48. doi:10.1002/1097-0134(20010101)42:1<38::aid-prot50>3.0.co;2-3
5. Zhao B, Kurgan L. Compositional bias of intrinsically disordered proteins and regions and their predictions. *Biomolecules*. 2022;12:888. doi:10.3390/biom12070888
6. Wang Y, Harrison PM. Homopeptide and homocodon levels across fungi are coupled to GC/AT-bias and intrinsic disorder, with unique behaviours for some amino acids. *Sci Rep*. 2021;11:10025. doi:10.1038/s41598-021-89650-1
7. Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*. 2003;25:847-855. doi:10.1002/bies.10324
8. Kastano K, Mier P, Dosztanyi Z, Promponas VJ, Andrade-Navarro MA. Functional tuning of intrinsically disordered regions in human proteins by composition bias. *Biomolecules*. 2022;12:1486. doi:10.3390/biom12101486
9. Cascarina SM, King DC, Osborne Nishimura E, Ross ED. LCD-Composer: an intuitive, composition-centric method enabling the identification and detailed functional mapping of low-complexity domains. *NAR Genom Bioinform*. 2021;3:lqab048. doi:10.1093/nargab/lqab048
10. Mier P, Andrade-Navarro MA. Assessing the low complexity of protein sequences via the low complexity triangle. *PLoS ONE*. 2020;15:e0239154. doi:10.1371/journal.pone.0239154
11. UniProt C. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523-D531. doi:10.1093/nar/gkac1052
12. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.
13. Harrison PM. Compositionally biased dark matter in the protein universe. *Proteomics*. 2018;18:e1800069. doi:10.1002/pmic.201800069
14. Harrison PM. FLPS: fast discovery of compositional biases for the protein universe. *BMC Bioinformatics*. 2017;18:476. doi:10.1186/s12859-017-1906-3
15. Harrison PM. fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences. *PeerJ*. 2021;9:e12363. doi:10.7717/peerj.12363
16. Harrison PM. Optimizing strategy for the discovery of compositionally-biased or low-complexity regions in proteins. *Sci Rep*. 2024;14:680. doi:10.1038/s41598-023-50991-8
17. Piovesan D, Del Conte A, Clementel D, et al. MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Res*. 2023;51:D438-D444. doi:10.1093/nar/gkac1065
18. Su TY, Harrison PM. Conservation of prion-like composition and sequence in prion-formers and prion-like proteins of Saccharomyces cerevisiae. *Front Mol Biosci*. 2019;6:54. doi:10.3389/fmolb.2019.00054
19. Alberti S, Halfmann R, King O, Kapila A, Lindquist S. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell*. 2009;137:146-158.
20. Chakrabortee S, Byers JS, Jones S, et al. Intrinsically disordered proteins drive emergence and inheritance of biological traits. *Cell*. 2016;167:369-381.e12. doi:10.1016/j.cell.2016.09.017
21. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*. 2014;30:2501-2502. doi:10.1093/bioinformatics/btu310
22. Su WC, Harrison PM. Deep conservation of prion-like composition in the eukaryotic prion-former Pub1/Tia1 family and its relatives. *PeerJ*. 2020;8:e9023. doi:10.7717/peerj.9023
23. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*. 2019;35:2790-2795. doi:10.1093/bioinformatics/bty1062
24. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*. 2014;42:D304-D309. doi:10.1093/nar/gkt1240

25. Bernhofer M, Goldberg T, Wolf S, et al. NLSdb-major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Res*. 2018;46:D503-D508. doi:10.1093/nar/gkx1021
26. Quaglia F, Meszaros B, Salladini E, et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res*. 2022;50:D480-D487. doi:10.1093/nar/gkab1082
27. Felsenstein J. PHYLIP-phylogeny inference package. *Cladistics*. 1989;5: 164-166.
28. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406-425. doi:10.1093/oxfordjournals.molbev.a040454
29. Subramanian B, Gao S, Lercher MJ, Hu S, Chen WH. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res*. 2019;47:W270-W275. doi:10.1093/nar/gkz357
30. Gene Ontology Consortium; Aleksander SA, Balhoff J, et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224:iyad031. doi:10.1093/genetics/iyad031
31. An L, Fitzpatrick D, Harrison PM. Emergence and evolution of yeast prion and prion-like proteins. *BMC Evol Biol*. 2016;16:24. doi:10.1186/s12862-016-0594-3
32. Shorter J, Lindquist S. Prions as adaptive conduits of memory and inheritance. *Nat Rev Genet*. 2005;6:435-450. doi:10.1038/nrg1616
33. Serio TR, Lindquist SL. [PSI+]: an epigenetic modulator of translation termination efficiency. *Annu Rev Cell Dev Biol*. 1999;15:661-703. doi:10.1146/annurev.cellbio.15.1.661
34. Franzmann TM, Alberti S. Protein phase separation as a stress survival strategy. *Cold Spring Harb Perspect Biol*. 2019;11:a034058. doi:10.1101/cshperspect.a034058
35. Franzmann TM, Jahnel M, Pozniakovsky A, et al. Phase separation of a yeast prion protein promotes cellular fitness. *Science*. 2018;359:eaao5654. doi:10.1126/science.aao5654
36. Vitrenko YA, Pavon ME, Stone SI, Liebman SW. Propagation of the [PIN+] prion by fragments of Rnq1 fused to GFP. *Curr Genet*. 2007;51:309-319. doi:10.1007/s00294-007-0127-0
37. Halfmann R, Jarosz DF, Jones SK, Chang A, Lancaster AK, Lindquist S. Prions are a common mechanism for phenotypic inheritance in wild yeasts. *Nature*. 2012;482:363-368. doi:10.1038/nature10875
38. Sondheimer N, Lopez N, Craig EA, Lindquist S. The role of Sis1 in the maintenance of the [RNQ+] prion. *EMBO J*. 2001;20:2435-2442. doi:10.1093/emboj/20.10.2435
39. Gemayel R, Chavali S, Pougach K, et al. Variable glutamine-rich repeats modulate transcription factor activity. *Mol Cell*. 2015;59:615-627. doi:10.1016/j.molcel.2015.07.003
40. Ripaud L, Chumakova V, Antonin M, et al. Overexpression of Q-rich prion-like proteins suppresses polyQ cytotoxicity and alters the polyQ interactome. *Proc Natl Acad Sci U S A*. 2014;111:18219-18224. doi:10.1073/pnas.1421313111
41. Deinert K, Fasiolo F, Hurt EC, Simos G. Arc1p organizes the yeast aminoacyl-tRNA synthetase complex and stabilizes its interaction with the cognate tRNAs. *J Biol Chem*. 2001;276:6000-6008. doi:10.1074/jbc.M008682200
42. Galani K, Grosshans H, Deinert K, Hurt EC, Simos G. The intracellular location of two aminoacyl-tRNA synthetases depends on complex formation with Arc1p. *EMBO J*. 2001;20:6889-6898. doi:10.1093/emboj/20.23.6889
43. Ukmar-Godec T, Hutten S, Grieshop MP, et al. Lysine/RNA-interactions drive and regulate biomolecular condensation. *Nat Commun*. 2019;10:2909. doi:10.1038/s41467-019-10792-y
44. Sheu YJ, Santos B, Fortin N, Costigan C, Snyder M. Spa2p interacts with cell polarity proteins and signaling components involved in yeast cell morphogenesis. *Mol Cell Biol*. 1998;18:4053-4069. doi:10.1128/MCB.18.7.4053
45. Yorihuzi T, Ohsumi Y. Saccharomyces cerevisiae MATa mutant cells defective in pointed projection formation in response to alpha-factor at high concentrations. *Yeast*. 1994;10:579-594. doi:10.1002/yea.320100503
46. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. *Nature*. 2009;461:1248-1253. doi:10.1038/nature08473
47. King MR, Ruff KM, Pappu RV. Emergent microenvironments of nucleoli. *Nucleus*. 2024;15:2319957. doi:10.1080/19491034.2024.2319957
48. Song Q, Johnson C, Wilson TE, Kumar A. Pooled segregant sequencing reveals genetic determinants of yeast pseudohyphal growth. *PLoS Genet*. 2014;10:e1004570. doi:10.1371/journal.pgen.1004570
49. Weider M, Machnik A, Klebl F, Sauer N. Vhr1p, a new transcription factor from budding yeast, regulates biotin-dependent expression of VHT1 and BIO5. *J Biol Chem*. 2006;281:13513-13524. doi:10.1074/jbc.M512158200
50. Hossain KA, Kogut M, Slabonska J, Sappati S, Wieczor M, Czub J. How acidic amino acid residues facilitate DNA target site selection. *Proc Natl Acad Sci U S A*. 2023;120:e2212501120. doi:10.1073/pnas.2212501120
51. Wang X, Bigman LS, Greenblatt HM, Yu B, Levy Y, Iwahara J. Negatively charged, intrinsically disordered regions can accelerate target search by DNA-binding proteins. *Nucleic Acids Res*. 2023;51:4701-4712. doi:10.1093/nar/gkad045
52. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157:105-132. doi:10.1016/0022-2836(82)90515-0
53. Taylor WR. The classification of amino acid conservation. *J Theor Biol*. 1986;119:205-218. doi:10.1016/s0022-5193(86)80075-3