Protein intrinsically disordered regions have a non-random, modular architecture.

Brendan S. McConnell[1], Matthew W. Parker[1]*

[1]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA.

*Correspondence should be addressed to: matthew.parker@utsouthwestern.edu

Short title: Protein disordered regions are non-randomly organized.

ABSTRACT

Motivation

Protein sequences can be broadly categorized into two classes: those which adopt stable secondary structure and fold into a domain (i.e., globular proteins), and those that do not. The sequences belonging to this latter class are conformationally heterogeneous and are described as being intrinsically disordered. Decades of investigation into the structure and function of globular proteins has resulted in a suite of computational tools that enable their sub-classification by domain type, an approach that has revolutionized how we understand and predict protein functionality. Conversely, it is unknown if sequences of disordered protein regions are subject to broadly generalizable organizational principles that would enable their sub-classification.

Results

Here we report the development of a statistical approach that quantifies linear variance in amino acid composition across a sequence. With multiple examples we provide evidence that intrinsically disordered regions are organized into statistically non-random modules of unique compositional bias. Modularity is observed for both low and high complexity sequences and, in some cases, we find that modules are organized in repetitive patterns. These data demonstrate that disordered sequences are non-randomly organized into modular architectures and motivate future experiments to comprehensively classify module types and to determine the degree to which modules constitute functionally separable units analogous to the domains of globular proteins.

Availability and implementation

The source code, documentation, and data to reproduce all figures is freely available at https://github.com/MWPlabUTSW/Chi-Score-Analysis.git. The analysis is also available as a Google Colab Notebook (https://colab.research.google.com/github/MWPlabUTSW/Chi-Score-Analysis/blob/main/ChiScore_Analysis.ipynb).

1 INTRODUCTION

The basic functional unit of a globular protein is the domain, a polypeptide region that folds into a stable three-dimensional structure (1). Many eukaryotic proteins possess a modular, multi-domain architecture resulting from the genetic duplication and shuffling (2) of the approximately 6,000 protein domain superfamilies (3). In this way evolution has produced a vast repertoire of architecturally distinct and functionally diverse multi-domain proteins. The domain architecture of a protein has traditionally been the realm of structural biology but modern bioinformatic algorithms now enable rapid identification of a protein's separable domains (4). Understanding the modular architecture of multi-domain proteins has proven key to understanding their function.

Approximately 40% of the eukaryotic proteome does not fold into globular domains (5). These protein sequences, which do not possess stable secondary structure, exist in an ensemble of dynamic configurations and are described as being intrinsically disordered (6). Despite their lack of structure, protein intrinsically disordered regions (IDRs) are known to play essential roles in many physiological and pathophysiological pathways, including transcription (7,8), DNA replication (9), as the etiological agents in certain neurodegenerative proteinopathies (10), and certain IDRs drive protein phase separation which helps organize the cell (11). Regions of protein disorder can be discriminated from globular domains by sequence composition alone (12,13) and there exist many bioinformatic algorithms to predict the disorder propensity of a polypeptide (14).

By definition, IDRs lack a defined spatial architecture and, with the exception of short linear motifs (SLiMs) (15), are thus relatively unrestrained in primary structure. As a result, there are generally lower levels of conservation between orthologous IDRs compared to folded regions (16,17) and IDR primary structure often appears to lack organization. One exception to this are seemingly non-random regions of disordered sequences that are locally enriched in only a small subset of amino acids, so-called Low Complexity Regions (LCRs). Sequence gazing readily identifies LCRs

by their conspicuous local sequence bias but these can also be quantitatively and unbiasedly discriminated on the basis of informational entropy (18). Some IDRs also possess multiple LCR types with distinct functionalities (19,20) and bioinformatic approaches to demarcate unique LCR subsequences on the basis of their composition and amino acid dispersion have recently been reported (20,21). LCRs, however, represent but a fraction of all disordered sequences and it is unknown if sequence-spanning organizational principles are operative in IDRs generally.

The sequence bias observed in many IDR sequences has motivated the development of bioinformatic algorithms to classify disordered protein sequences on the basis of composition. These approaches have focused on annotating IDRs and their sub-regions according to a limited set of *known* compositional varieties (i.e., “flavors”), such as by charged residue content (e.g., polyampholyte or strong polyelectrolyte) (22,23) or bias for a given amino acid type (e.g., polar residues) (24,25). Guided by this concept, sequence analysis tools have been built that will demarcate regions within a protein that match a user-defined composition (26). Although useful, these techniques require *a posteriori* knowledge of IDR flavors. In this sense, these methods are *candidate-based* analysis tools, being very good at determining whether a sequence is or isn’t of the candidate class but not useful in identifying new compositional varieties that exist undiscovered within a sequence.

Here we report the development of a statistically robust computational algorithm that unbiasedly maps compositionally-distinct subsequences within an IDR. Relying on the Chi-Square (χ^2) Test of Homogeneity, our Chi-Score Analysis quantifies variability in the fractional composition of amino acids between two sequences. Applied intramolecularly in a moving-window, matrix-based approach, this method can identify sequence-spanning compositional heterogeneity to parse a protein sequence into regions of distinct amino acid composition, regardless of what those compositions are. With multiple examples we show that IDRs of both low and high sequence

complexity possess local compositional bias that bestows disordered sequences with a non-random, modular architecture. Analogous to the domain architecture of globular proteins, we propose that modules (i.e., compositionally distinct subsequences) represent functionally separable units of disordered sequences, and our unbiased, discovery-based approach to their identification represents a promising new direction in IDR classification. Altogether, these data demonstrate that high-level organizational principles are at work in disordered sequences and motivate future functional studies to understand the role of modules in biology.

2 RESULTS & DISCUSSION

2.1 Development of a bioinformatic algorithm to identify local compositional bias in IDRs.

Protein disordered regions lack stable secondary structure and are thus relatively unrestrained in primary structure. Consistently, IDRs often have weak linear sequence conservation (17,27) and no visually discernible sequence patterns, suggesting they lack a defined organization. We hypothesized that local compositional bias may bestow disordered sequences with a sequence-spanning level of organization undetectable by current methods. We therefore developed a bioinformatic algorithm to determine in an unbiased and quantitative fashion if amino acids are non-randomly distributed across the length of a protein sequence and to map this information back onto the sequence. Our approach implements the Chi-Square (χ^2) test statistic to compare the fractional content of amino acids between two sequences as a measure of compositional dispersion (see **METHODS AND ALGORITHMS**). Applied in this way, the chi-score quantifies how different the amino acid proportions (i.e., composition) are between two sequences.

The chi-score metric can be applied intramolecularly to identify compositionally biased regions (**Figure 1A**). In the first step of our algorithm, the sequence is broken into all possible subsequences for a specified window size and all pairwise chi-scores are calculated (**Figure 1A-a**). Each chi-score is then converted to a Pearson's correlation coefficient which better resolves

compositionally biased regions and sequence-spanning patterns (**Figure 1A-b**). Subsequently, the mean correlation coefficient of each subsequence is calculated from a subsequence-centered square window and these “insulation scores” are plotted against residue position (**Figure 1A-c**). Finally, local minima from the insulation score plot are calculated and recorded as potential boundaries between compositionally distinct regions (**Figure 1A-d**). These steps are completed for nine sets of pairwise chi-scores – each using a different even integer window size between 6 and 22 to define the original subsequences – which results in nine sets of boundaries (**Figure 1A-d**).

The boundaries, which were originally grouped by window size, are then clustered by residue proximity (**Figure 1A-d**) and the optimal boundary positions are determined by selecting the placements within each group that maximize the mean chi-score between the surrounding regions (**Figure 1A-e**). The statistical strength of each boundary is then determined by calculating a z-score. To do this, the two juxtaposed regions separated by each boundary are scrambled 500 times and the boundary position resulting in the maximum chi-score identified. This results in a set of 500 chi-scores, one for each scramble, from which the mean (and standard deviation) is determined and used to calculate the z-score for the corresponding boundary (**Figure 1A-f**). Finally, low scoring boundaries are iteratively removed and those that remain are re-optimized and scored until only high-confidence boundaries remain.

To demonstrate the utility of this algorithm in identifying compositional bias, we first tested its effectiveness at differentiating human languages (**Figure 1B**). We translated a quote into English (Eng), Japanese (Jpn), Yoruba (Yrb), Spanish (Spn), and Polish (Pol), appended the translations one after another, and then analyzed the resulting string of text with the Chi-Score Analysis. This approach proved highly effective at discriminating between languages based on their differential character usage (all texts were Romanized and composed of the same 26 letter alphabet).

Regions off the diagonal with relatively high correlation reveal languages with more similar alphabet usage, such as English and Spanish or Yoruba and Japanese. Conversely, regions off the diagonal with relatively low correlation reveal languages with differential alphabet usage, such as Polish and Yoruba. We next applied the method to a protein sequence to determine whether it can parse a sequence by "molecular language" (**Figure 1C**). Fused in Sarcoma (FUS) is known to possess multiple compositionally biased regions with unique functionality (28,29) and our method accurately identifies these, revealing three major language types: G/S-Y-G/S repeats, RGG repeats, and a more complex sequence type which corresponds to the folded domains of the protein (RRM and Zn finger domains). Analysis of off-diagonal correlated regions reveals homology amongst the three RGG-enriched regions and the two regions which have a globular structure. These data establish the utility of the Chi-Score Analysis in parsing sequences by local compositional bias.

2.2 IDRs have a non-random, modular organization.

Having established the algorithm on a sequence with conspicuous local sequence bias (**Figure 1C**), we next applied it to a disordered region that lacks recognizable sequence patterning. The Origin Recognition Complex (ORC, composed of Orc1-6) is an essential DNA replication initiation factor that contains an IDR in the Orc1 subunit that is necessary for recruitment to chromatin (9). In *C. elegans*, the Orc1 IDR is predicted to be 245 amino acids long and, except for two short, low complexity regions, has no apparent sequence organization (**Figure 2A**). However, Chi-Score Analysis reveals a strikingly non-random, sequence-spanning level of organization with regions of distinct compositional bias juxtaposed in a repetitive pattern (**Figure 2B**). Given the modular appearance, we will refer to these compositionally biased regions as "modules". The Orc1 IDR has module types which can be loosely categorized as either basic, neutral, or acidic. In this sequence the basic modules (residues 1-51, 106-167, 186-232) and acidic modules (residues 52-68, 80-105, 168-185, 233-245) alternately repeat throughout the sequence, with the neutral

module type appearing only once (residues 69-79). While charge-based classification is convenient, a more careful investigation of module sequences (**Table 1**) reveals greater complexity than a simple alternation of charge, with non-charged residues also being differentially patterned. Notably, the majority of Orc1 IDR modules identified by the Chi-Score Analysis are not annotated in Uniprot as having "Compositional bias", emphasizing the importance of our discovery-based approach (see **Supplemental File 1** for a list of module sequences and a comparison with Uniprot annotations).

The discovery of a statistically non-random, sequence-spanning level of organization in the Orc1 IDR was unexpected. Importantly, amino acid content alone does not produce this type of organization. To demonstrate this, we reran the analysis on a random scramble of the Orc1 IDR (**Figure 2C**). In this example, modules are not only visually absent but there were additionally no statistically significant boundaries output by our algorithm. Likewise, compositionally biased modules are absent from Orc1's folded domains (**Figure 2 – figure supplement 1A**) which, compared to the IDR, have a far more uniform sequence landscape. This is consistent with prior work showing that the lengthwise distribution of amino acids in globular domains does not differ substantially from randomized sequences containing equivalent proportions of amino acids (30–32). These data suggest that local compositional bias is an organizational principle unique to disordered sequences.

2.3 Many disordered sequences have local compositional bias.

The strikingly modular architecture of FUS (**Figure 1C**) – a low complexity IDR – as well as Orc1 (**Figure 2B**) – a high complexity sequence – prompted us to investigate whether this type of organization is widely operative in disordered sequences. We therefore extended our studies to assess modularity of several other protein disordered regions with known biological and pathological significance, including human TDP-43 (**Figure 3A**), Spt6 (**Figure 3B**), Nucleolin

(NCL, **Figure 3C**), KMT2B (**Figure 3D**), and *Caulobacter crescentus* Ribonuclease (RNase) E (**Figure 3E**). These analyses, which we briefly describe below, show that local compositional bias is pervasive amongst IDRs and bestows disordered sequences with a modular architecture. Conversely, local compositional bias appears largely absent from folded domains, at least for the proteins under consideration here (analysis of full-length sequences is shown in **Figure 2 – figure supplement 1B-F**).

We analyzed an internal IDR of the protein TDP-43 (residues 259-414, **Figure 3A**), an RNA processing factor that forms pathological aggregates in neurodegenerative disease. Our analysis identified five modules, one of which corresponds precisely to a region harboring an abundance of pathological missense mutations and which functions independently as TDP-43's amyloidogenic core (residues 320-PAMMAAAQAALQSSWGMMGMLAS-342) (33–36). These data suggest that modules, like a folded domain, can behave as functionally separable units. Interestingly, none of the TDP-43 modules identified by our analysis are annotated in Uniprot (**Supplemental File 1**). Chi-Score Analysis of the N-terminal IDR of Spt6 reveals several highly distinct modules (residues 1-310, **Figure 3B**) which, together with the prior literature (8), suggests that modules can also demonstrate emergent behavior, with their functionality derived from the cooperative interactions of modules. Specifically, we identified a repetitive pattern of basic and acidic modules, and recent data show that these alternating blocks of charge mediate selective partitioning of Spt6 into MED1 condensates to control transcriptional activation (8). We anticipate that understanding module types and patterning will help elucidate the rules underlying selective partitioning in biomolecular condensates. This idea is supported by Chi-Score Analysis of full-length Nucleolin (NCL, **Figure 3C**), where we identify a pattern of modules within its N-terminal IDR that was recently shown to mediate the protein's selective partitioning into the nucleolus (37). Finally, we identified numerous modules of variable composition and size within the N-terminal

IDR of KMT2B (**Figure 3D**) and the C-terminal IDR of RNase E (**Figure 3E**). Module sequences for each protein can be found in **Supplemental File 1**.

Altogether, these data suggest that local compositional bias represents an organizational principle that is widely operative in disordered sequences and inspires fundamental questions about the role of IDR modularity in biology and disease. A significant advantage of the Chi-Score Analysis method is that compositionally-distinct regions can be discovered solely on the basis of amino acid bias without user-defined search criteria. Even in the small set of sequences analyzed here, this unbiased approach to module identification suggests a level of IDR compositional diversity (i.e., flavors) that dwarfs existing classification paradigms (**Supplemental File 1**).

3 CONCLUSION

We find that the sequences of both low and high complexity IDRs are non-randomly organized into regions with local compositional bias. This concept, like the structural hierarchy (1° – 4° structure), bears the hallmarks of a fundamental organizational principle: it is generalizable to any sequence, it appears to be broadly operative, and it enables the unbiased sub-division of a sequence into component parts. These features have enabled the structural hierarchy to provide a comprehensive classification of folded sequence space, a long-standing objective in the field of protein disorder. Collectively, these findings warrant a conceptual shift away from IDRs as indivisible functional units to viewing them as modular sequences with region-specific physical, chemical, and functional properties (**Figure 3F**). To this extent, individual IDR modules and their combinations may represent the functional analog of the globular protein's domain, whose many types and arrangements produce the diversity of cellular functions that are needed for life. This concept has empirical support from studies of transcription factor transactivation domains (7) and prion domains (37,38) where sub-sequences in extended regions of disorder contribute specific functionality strictly on the basis of their composition.

This work calls for the comprehensive classification of module types and their combinations to determine whether there exist distinct classes or a continuum of compositional varieties. Such work will benefit from other sequence characterization parameters, such as charge distribution (22) and binary sequence patterning (39). While the evolutionary preservation of modules provides strong *a priori* evidence that modules are important for biology, future studies are needed to systematically relate module type with functionality. Beyond this, many other important questions remain, including the use of genetic mechanisms to shuffle modules and produce novel functionalities, the biophysical properties of isolated modules and emergent properties of multi-module sequences, and whether a modular view of IDRs clarifies the mechanism of disease-associated mutations and rationalizes specificity in IDR-enriched biomolecular condensates. In the immediate, we hope the concept of IDR modularity as a generalizable organizational principle serves as a framework for generating hypotheses for IDR functional mechanisms and provides a rational approach for dissecting this enigmatic class of sequences (**Figure 3F**). To facilitate these types of studies, the code required to run these analyses is freely available and easily implemented (see **METHODS AND ALGORITHMS**).

4 METHODS AND ALGORITHMS

4.1 Applying the Chi-square test to compare sequence composition.

The Chi-Square Test of Homogeneity is used to determine whether two distributions are from, or were sampled from, the same population. Traditionally, the Chi-Square test statistic is calculated and used to either reject or accept the null hypothesis. Here, the test statistic is instead used as a metric scoring the compositional difference between two sequences; a high-test statistic indicates a high degree of compositional distinction.

The Chi-square test can be applied both intermolecularly, comparing the amino acid content of different protein sequences, and intramolecularly, comparing the amino acid content of subsequences within a single protein. This latter application, which is elaborated on in the following section, applies a matrix-based approach to parse a sequence into compositionally distinct regions from pairwise subsequence comparisons. When two sequences are compared, the number of each residue is first taken as the observed values (O) for the Chi-Square formula. For each observed value, a corresponding expected value (E) is calculated with the following equation.

$$E_{r,n} = \frac{\sum_r O_{r,n}}{\sum_r O_{r,1} + \sum_r O_{r,2}} \left(O_{r,1} + O_{r,2}\right)$$

In this formula, *n* refers to the sequence (either 1 or 2) and *r* refers to the residue (one of twenty amino acids). To determine the expected value for alanine residues in the first sequence, the total number of alanine residues in the two sequences is multiplied by the ratio of that sequence's length to the total length of both sequences. The test statistic can then be calculated with each observed/expected pair using the following equation.

$$\chi^2_{r,n} = \frac{\left(O_{r,n} - E_{r,n}\right)^2}{\left(E_{r,n}\right)^2}$$

Finally, the 40 unique scores – one for each amino acid and sequence pair – are summed, and the test statistic is normalized between zero and 1. This is done by dividing this sum by the maximum possible score for those two sequences, which is equal to the sum of their lengths and occurs when they have no residues in common. Sequences that have no residues in common will

always receive a normalized score of one, and sequences with identical amino acid compositions will receive a normalized score of zero.

In addition to the chi-score method, Euclidean distance has also been applied to quantitate the similarity in amino acid composition between disordered sequences (40,41). This method takes the fractional content of amino acids as Cartesian coordinates in a 20-dimensional Euclidean space and the “distance” between any two sequences is quantified. Our chi-score method builds on Euclidean distance in several ways. First, chi-score values can be readily decomposed to see the contribution of each amino acid to the overall score, thereby determining the residue type(s) that most distinguishes one sequence from another. Second, Euclidean distance is influenced by sequence complexity while chi-score is not. Finally, the Chi-Square test possesses inherent statistical power which we apply to determine whether subsequences are compositionally distinct compared to random scrambles of the same sequence.

4.2 Applying the Chi-Score Analysis intramolecularly to identify regions of local compositional bias.

The Chi-Score Method can be applied intramolecularly to identify regions of distinct amino acid compositional bias. During the first step of this analysis, the sequence is broken up into all possible subsequences of nine different window sizes (all even integers between 6 and 22) and for each window size the chi-score is calculated for all subsequence pairs. The pairwise scores are then converted to Pearson’s Correlation Coefficients with the following equation:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

After plotting the coefficients as a two-dimensional matrix with the subsequence positions along each axis, we then calculate insulation scores (as used in Hi-C data processing (42)) for each residue position. These scores are calculated by taking the average value in a square window that slides along the main diagonal. Insulation score values are high when this square is within a region of distinct amino acid composition and low when the square is near the boundaries between them. Therefore, we then take the residue positions of the local minima of these insulation scores as potential boundaries between compositionally distinct modules.

Because this is done with nine different window sizes, we now have multiple sets of potential boundaries grouped by the window size used to calculate them. Modules of different length and/or complexity will be best identified with different window sizes, so the precise positions for each boundary will vary between the nine groups. We then regroup these boundaries spatially so that they are clustered with other potential positions for the same boundary. The optimal position for each boundary is then determined by selecting those that maximize the chi-scores between the resulting modules.

The boundaries can now be statistically verified by calculating a z-score for each. First, the two modules separated by a boundary are juxtaposed and randomly scrambled 500 times. Then, the greatest chi-score between two modules that can be achieved with a comparable boundary is determined for each scramble. Finally, these scores are used to convert the raw chi-score for that boundary into a z-score, which tells us how likely it is that the predicted boundary separates truly distinct modules or simply identifies local biases occurring by chance. Boundaries with low z-scores are iteratively removed until only those with significant scores remain. After a boundary is removed, the positions are reoptimized and the z-scores are recalculated for those that remain; the placements and z-scores corresponding to each iteration are also stored so that they can be recalled as desired.

4.3 Python implementation and accessibility.

The code to perform these analyses was designed with Python Version 3.10.8 and executed in Jupyter Notebook. To allow for easy access and implementation of the analysis, we have made it freely available in a number of formats: 1) the source code, which contains all functions necessary to perform and manipulate the analysis as desired, 2) a streamlined Python notebook that installs the algorithm and performs the analysis on a single protein sequence, 3) a Python notebook that performs the analysis on a step-by-step basis so that the outputs of each can be recalled as desired, and 4) a Google Colab notebook that lets the user input a sequence and adjust optional parameters. For the easiest implementation of this analysis, we recommend using the Google Colab notebook, which can be found at: https://colab.research.google.com/github/MWPlabUTSW/Chi-Score-Analysis/blob/main/ChiScore_Analysis.ipynb. For instructions on how to use the other implementations, as well as the code required to reproduce all matrices shown throughout the paper, please see: https://github.com/MWPlabUTSW/Chi-Score-Analysis.git.

ACKNOWLEDGEMENTS

We thank Xiaochen Bai, Nick Grishin, Michael Rosen, and Weiwei Wang for critical reading of the manuscript. M.W.P. is the Cecil H. and Ida Green Endowed Scholar in Biomedical Computational Science. This work was supported by The Welch Foundation (I-2074-20210327, to M.W.P.), the Cancer Prevention and Research Institute of Texas (CPRIT, RR200070, to M.W.P.), and the National Science Foundation (NSF, 2308642, to M.W.P). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of manuscript.

FIGURES, FIGURE TITLES AND LEGENDS

FIGURE 1

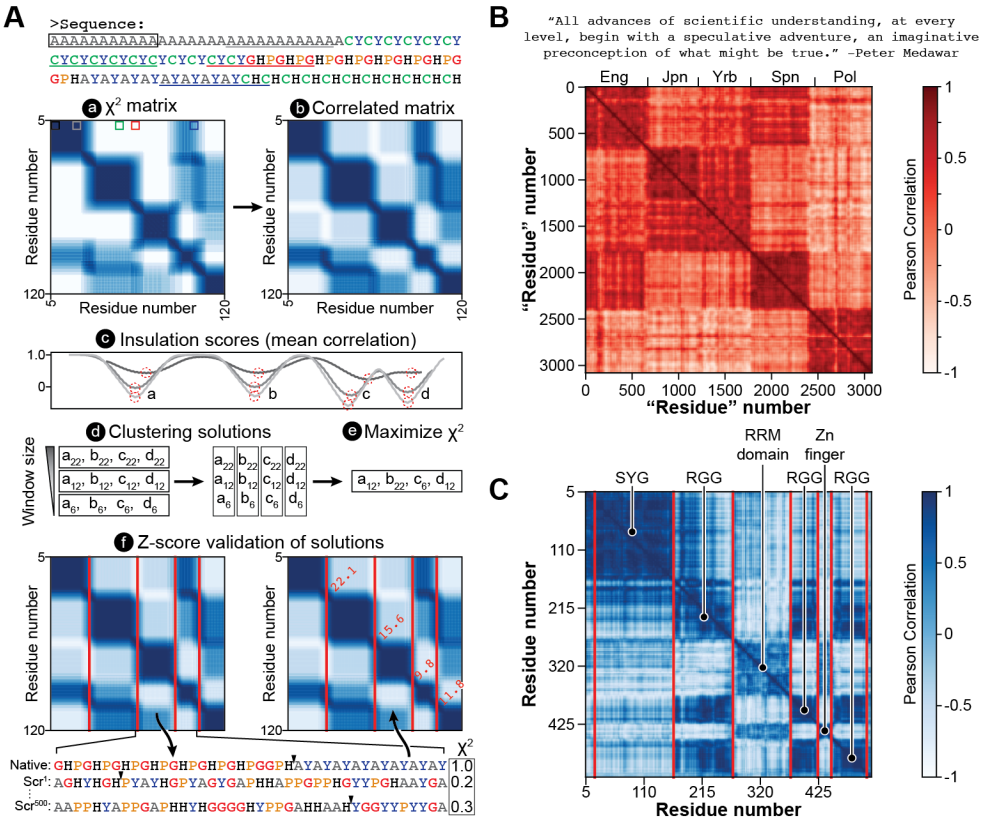


Figure 1: A bioinformatic algorithm to measure compositional dispersion across a sequence. A) For a given sequence, a) pairwise chi-scores are calculated for all window-defined subsequences and plotted as a matrix. Window sizes include every even size from 6-22. Then, for each window size, b) chi-scores are converted to Pearson's correlation coefficients, c) insulation scores are calculated to define local minima which represent compositional boundaries, and d) the minima identified for all window sizes are grouped according to proximity. From these groups, e) each boundary placement is optimized to maximize chi-score and f) z-scores are calculated for each boundary element and are used to iteratively remove non-significant boundaries. B) The Chi-Score Analysis method can distinguish between human languages on the basis of alphabet usage. A quote from Peter Medawar's Romane Lecture was translated into five languages and each translation was strung end-to-end for analysis. "Eng" = English, "Jpn" = Japanese, "Yrb" = Yoruba, "Spn" = Spanish, and "Pol" = Polish. C) The Chi-Score Analysis method can distinguish

between regions of sequence bias in the human protein FUS. Boundaries are shown for 95% confidence level.

FIGURE 2

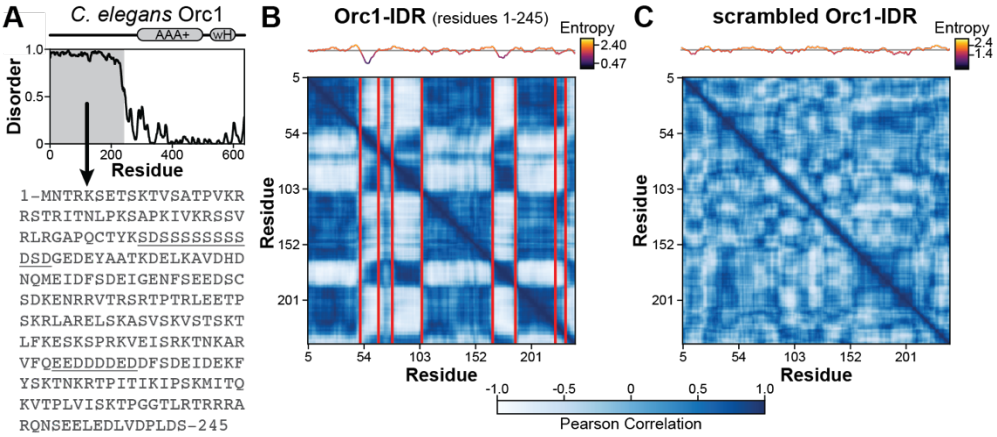


Figure 2: High-complexity disordered regions are non-randomly organized. A) The *C. elegans* Orc1 IDR (Orc1-IDR) has a long N-terminal IDR (residues 1-245) as predicted by Metapredict (43) but no obvious sequence patterning. B) Chi-Score Analysis reveals a strikingly modular architecture for Orc1-IDR with alternately repeating blocks of like-type sequences. Boundaries are shown for 95% confidence level and a sequence entropy plot is shown above the matrix. C) Modularity is lost when the sequence of Orc1-IDR is randomly scrambled.

FIGURE 3

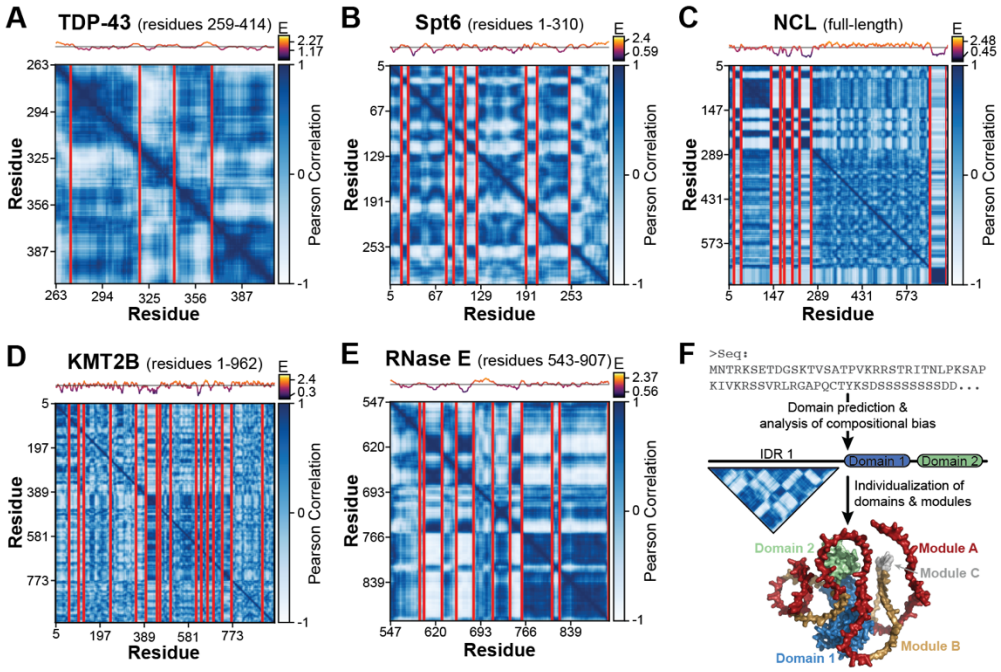


Figure 3: Many intrinsically disordered regions have a modular architecture. Chi-Score Analysis for A) human TDP-43 IDR (residues 259-414), B) human Spt6 N-IDR (residues 1-310), C) human Nucleolin (NCL), D) human KMT2B N-IDR (residues 1-962), and E) *C. crescentus* RNaseE C-IDR (residues 543-907). Boundaries are shown for 95% confidence level and a sequence entropy ("E") plot is shown above each matrix. NCL contains four tandem RNA binding domains which are contained within the largest module in the matrix and appear compositionally uniform. F) IDRs should not be thought of as indivisible units but as modular sequences with region-specific physicochemical features and biological function.

FIGURE 2 – FIGURE SUPPLEMENT 1

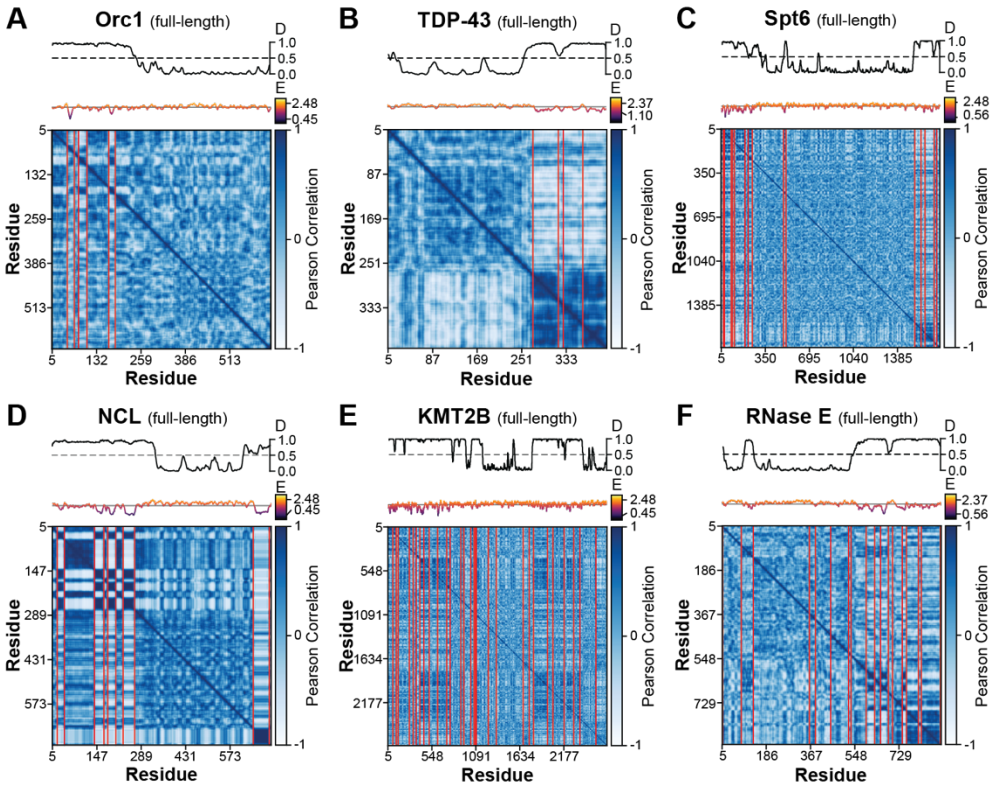


Figure 2 – figure supplement 1: Compositionally-defined modules are largely absent from protein folded domains. Chi-score analysis for full-length A) *C. elegans* Orc1, B) human TDP-43, C) human Spt6, D) human Nucleolin (NCL), E) human KMT2B, and F) *C. crescentus* RNaseE. Boundaries are shown for 95% confidence level and a sequence disorder ("D") and entropy ("E") plot are shown above each matrix.

TABLE 1

Module #	Module Sequence
1	MNTRKSETSKTVSATPVKRRSTRITNLPKSAPKIVKRSSVRLRGAPQCTYK
2	SDSSSSSSSSDSDGED
3	EYAATKDELKAV
4	DHDNQMEIDFSDEIGENFSEEDSCSD
5	KENRRVTRSRTPTRLEETPSKRLARELSKASVSKVSTSKTLFKESKSPRKVEISRKTNKARV
6	FQEEDDDDEDDFSDEIDEKF
7	YSKTNKRTPITIKIPSKMITQKVTPLVISKTPGGT
8	LRTRRRARQ
9	NSEELEDLVDPLDS

Table 1: Module sequences for the Orc1 IDR.

REFERENCES

1. Wetlaufer DB. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins (protein structure/chain continuity/independent regions/self-assembly). Vol. 70. 1973.
2. Patthy L. Introns and exons. Curr Opin Struct Biol. 1994 Jun 1;4(3):383–92.
3. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: Increased structural coverage of functional space. Nucleic Acids Res. 2021 Jan 8;49(D1):D266–73.
4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.
5. Uversky VN. Intrinsically disordered proteins and their “Mysterious” (meta)physics. Vol. 7, Frontiers in Physics. Frontiers Media S.A.; 2019. p. 10.
6. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. J Mol Graph Model. 2001 Feb 1;19(1):26–59.
7. Sigler PB. Acid blobs and negative noodles. Vol. 333, Nature. Nature Publishing Group; 1988. p. 210–2.
8. Lyons H, Veettil RT, Pradhan P, Fornero C, De La Cruz N, Ito K, et al. Functional partitioning of transcriptional regulators by patterned charge blocks. Cell. 2023 Jan 19;186(2):327-345.e28.
9. Parker MW, Bell M, Mir M, Kao JA, Darzacq X, Botchan MR, et al. A new class of disordered elements controls DNA replication through initiator self-assembly. Elife. 2019;
10. Uversky VN. Intrinsic disorder in proteins associated with neurodegenerative diseases. Front Biosci. 2009 Jun 1;14(14):5188–238.
11. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Vol. 44, Biochemical Society Transactions. Portland Press Ltd; 2016. p. 1185–200.

12. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins Struct Funct Genet. 2000 Nov 15;41(3):415–27.
13. Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett. 2004 Oct 22;576(3):348–52.
14. Katuwawala A, Kurgan L. Comparative assessment of intrinsic disorder predictions with a focus on protein and nucleic acid-binding proteins. Vol. 10, Biomolecules. MDPI AG; 2020. p. 1–18.
15. Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol Biosyst. 2012;8(1):268–81.
16. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Vol. 21, Current Opinion in Structural Biology. Elsevier Current Trends; 2011. p. 441–6.
17. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 2002;55(1):104–10.
18. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem. 1993;17(2):149–63.
19. Kim GH, Kwon I. Distinct roles of hnRNPH1 low-complexity domains in splicing and transcription. Proc Natl Acad Sci U S A. 2021 Dec 14;118(50):e2109668118.
20. Lee B, Jaberi-Lashkari N, Calo E. A unified view of low complexity region(LCRs) across species. Elife. 2022 Sep 1;11.
21. 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 Genomics Bioinforma. 2021 Apr 9;3(2):1–19.
22. Das RK, Pappu R V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A.

2013;

23. Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu R V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. Biophysj. 2017;112:16–21.

24. Necci M, Piovesan D, Tosatto SCE. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. Protein Sci. 2016 Dec 25;25(12):2164–74.

25. Necci M, Piovesan D, Clementel D, Dosztányi Z, Tosatto SCE. MobiDB-lite 3.0: Fast consensus annotation of intrinsic disorder flavors in proteins. Bioinformatics. 2020 Dec 1;36(22–23):5533–4.

26. Millard PS, Bugge K, Marabini R, Boomsma W, Burow M, Kragelund BB. IDDomainSpotter: Compositional bias reveals domains in long disordered protein regions—Insights from transcription factors. Protein Sci. 2020 Jan 11;29(1):169–83.

27. Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife. 2019;

28. Murthy AC, Tang WS, Jovic N, Janke AM, Seo DH, Perdikari TM, et al. Molecular interactions contributing to FUS SYGQ LC-RGG phase separation and co-partitioning with RNA polymerase II heptads. Nat Struct Mol Biol. 2021 Nov 1;28(11):923–35.

29. Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, et al. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell. 2018;

30. White SH, Jacobs RE. The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J Mol Evol. 1993 Jan;36(1):79–95.

31. Mitra CK, Rani M. Protein sequences as random fractals. J Biosci. 1993 Jun;18(2):213–20.

32. White SH. The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. J Mol Evol. 1994;38(4):383–94.

33. Jiang LL, Che MX, Zhao J, Zhou CJ, Xie MY, Li HY, et al. Structural transformation of the amyloidogenic core region of TDP-43 protein initiates its aggregation and cytoplasmic inclusion. J Biol Chem. 2013 Jul 5;288(27):19614–24.

34. Cao Q, Boyer DR, Sawaya MR, Ge P, Eisenberg DS. Cryo-EM structures of four polymorphic TDP-43 amyloid cores. Nat Struct Mol Biol. 2019 Jul 1;26(7):619–27.

35. Lin Y, Zhou X, Kato M, Liu D, Ghaemmaghami S, Tu BP, et al. Redox-mediated regulation of an evolutionarily conserved cross-β structure formed by the TDP43 low complexity domain. Proc Natl Acad Sci U S A. 2020 Nov 17;117(46):28727–34.

36. Conicella AE, Zerze GH, Mittal J, Fawzi NL. ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. Structure. 2016 Sep 6;24(9):1537–49.

37. Ross ED, Edskes HK, Terry MJ, Wickner RB. Primary sequence independence for prion formation. Proc Natl Acad Sci U S A. 2005 Sep 6;102(36):12825–30.

38. Ross ED, Baxa U, Wickner RB. Scrambled Prion Domains Form Prions and Amyloid. Mol Cell Biol. 2004 Aug 15;24(16):7206–13.

39. Cohan MC, Shinn MK, Lalmansingh JM, Pappu R V. Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins. J Mol Biol. 2022 Jan 30;434(2):167373.

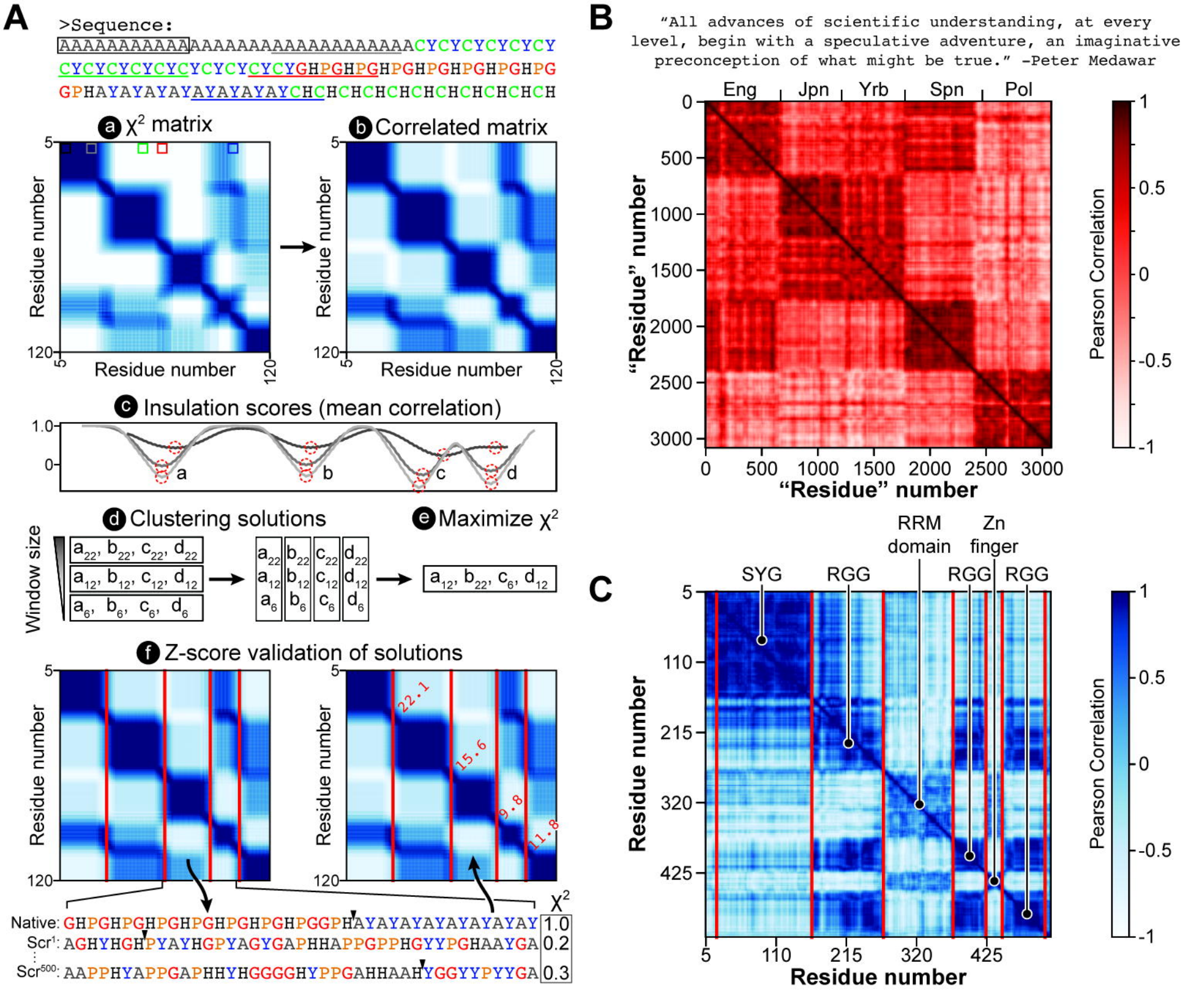
40. Patil A, Teraguchi S, Dinh H, Nakai K, Standley DM. Functional annotation of intrinsically disordered domains by their amino acid content using IDD navigator. In: Pacific Symposium on Biocomputing. 2012. p. 164–75.

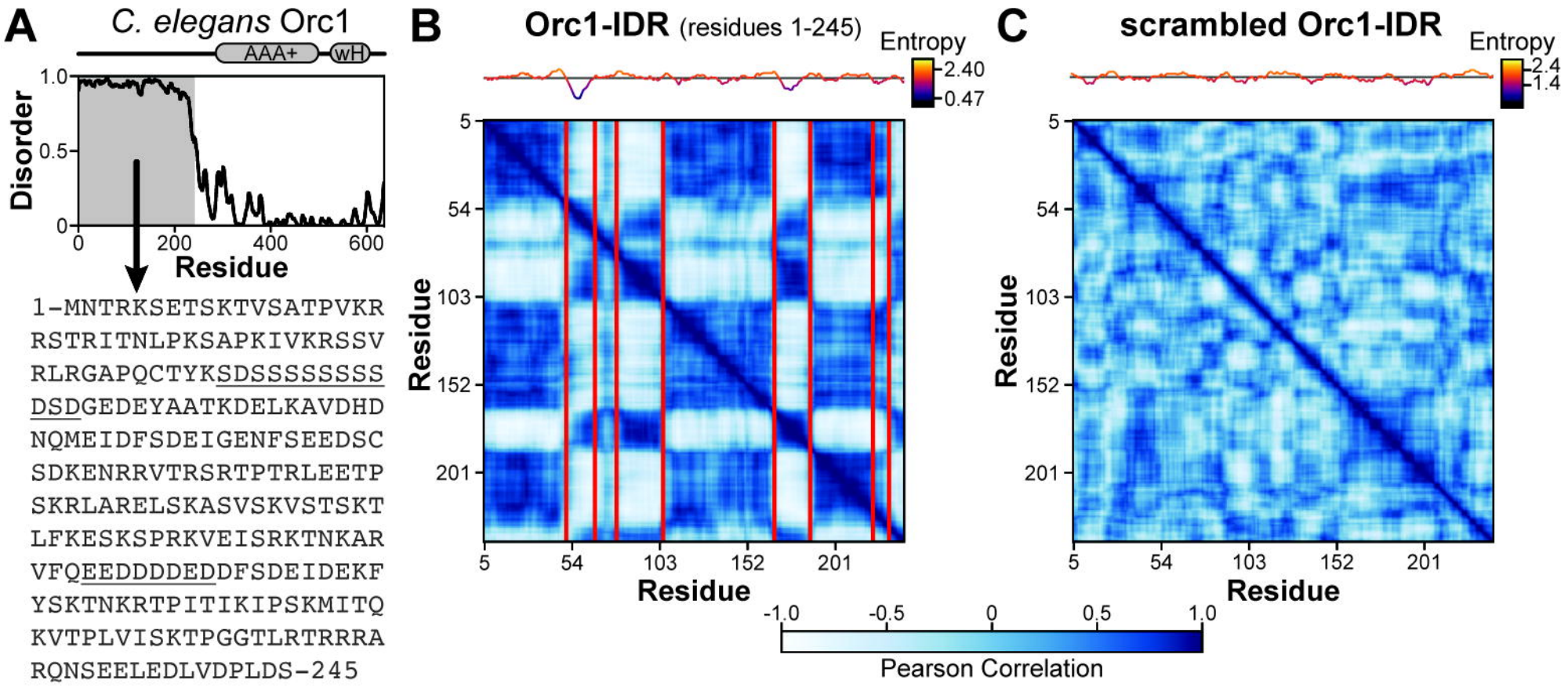
41. Moesa HA, Wakabayashi S, Nakai K, Patil A. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their

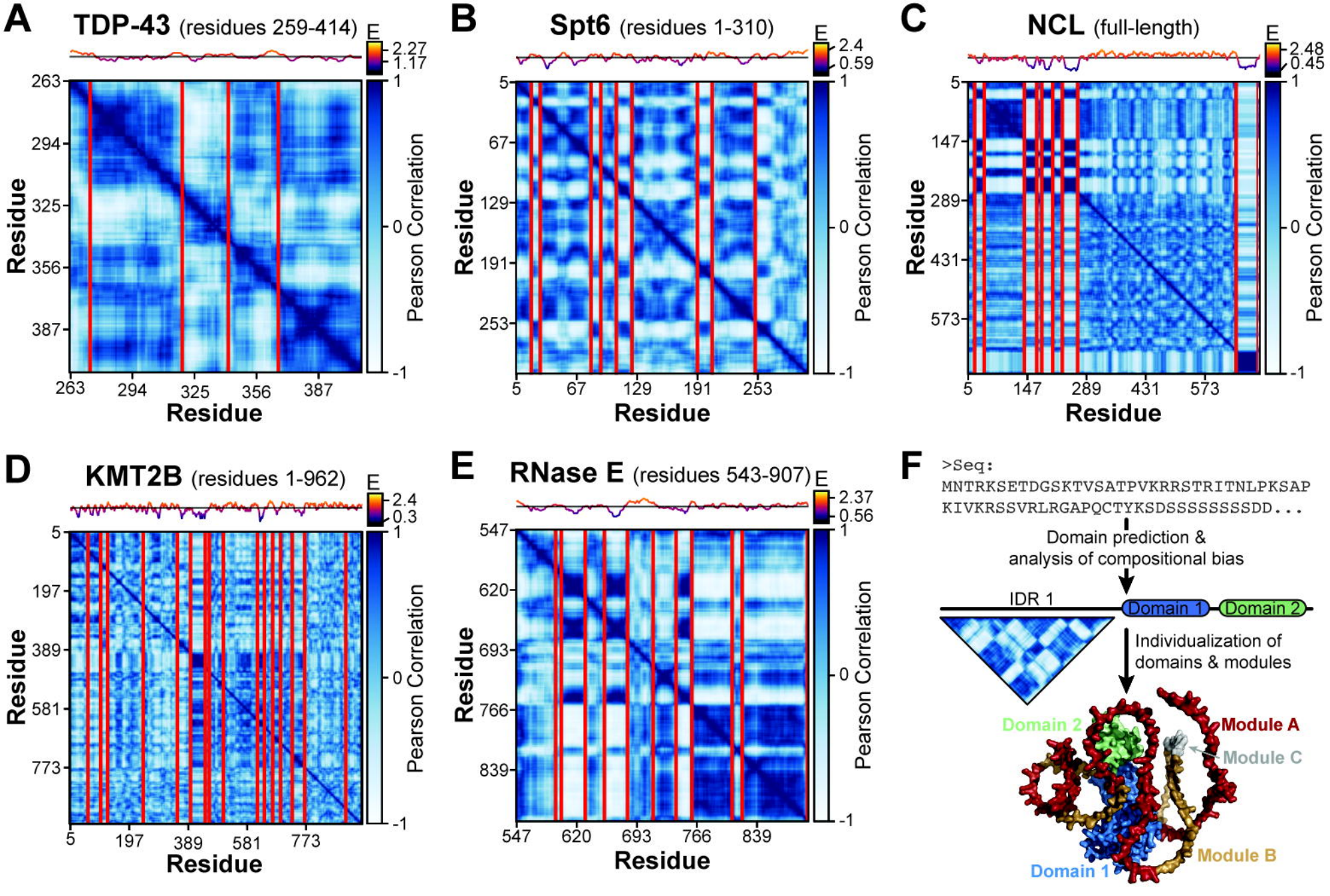
classification. Mol Biosyst. 2012;8(12):3262–73.

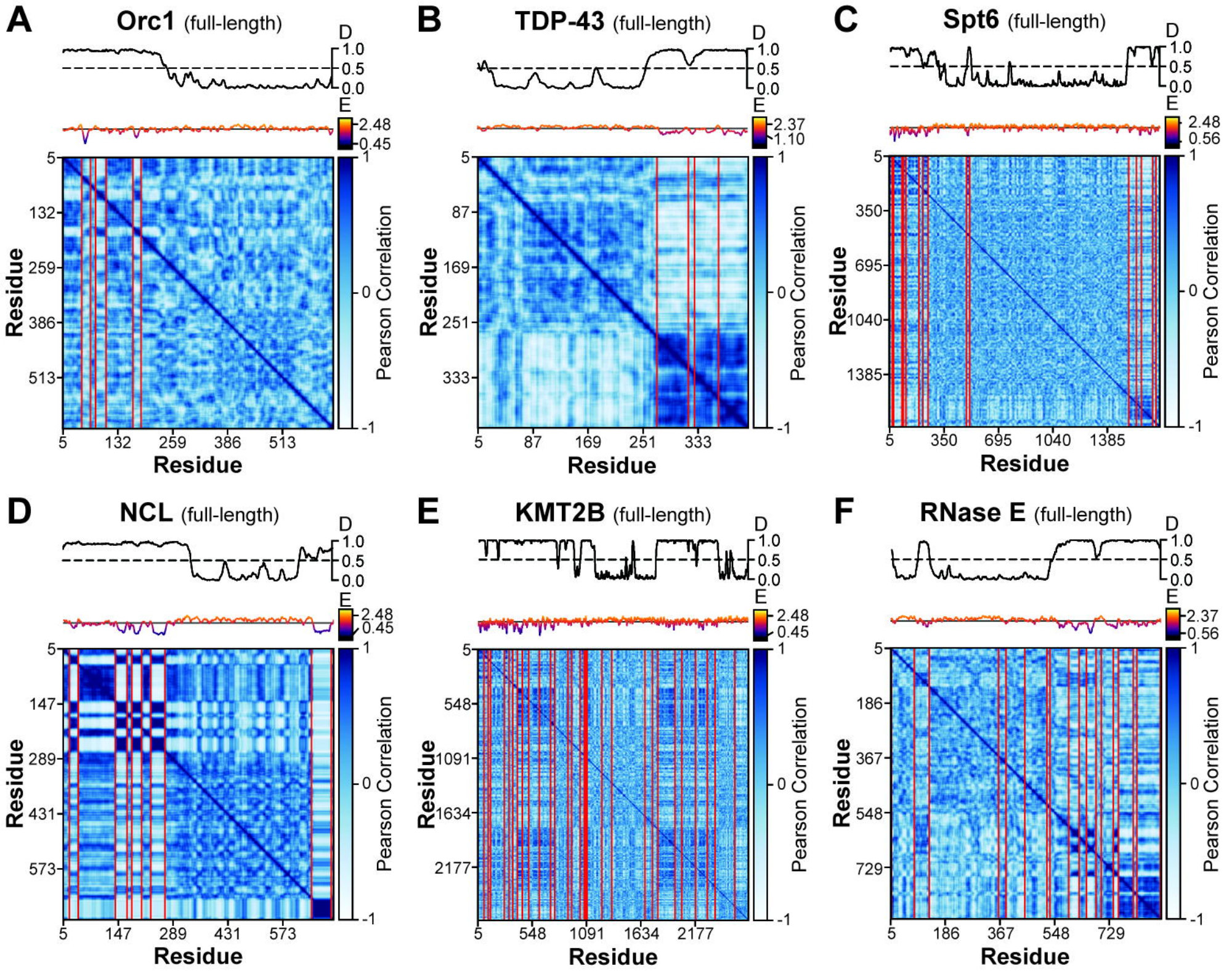
42. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods. 2015;72(C):65–75.

43. Emenecker RJ, Griffith D, Holehouse AS. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Biophys J. 2021 Oct 19;120(20):4312–9.


A
>Sequence:
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACYCYCYCYCYCY
CYCYCYCYCYCYCYCYCYCYGHPGHPGHPGHPGHPGHPGHPG
GPHAYAYAYAYAYAYAYAYAYCHCHCHCHCHCHCHCHCHCHCH
a χ2 matrix
b Correlated matrix
Residue number
c Insulation scores (mean correlation)
a
b
c
d
d Clustering solutions
e Maximize χ2
Window size
f Z-score validation of solutions
22.1
15.6
9.8
11.8
χ2
Native: GHPGHPGHPGHPGHPGHPGHPGHPGGPHAYAYAYAYAYAYAYAY 1.0
Scr1: AGHYHGHPYAYHGPYAGYGAPHHAPPGPPHGYYPGHAAYGA 0.2
Scr500: AAPPHYAPPGAPHHYHGGGGHYPPGAHHAAHYGGYYPYYGA 0.3
B
"All advances of scientific understanding, at every level, begin with a speculative adventure, an imaginative preconception of what might be true." -Peter Medawar
Eng
Jpn
Yrb
Spn
Pol
"Residue" number
Pearson Correlation
C
SYG
RGG
RRM domain
Zn finger
RGG
RGG
Residue number
Pearson Correlation


A
C. elegans Orc1
AAA+
wH
Disorder
1.0
0.5
0
0
200
400
600
Residue
1-MNTRKSETSKTVSATPVKR
RSTRITNLPKSAPKIVKRSSV
RLRGAPQCTYKSDSSSSSSSS
DSDGEDEYAATKDELKAVDHD
NQMEIDFSDEIGENFSEEDSC
SDKENRRVTRSRTPTRLEETP
SKRLARELSKASVSKVSTSKT
LFKESKSPRKVEISRKTNKAR
VFQEEDDDDEDDFSDEIDEKF
YSKTNKRTPITIKIPSKMITQ
KVTPLVISKTPGGTLRTRRRA
RQNSEELEDLVDPLDS-245
B
Orc1-IDR (residues 1-245)
Entropy
2.40
0.47
Residue
5
54
103
152
201
Residue
C
scrambled Orc1-IDR
Entropy
2.4
1.4
Residue
5
54
103
152
201
Residue
-1.0
-0.5
0
0.5
1.0
Pearson Correlation


A
TDP-43 (residues 259-414)
E
2.27
1.17
263
294
325
356
387
Residue
Residue
1
0
-1
Pearson Correlation
B
Spt6 (residues 1-310)
E
2.4
0.59
5
67
129
191
253
Residue
Residue
1
0
-1
Pearson Correlation
C
NCL (full-length)
E
2.48
0.45
5
147
289
431
573
Residue
Residue
1
0
-1
Pearson Correlation
D
KMT2B (residues 1-962)
E
2.4
0.3
5
197
389
581
773
Residue
Residue
1
0
-1
Pearson Correlation
E
RNase E (residues 543-907)
E
2.37
0.56
547
620
693
766
839
Residue
Residue
1
0
-1
Pearson Correlation
F
>Seq:
MNTRKSETDGSKTVSATPVKRRSTRITNLPKSAP
KIVKRSSVRLRGAPQCTYKSDSSSSSSSSDD...
Domain prediction &
analysis of compositional bias
IDR 1
Domain 1
Domain 2
Individualization of
domains & modules
Domain 2
Module A
Module C
Module B
Domain 1


A
Orc1 (full-length)
B
TDP-43 (full-length)
C
Spt6 (full-length)
D
NCL (full-length)
E
KMT2B (full-length)
F
RNase E (full-length)
D
E
Residue
Pearson Correlation