bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Intrinsically disordered protein ensembles shape evolutionary rates revealing conformational patterns

Nicolas Palopoli[1,*], Julia Marchetti[1,*], Alexander Miguel Monzon[2], Diego Javier Zea[3], Silvio C.E. Tosatto[2], Maria Silvina Fornasari[1] and Gustavo Parisi[1]

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina
[2] Department of Biomedical Sciences, University of Padua, Padua, Italy
[3] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Paris, France

* These authors contributed equally to this manuscript.

Corresponding author: Gustavo Parisi <gusparisi@gmail.com>

Abstract

Intrinsically disordered proteins (IDPs) lack stable tertiary structure under physiological conditions. The unique composition and complex dynamical behaviour of IDPs make them a challenge for structural biology and molecular evolution studies. Using NMR ensembles, we found that IDPs evolve under a strong site-specific evolutionary rate heterogeneity, mainly originated by different constraints derived from their inter-residue contacts. Evolutionary rate profiles correlate with the experimentally observed conformational diversity of the protein, allowing the description of different conformational patterns possibly related to their structure-function relationships. The correlation between evolutionary rates and contact information improves when structural information is taken not from any individual conformer or the whole ensemble, but from combining a limited number of conformers. Our results suggest that residue contacts in disordered regions constrain evolutionary rates to conserve the dynamic behaviour of the ensemble and that evolutionary rates can be used as a proxy for the conformational diversity of IDPs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

The native state of proteins is composed of multiple conformers in a dynamical equilibrium, collectively known as the native ensemble [1]. Ensembles could differ in the accessibility of each conformation to the rest of conformers (limited by the interchange free energy barriers) and how much structurally different the conformers are as a whole (i.e., the conformational diversity of the protein). For globular ("ordered") proteins or domains, the energy barriers for conformational transitions are relatively high, allowing the experimental exploration of discrete and well-established conformations with several techniques [2]. In one extreme, conformers may only account for tiny differences like those observed in the so-called rigid proteins, where slight residue movements or rotations allow the transit of ligands to binding sites, opening tunnels or enlarging cavities without backbone translations [3]. Loops movements, displacements of secondary structure elements and relative rotations of domains commonly contribute to further increase the structural differences between conformers [4]. At the other end of the protein space, in a continuous increase of conformational diversity, proteins display high plasticity and very low conformational energy barriers between conformers [5]. Consequently, these conformers are easily interchanged, difficulting the selection of a representative group that comprehensively describes the native ensemble. These proteins of large and heterogeneous conformational diversity are commonly known as Intrinsically Disordered Proteins (IDPs) or intrinsically disordered regions (IDRs). Their main characteristic is the lack of structurally well-defined conformations under native conditions [6,7] (Figure 1A) and structurally complex native ensembles. Their distinctive properties do not prevent IDPs from being widely distributed and relatively abundant across taxa. For example, ~40% of eukaryotic proteomes [8] and ~46% of Uniprot entries contain short disordered regions [9]. Moreover, they play a crucial role in many neurodegenerative and systemic disorders [10].

As ordered proteins, the native structural ensemble of IDPs is a key concept in two alternative explanations describing the complex series of events associated with ligand binding [11]. In the conformational selection model, a low-populated conformation shows high binding affinity and specificity towards a given biological ligand or partner. Binding events produce a shift in the conformational equilibrium that increases the relative concentration of this competent conformation. Alternatively, the induced-fit model proposes that the adoption of the preferred conformation occurs after binding, in a process that may involve conformational changes and disorder-order transitions. Whatever the relative importance of these models to explain protein biology[12], pre-existence of low-represented competent conformations or conformational

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

rearrangements after ligand binding, raise the notion that IDPs ensembles are not fully disordered and that conformations with residual structure content play key biological roles [5,13–16].

One of the stronger predictors of the global evolutionary rate in proteins is the expression level of the gene [17]. However, each position evolves under different selection pressure. Several studies found correlations between per-site evolutionary rates and different structural features (for a review see [18]). For ordered proteins in general, sites with a large number of contacts, high packing densities or low solvent exposure tend to evolve more slowly [19–21]. Early studies showed that IDPs and disordered regions evolve faster than ordered proteins [22], as expected from their low number of contacts per site [23] and high solvent exposure [13]. Although previous works have detected structurally constrained evolutionary patterns in IDPs [24–26] , their role with IDPs typical higher flexibility remains unclear. In proteins with a significant degree of conformational diversity like IDPs, it could also be expected that different partially folded conformers with transient secondary structure elements or preferred conformations in the ensemble may impose specific and additive constraints on per-site evolutionary rates. Furthermore, these additional constraints derived from the ensemble could reveal functional types of IDPs, particularly associated with order/disorder transitions upon binding [13]. In this manuscript, we address these hypotheses by assessing the impact on evolutionary rates of the structural information derived from experimentally determined IDP ensembles.

Results

Evolutionary rate profiles allow the characterization of IDPs organization

Our dataset consists of data from 310 NMR ensembles of IDPs with at least 60% of disordered positions (see Methods). Figure 1A illustrates the disordered and ordered (globular) regions in these proteins. As expected, a comparison of their normalized evolutionary rate distributions (Figure 1B) shows that disordered positions of IDPs evolve significantly faster than sites from their ordered regions. This heterogeneity in evolutionary rates of IDPs is better understood by inspecting the relationship of position-specific rate profiles with their structural variation in the ensemble. We found that per-site evolutionary rates follow the structural variability between conformers measured as the RMSF (Root Mean Square Fluctuation) of their alpha carbons. Each panel in Figure 2 collects proteins that comply with a given structure-function relationship and thus have similar evolutionary rate (left column) and RMSF (middle column) profiles patterns.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Cartoon depictions of representative ensembles of conformers (right column in each panel) highlight that disordered positions could occur in different regions of the protein. Disordered regions can be found in the middle or in the terminal ends of the PDB polypeptide chain (here, “terminal ends” could refer to either one of the amino or carboxyl ends of a protein or domain boundaries, as about one-third of the polypeptides in our dataset are domains occurring at least 30 residues away from the protein extremes).

Evolutionary rate profiles can be used to distinguish at least five different protein organizations. Two of them that can be easily recognized from their rate profiles comprise an N-terminal disordered region followed by a globular domain, or vice versa (Figure 2 panels A and B, respectively). These protein chains display higher evolutionary rates in regions of high flexibility. The N-terminal domain of the murine heterochromatin protein 1 is an example of a protein with a disorder-order pattern [27] (PDB: 2rvl; figure 2A). The N-terminal end of this domain is polyampholytic and its flexibility is regulated by phosphorylation of its serines, which changes the ensemble distribution promoting histone H3 binding. Figure 2B shows data for protein with an order-disorder organization, such as the alpha-helical MA domain from the HIV-1 Gag polyprotein precursor immediately followed by a highly flexible region (PDB: 2hmx). In the polyprotein, this disordered region links MA to the CA protein N-terminal domain. It has recently been suggested that the linker contraction may enable interactions between the immature MA and CA domains regulating the accessibility of the viral protease to release the mature proteins [28]. Some protein structures display an alternating disorder-order-disorder organization (Figure 2C). The trypanosome protein UBI-1 is one of them, with a single RNA-recognition motif domain flanked by N-terminal and C-terminal disordered regions (PDB: 1u6f) [29]. The ordered region is evidenced by a concomitant depression in both the evolutionary rate and local flexibility profiles. Other proteins have a disordered region as a linker between two globular domains (Figure 2D). As an example, the nuclear immunophilin FKBP25 (PDB: 2mph) has an N-terminal HLH domain and a C-terminal FK506-binding domain, joined by a long and flexible disordered linker that allows proper recognition of DNA by the globular domains [30]. Finally, Figure 2E combines the complex evolutionary rate and RMSF profiles derived from a heterogeneous set of protein chains. They share the presence of several short and flexible regions scattered through the sequence, disordered segments embedded in globular domains, or regions of order-disorder transitions. As an example of this structure-function relationship, the cartoon representation of the nuclear pore targeting domain of yeast nucleoporin NUP116P (PDB: 2aiv) [31] shows regions with defined secondary structure interspersed with flexible loops.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Effects of inter-residue contacts on evolutionary rates

Figure 1 indicates that some disordered positions can also evolve as slowly as ordered positions. To study this evolutionary rate heterogeneity in disordered regions, we estimated the average rates in disordered positions as a function of their sequence distance from the closest ordered residue in globular domains. Disordered positions have an average distance to order of 14.4 amino acids, with a maximum of 161 amino acids. In general, disordered positions tend to evolve faster as the distance to globular regions increases. However, if bigger distances between disordered and ordered regions are considered (>20 residues), relative evolutionary rates become slower and eventually stabilize (Figure 3A). A similar behaviour is observed on the normalized values of RMSF (Figure S1). These trends might be due to the additional constraints imposed by an increased chance of establishing transient long-range interactions with distant regions in the protein, such as those responsible for multisteric regulation (for a review see [32]).

The global evolutionary rates in Figures 1 and 2 and the rate distributions in disordered regions shown in Figure 3A are clear examples of the observed evolutionary rate heterogeneity in IDPs. The significantly different rates in ordered and disordered positions shown in Figure 1 and Figure 2 can be explained by considering the number of contacts established by each residue. The average number of contacts per position in ordered regions doubles that of disordered regions (3.44 vs 1.75, respectively). This difference evidences the higher influence of the structural constraints on the evolutionary rates in ordered regions. The influence of inter-residue contacts on rate heterogeneity extends to disordered positions. The number of contacts per site, averaged between conformers in each ensemble, decreases with their separation from ordered residues (Figure 3B). These contact number distributions show the inverse pattern to the evolutionary rates in Figure 3A.

As we mentioned above, several works studied the influence of protein structure in evolutionary rates (for a review see [18]). However, IDP ensembles are different from the globular protein ensembles analysed in those studies, since their heterogeneous and rapidly interchanging conformations make the number of residue contacts per-site to change substantially among different conformers. Considering the average of the contact number per position along all the conformers in the ensemble, we observed the same negative trend between contacts and normalized evolutionary rates regardless of the order/disorder state (Figure 4A). Interestingly, ordered and disordered regions seem to evolve at more similar rates as the number of contacts increases. We obtained similar results using other representations of per-site contacts, in

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

particular, the maximum, minimum and mode of the number of contacts per position derived from all the conformers in the ensemble (See Supplementary Figure S2).

As IDPs are highly dynamical, some conformers may have a stronger structural relevance than others. We studied the relationship of evolutionary rates with the fraction of conformers having at least one contact per site in at least one conformer. For example, a value of 0.5 means that half of the observed conformers have at least one contact in that position. This fraction can be interpreted as the robustness of a position to maintain at least one contact in the ensemble; therefore, we adopted the term *contact robustness* to represent this fraction. As expected, disordered positions have lower average values of contact robustness (~0.67) than almost all the ordered positions (~0.92) (See Supplementary Figure S3). Figure 4B represents evolutionary rates as a function of contact robustness, showing that rates become slower with increasing contact robustness. Although disordered residues evolve faster regardless of the robustness, similar constraints may be acting on both disorder and ordered positions when a given residue establishes a contact in the vast majority of the conformers (contact robustness >0.9).

Rate profiles and contacts

To further explore the relationship between conformational ensembles and evolutionary rates, we sought to determine if site-specific protein rates are better correlated with the contacts in any single conformer or with contact information derived from the whole ensemble. For each protein, the Pearson correlations of the contact average and the contact robustness per position against the evolutionary rates were calculated. Then, the best correlation for each protein among the different parameters describing the contacts was selected. For ~75% of the proteins in our dataset, the strongest correlation is observed with a single conformer from the ensemble (average ρ = -0.3692). It could be expected that conformers with the maximum number of contacts overall (assumed to be the closed form of the protein) or those with the minimum (likely the open form) may impose relevant functional constraints on evolution. However, none of these accounted for any of the best correlations in our dataset. The remaining 25% of the proteins were more associated with the contact distribution in the ensemble. Contact robustness correlations (average ρ = -0.5448) are highest for 21% of all proteins, largely surpassing the ~3% with the best correlation provided by the mean number of contacts per position (average ρ = -0.4683). Overall, for a quarter of the IDPs in our dataset, simultaneous consideration of the structural information obtained from all the available conformers in the ensemble allows for a better explanation of evolutionary change.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Effect of conformational ensembles on evolutionary rates

In a previous work we found that IDP ensembles are redundant in terms of their structural constraints on evolution, since ~10 conformers are required on average to explain the structural constraints observed in homologous alignments of IDPs [25]. Following this idea, we explored all possible combinations of different numbers of conformers from each ensemble to study if there is a particular subset of conformers establishing contacts that better explain the evolutionary rate profiles. To this end, we first obtained all possible combinations between any number of conformers in each ensemble. Each combination was calculated as the average of residue contacts or the contact robustness per position between the selected conformers. Then, we derived the Pearson correlation coefficient rho (ρ) against the evolutionary rates by using either the contact average (ρ_{ave}) or the contact robustness (ρ_{crob}). We estimated the statistical significance for each combination and protein using a t-test that determines whether the obtained ρ distribution is different from the ρ distribution of a random (bootstrapped) sample (see Methods for more details). We found that in ~52% of the proteins, the evolutionary rates are better correlated with the contact robustness than with any other single conformer- or ensemble-based description of contacts (average ρ_{crob} = -0.5254). In the same sense, in ~30% of the cases, the correlations with the average of contact for a certain combination outperformed the rest (average ρ_{ave} = -0.4154). Individual conformers only provide the best correlations for ~17% of the studied proteins, and with lower ρ values (average ρ_{indiv} = -0.3695). For both the contact average and contact robustness, the correlations can be maximized by considering only a subset from the available conformers in the ensemble. The best number of combined conformers is 4.0 on average using the contact robustness and 2.8 with contact average (Figure 5).

The previous results highlight the importance of the ensemble in shaping evolutionary rates, as the correlations for ~82% of the proteins were improved by using different descriptions of the contacts in the conformational ensemble. Why do the evolutionary rates better correlate with the contact robustness for some proteins, while the contact average provides stronger correlations for others? While contact robustness should better represent IDPs with a considerable structural variation between conformers, contact averages could only capture the behaviour of an ensemble as a whole. We hypothesize that proteins with better ρ_{crob} (Figure 6A) could be more flexible, and their evolutionarily constrained positions could establish contacts with different positions in a given number of conformers. On the contrary, proteins better represented by ρ_{ave} (Figure 6B) might form more closed structures where the number of contacts is relatively constant along the conformers

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in the ensemble. To test this hypothesis, we estimated the Root Mean Square Deviation (RMSD) between all conformers in the ensemble and registered the average and the maximum. Interestingly, for the proteins for which the contact robustness provides the best correlation against evolutionary rates, the mean of the average RMSD is 19.77 Å, and the mean of their maximum RMSD is 35.2 Å. The proteins showing the best correlation against the average number of contacts per position have lower average and maximum RMSD values (with means of 11.12 Å and 20.28 Å, respectively).

Discussion

In a first approximation, the evolutionary rate heterogeneity in IDPs (Figures 1, 2 and 3A) can be explained as a function of the tertiary contacts per site in different conformers of the ensemble (Figure 3B and 4). These results suggest that the state-of-the-art about evolutionary rates in globular proteins could be extended to unstructured proteins. However, IDPs share certain peculiarities. The major difference between globular proteins and IDPs is the high degree of conformational diversity in the latter and their relative ease to interchange between alternative conformers. This particular behaviour implies a marked variability in the tertiary contacts per position between the different conformers. According to our results, ~80% of the IDPs studied show better correlations between evolutionary rates and residue contact profiles when the contacts are extracted from a set of conformers instead of coming from a particular one. The best correlations are found by combining a small (<8) number of conformers rather than the whole ensemble, reflecting distinct contributions of preferred conformations, possibly related to the structure-function relationship of the protein [33,34]. These results agree with previous works describing IDP ensembles as highly redundant [25] and open the opportunity for the development of bioinformatics tools to curate IDPs ensembles databases [35]. Largely ignored in evolutionary studies, except for a few exceptions (for example see [36–38]), contact variations due to conformational changes are revealed as impossible to neglect in IDP ensembles. These variations impose cumulative constraints on evolutionary rates [39] that may reflect the importance of transitory or preferred conformations in the ensemble [13].

It is possible that the observed rate heterogeneity in disordered regions may be partially influenced by the presence of functional modules involved in protein-protein interactions, ligand binding or post-translational modifications. Among these, short linear motifs (SLiMs) are typical interaction sites found in IDPs, usually more conserved than residues in their surroundings [40].

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While we cannot rule these out as possible modulators of evolutionary rates in our dataset, no binding sites were found when searching for matches to known motif patterns in the ELM database [41]. Still, preferred conformations in the ensemble, detected using correlation analysis on combinations of conformers, may evidence residual secondary structures linked with their interactions.

The IDPs in our dataset have different patterns of alternating disordered and ordered regions that were evidenced by their evolutionary rate profiles. Since these profiles reflect the conformational preferences of the proteins, they could be used to infer global behaviours of IDP ensembles at the proteomic level. Interestingly, we observed the modulating effect of folded domains on the rates of attached intrinsically disordered regions, raising the question about the occurrence of coevolutionary processes [42]. Since a majority of the highest correlations with evolutionary rates were obtained using contact robustness, this supports the current view of IDPs as highly dynamic ensembles with multiple weak interactions throughout the entire protein, possibly to allow the stabilization of the ensemble and the interaction with other proteins [43–46]. In general, contact robustness reflects the evolutionary rates better than the contact number average, possibly due to the highly dynamic interactions needed to maintain the ensemble [42].

Our results suggest that structurally-based evolutionary studies are not only feasible but also a promising approach to unveil the biology and evolution of intrinsically disordered proteins.

Materials and Methods

Dataset construction

We built a dataset of IDPs with known ensembles derived from Nuclear Magnetic Resonance (NMR) spectroscopy estimations deposited in the Protein Data Bank (PDB) [47]. We downloaded all PDB files from NMR experiments and selected as IDPs those that have at least 40% of their residues predicted to be disordered by Espritz [48]. We obtained pre-calculated homologous sequence alignments of these proteins from the HSSP database [49], filtering them to allow only ~5% of gaps in average per position and with identity percentages between 33 and 100 against the PDB reference sequence (average ~60%). All remaining proteins were visually inspected, keeping only those with an evident disorder content in their 3D structure. Chimeric proteins were eliminated. This resulted in a curated dataset of 310 IDPs, with an average length of ~130 residues (minimum= 58, maximum= 370) and an average number of conformers per protein of

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

~20 (minimum=6, maximum=60). In this dataset, disordered regions longer than 5 residues have an average length of 23.6 amino acids, with a maximum of 263 amino acids.

Estimation of evolutionary rates

We fed the program Rate4Site [50] with the alignments described above to obtain site-specific evolutionary rates for each protein. The raw (unnormalized) scores from Rate4Site were normalized for each protein separately by dividing by the protein average, following [51]. Wilcoxon rank sum tests were used to assess the statistical significance of the differences in the means of normalized evolutionary rates between ordered and disordered subsets.

Structural information of protein conformers and ensembles

The contact information for each protein conformer was derived as the absolute number of tertiary contacts per position (defined as the minimum distances between the Van der Waals spheres of any two side-chain or α-carbon atoms belonging to any two amino acids, employing a cutoff distance of 1.0 Å).

Contact information for protein ensembles was estimated in several ways. Average, mode, maximum and minimum number of contacts per position were calculated over the same position in all (or a subset) of the conformers in a given ensemble. Contact robustness per position was defined as the fraction of conformers in the ensemble that establish at least one contact in that position. For example, a value of 0.5 means that half of the observed conformers have at least one contact in that position.

For each disorder position, the distance to an order position was calculated as the minimum consecutive distance to the nearest ordered residue.

Per-residue flexibility was estimated using the RSMF of the α-carbons in the ensemble as calculated by MIToS [52]. RMSF values were normalized by dividing them by the protein average.

Calculation of evolutionary rate profiles and conformational patterns

For every protein in our dataset we estimated its evolutionary rate and RMSF profiles per position. As proteins have different lengths, to make proper comparisons between them, positions were renumbered to the range [0,1] by dividing by total protein length. We manually checked the order/disorder arrangement of structural elements in each protein to define its structural organization pattern. LOESS regression smoothing was performed to model the behaviour and variability of the data (normalized rates or normalized RMSF values) among selected proteins sharing the same structure-function relationship (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Analysis of correlations for single conformers and for whole ensembles

For each protein conformer, we calculated the Pearson correlation coefficient *ρ* between the normalized evolutionary rates and the number of contacts per position. We repeated this on an ensemble level by deriving the Pearson correlations between rates and the contact average or contact robustness per position among all conformers of the protein. We then determined which description of contacts was better related with the evolutionary rates of each protein by comparing their correlation values.

Analysis of correlations for combinations of subsets of conformers

For each protein and its corresponding ensemble we calculated all the possible combinations between its conformers using the formula $n!/k!\,(n-k)!$ where *k* is the number of combined conformers and *n* is the total number of conformers for each protein. For each of these combinations, we calculated the Pearson correlation coefficient *ρ* against the evolutionary rates in two ways: by using the average of residue contacts (ρ_{ave}) and by using the contact robustness (ρ_{crob}). For each combination and protein, we compared each obtained ρ with its bootstrapped distribution (10,000 resamplings) to assess if each obtained ρ is significantly different from a random distribution using a t-test (significance level 0.01).

Data and code availability

The data and code used in this manuscript are publicly available at https://gitlab.com/sbgunq/publications/palopoli-marchetti-2020-natcomm

Author contributions

NP, JM, AMM, DJZ and GP designed and performed experiments, analysed data and wrote the paper; MSF and SCET analysed data and wrote the paper.

Acknowledgements

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

NP, MSF and GP are researchers and JM is a PhD fellow from CONICET. AMM is funded by the research programme “MSCA Seal of Excellence @UniPD”. DJZ is funded by the Agence Nationale de la Recherche (ANR-17-CE12-0009-MASSIV). This work was supported by Universidad Nacional de Quilmes (PUNQ 1004/11), ANPCyT (PICT-2014-3430) and the European Union’s Horizon 2020 Research and Innovation Programme (Grant Agreement Nº 778247). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Figures

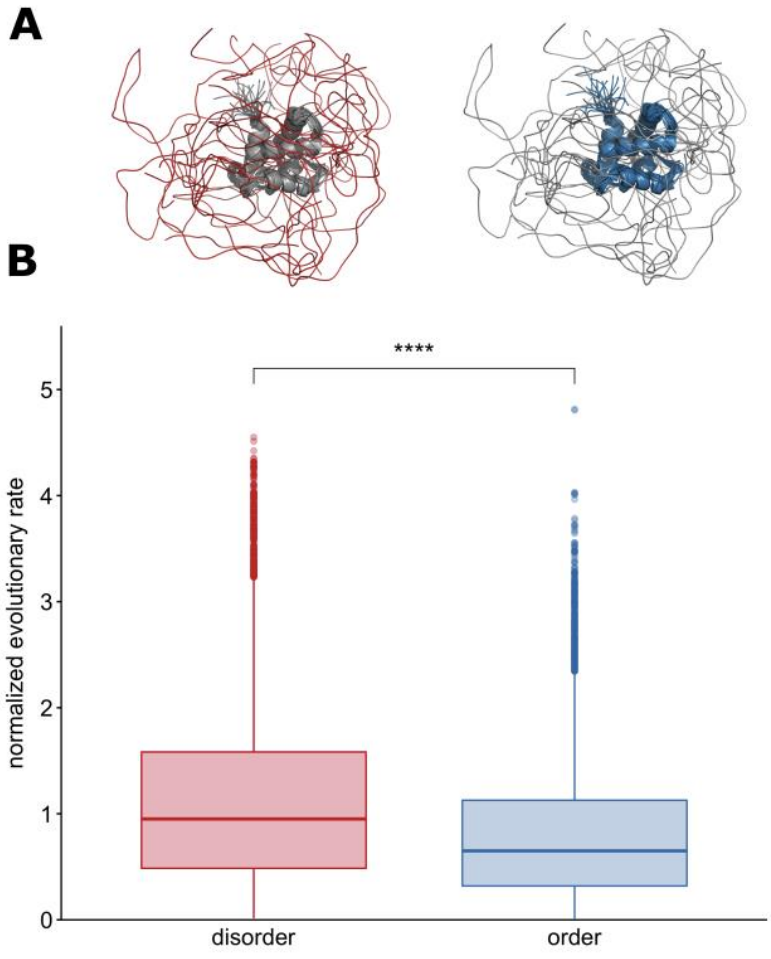
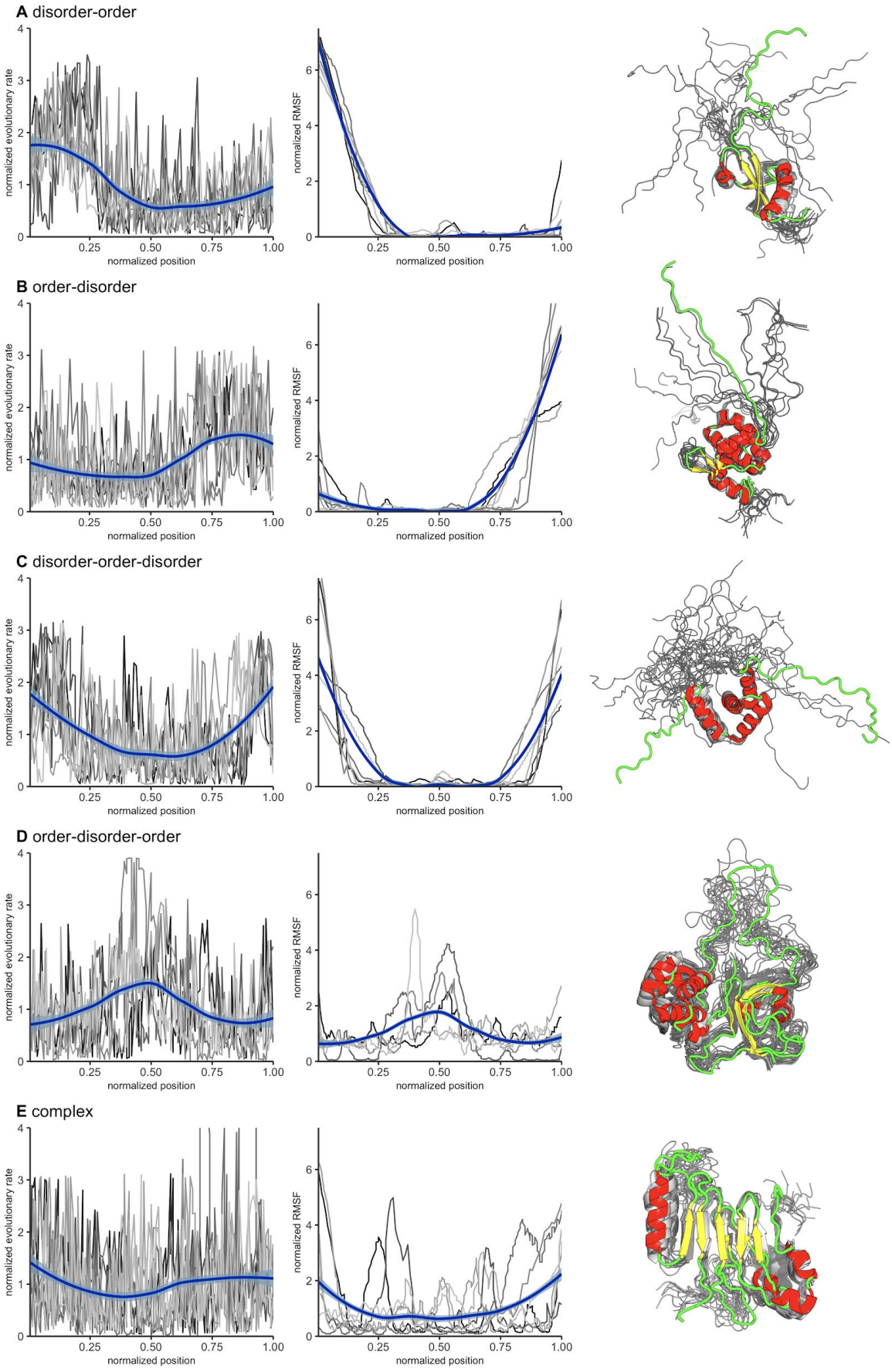


Figure 1. **Panel A:** An example of the proteins included in our dataset. Both protein figures show the ensemble of tubulin polymerization-promoting protein family member 3 (TPPP3) from *Homo sapiens* (PDB: 2jrf), a typical IDP. Figure on the left highlights disorder regions of TPPP3 in red;

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

such regions are included in the 'disorder' dataset. Figure on the right shows the TPPP3 globular domain in blue, positions that are part of the 'order' dataset. Proteins are represented using Pymol. **Panel B:** Boxplots showing distributions of normalized evolutionary rate per position for disordered positions of IDPs (red, median = 0.95) and for ordered positions of IDPs (blue, median = 0.65). Asterisks represent the statistical significance of the difference between a given pair of datasets (Wilcoxon rank sum test; **** indicates p <= 0.0001; ***: p<= 0.001; **: p<= 0.01; *: p <= 0.05; ns: p > 0.05).

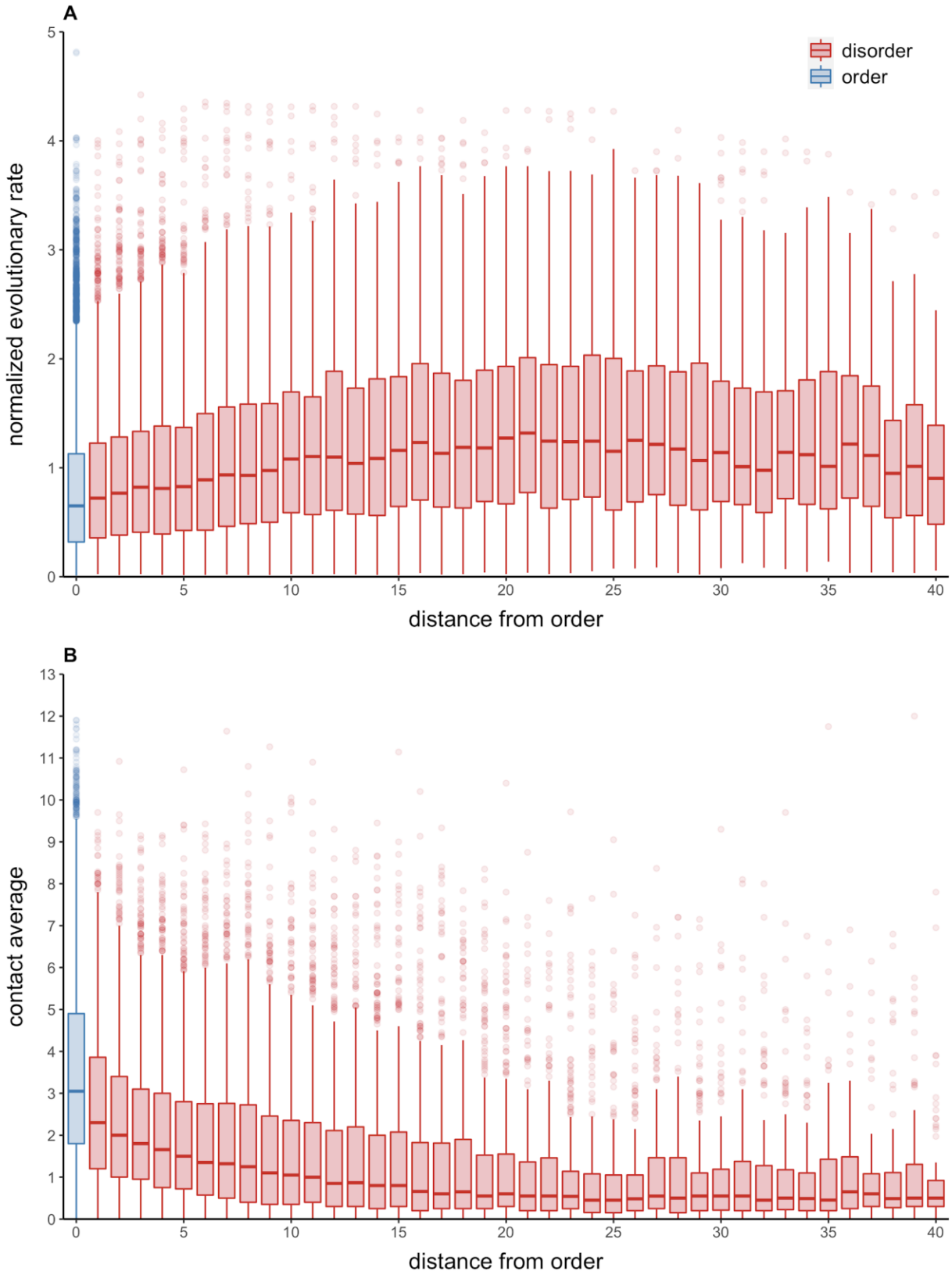
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 2. Normalized evolutionary rate profiles (left), normalized RMSF profiles (center) and cartoon image of a representative protein ensemble (right) for different structure-function patterns. Grey profiles correspond to selected protein chains from each subset. Their PDB identifiers are given below, with an asterisk marking proteins represented as cartoons (one NMR model colored by secondary structure, others in grey). Blue lines represent smooth curves obtained by LOESS regression. **Panel A**: Protein chains with a disorder-order structural pattern (PDBs: 2cqrA, 2k3aA, 2khmA, 2myxA, 2rsmA, 2rsoA, *2rvlA). **Panel B**: Proteins with an order-disorder pattern (PDBs: 1uffA, 1ug1A, 1jvrA, *2hmxA, 2k4kA, 2k5fA, 2ru8A). **Panel C**: Proteins with a disorder-order-disorder pattern (PDBs: 1ixdA, *1u6fA, 1wj7A, 2dh7A, 2ecbA, 2ejmA, 2k8pA). **Panel D**: Proteins with an order-disorder-order pattern (PDBs: 1qu6A, 2adzA, 2lbcA, 2l3sA, *2mphA). **Panel E**: Proteins with complex structural patterns (PDBs: 1mm4A, 1sm7A, 1tteA, 1vzsA, *2aivA, 2afjA, 2n1rA).

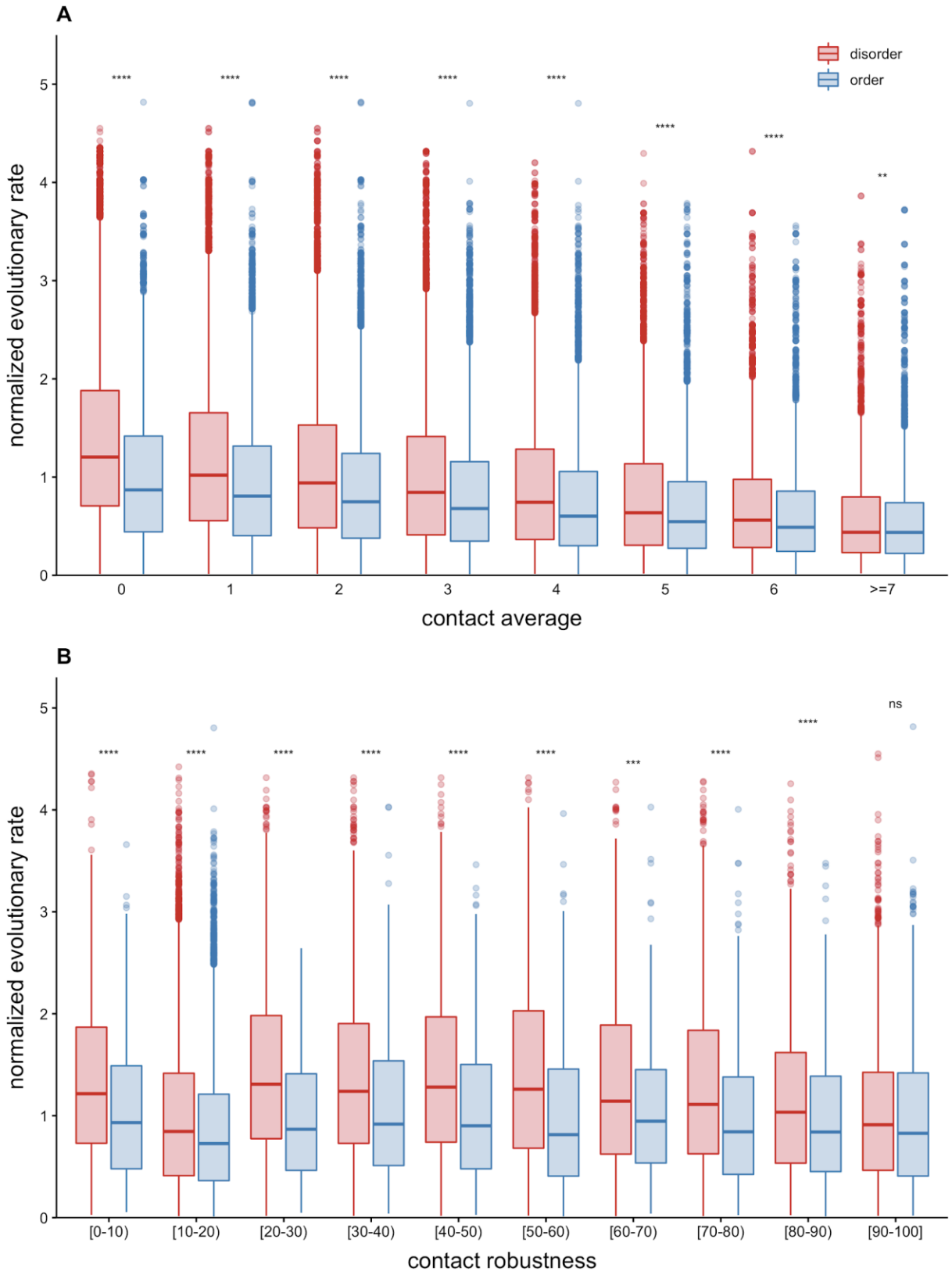
bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 3. Panel A: Boxplots showing distribution of normalized evolutionary rates per position as function of sequence distance from the nearest ordered residue. **Panel B:** Boxplots showing distribution of average number of contacts as function of sequence distance from the nearest ordered residue.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4. Panel A: Boxplots showing distribution of normalized evolutionary rates as function of the average number of contacts per position along the complete ensemble. **Panel B:** Boxplots showing distribution of normalized evolutionary rates as function of contact robustness. Asterisks represent the statistical significance of the difference between a given pair of datasets (Wilcoxon rank sum test; **** indicates p <= 0.0001; ***: p<= 0.001; **: p<= 0.01; *: p <= 0.05; ns: p > 0.05).

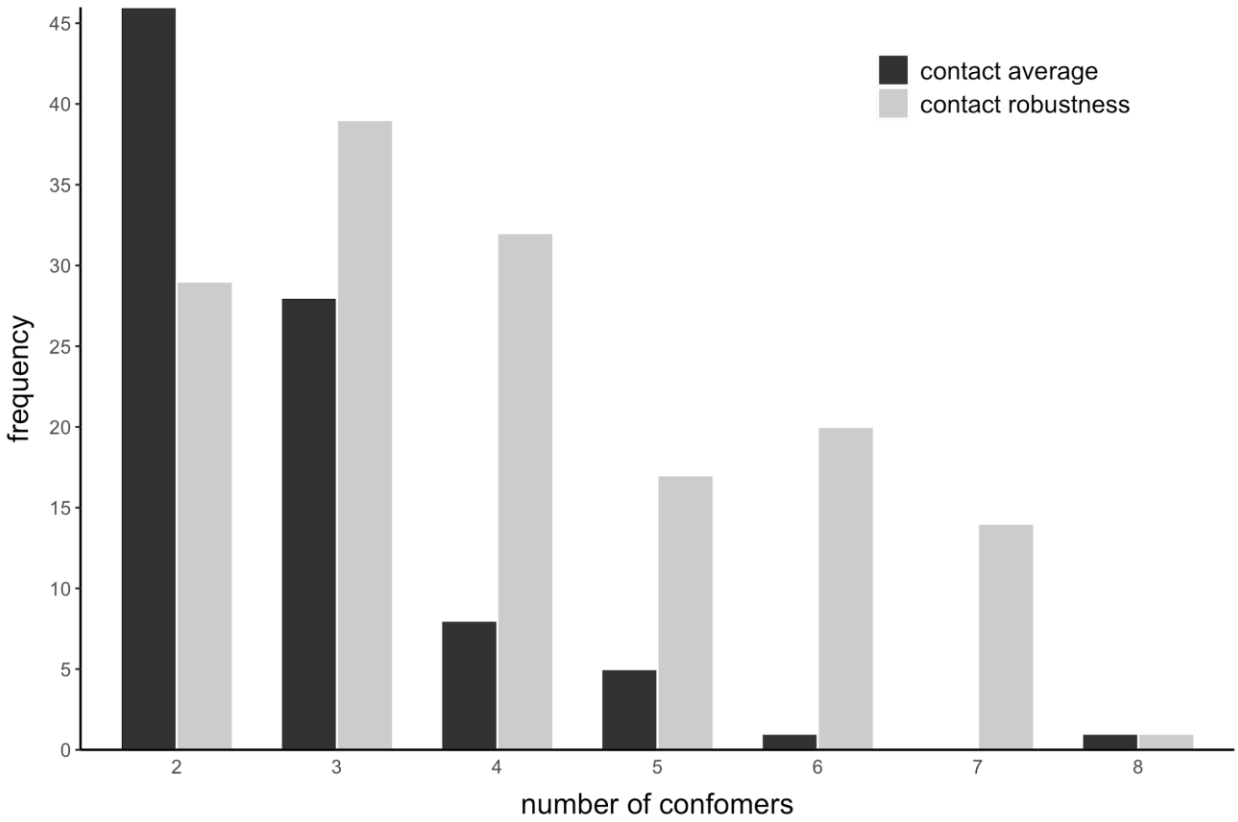


Figure 5. Histogram showing number of different conformers needed to find the best correlation between evolutionary rates and contact robustness (grey) or the average number of contacts (black).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

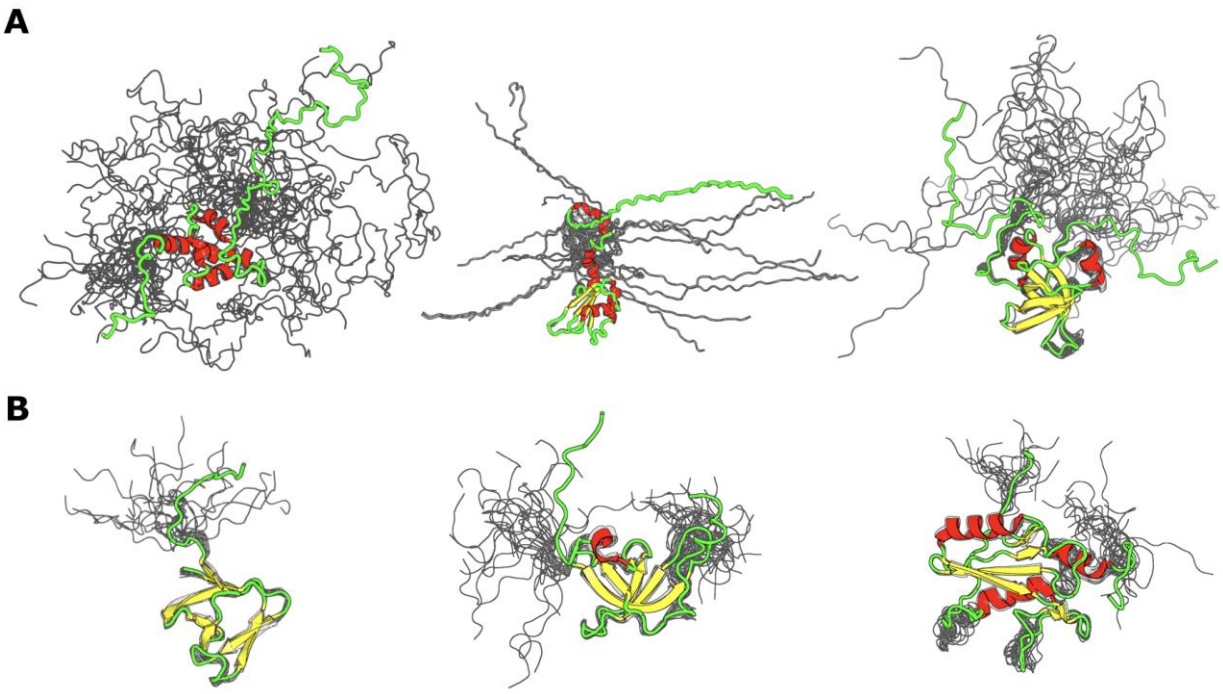


Figure 6. Panel A: Examples of IDP ensembles for which contact robustness provides the best correlation with evolutionary rates (from left to right, PDBs: 1g9lA, 2m70A, 1d7qA). **Panel B:** Examples of IDP ensembles for which contact average correlates best with evolutionary rates (from left to right, PDBs: 2krsA, 2budA, 2eo6A)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bibliography

1. Wei, G., Xi, W., Nussinov, R. & Ma, B. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chem. Rev.* **116,** 6516–6551 (2016).
2. Monzon, A. M., Rohr, C. O., Fornasari, M. S. & Parisi, G. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database (Oxford)* **2016,** (2016).
3. Monzon, A. M. *et al.* Conformational diversity analysis reveals three functional mechanisms in proteins. *PLoS Comput. Biol.* **13,** e1005398 (2017).
4. Gerstein, M. & Krebs, W. A database of macromolecular motions. *Nucleic Acids Res.* **26,** 4280–4290 (1998).
5. Berlow, R. B., Dyson, H. J. & Wright, P. E. Functional advantages of dynamic protein disorder. *FEBS Lett.* **589,** 2433–2440 (2015).
6. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27,** 527–533 (2002).
7. Siltberg-Liberles, J. Evolution of structurally disordered proteins promotes neostructuralization. *Mol. Biol. Evol.* **28,** 59–62 (2011).
8. Lobanov, M. Y. & Galzitskaya, O. V. How common is disorder? occurrence of disordered residues in four domains of life. *Int. J. Mol. Sci.* **16,** 19490–19507 (2015).
9. Necci, M., Piovesan, D. & Tosatto, S. C. E. Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein Sci.* **25,** 2164–2174 (2016).
10. Uversky, V. N. *et al.* Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* **10 Suppl 1,** S7 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

11. Nussinov, R., Ma, B. & Tsai, C.-J. Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys. Chem.* **186,** 22–30 (2014).
12. Hammes, G. G., Chang, Y.-C. & Oas, T. G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proc Natl Acad Sci USA* **106,** 13737–13741 (2009).
13. Davey, N. E. The functional importance of structure in unstructured protein regions. *Curr. Opin. Struct. Biol.* **56,** 155–163 (2019).
14. Reisen, F., Weisel, M., Kriegl, J. M. & Schneider, G. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *J. Proteome Res.* **9,** 6498–6510 (2010).
15. Mészáros, B. *et al.* Sequential, structural and functional properties of protein complexes are defined by how folding and binding intertwine. *J. Mol. Biol.* **431,** 4408–4428 (2019).
16. Zea, D. J. *et al.* Disorder transitions and conformational diversity cooperatively modulate biological function in proteins. *Protein Sci.* **25,** 1138–1146 (2016).
17. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* **102,** 14338–14343 (2005).
18. Echave, J., Spielman, S. J. & Wilke, C. O. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17,** 109–121 (2016).
19. Yeh, S.-W. *et al.* Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res. Int.* **2014,** 572409 (2014).
20. Franzosa, E. A. & Xia, Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26,** 2387–2395 (2009).
21. Tóth-Petróczy, A. & Tawfik, D. S. Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci USA* **108,** 11151–11156 (2011).
22. Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55,** 104–110 (2002).
23. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

proteins. *J. Mol. Biol.* **347,** 827–839 (2005).

24. Toth-Petroczy, A. *et al.* Structured States of Disordered Proteins from Genomic Sequences. *Cell* **167,** 158-170.e12 (2016).
25. Marchetti, J., Monzon, A. M., Tosatto, S. C. E., Parisi, G. & Fornasari, M. S. Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution. *J. Mol. Biol.* **431,** 1298–1307 (2019).
26. Pancsa, R., Zsolyomi, F. & Tompa, P. Co-Evolution of Intrinsically Disordered Proteins with Folded Partners Witnessed by Evolutionary Couplings. *Int. J. Mol. Sci.* **19,** (2018).
27. Shimojo, H. *et al.* Extended string-like binding of the phosphorylated HP1α N-terminal tail to the lysine 9-methylated histone H3 tail. *Sci. Rep.* **6,** 22527 (2016).
28. Samsudin, F., Gan, S. K.-E. & Bond, P. J. The Structural Basis for Gag Non-Cleavage Site Mutations in Determining HIV-1 Viral Fitness. *BioRxiv* (2020). doi:10.1101/2020.07.05.188326
29. Volpon, L., D'Orso, I., Young, C. R., Frasch, A. C. & Gehring, K. NMR structural study of TcUBP1, a single RRM domain protein from Trypanosoma cruzi: contribution of a beta hairpin to RNA binding. *Biochemistry* **44,** 3708–3717 (2005).
30. Prakash, A., Shin, J., Rajan, S. & Yoon, H. S. Structural basis of nucleic acid recognition by FK506-binding protein 25 (FKBP25), a nuclear immunophilin. *Nucleic Acids Res.* **44,** 2909–2925 (2016).
31. Robinson, M. A. *et al.* Multiple conformations in the ligand-binding site of the yeast nuclear pore-targeting domain of Nup116p. *J. Biol. Chem.* **280,** 35723–35732 (2005).
32. Tompa, P. Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.* **114,** 6715–6732 (2014).
33. Fuxreiter, M. Fuzziness: linking regulation to protein dynamics. *Mol. Biosyst.* **8,** 168–177 (2012).
34. Kim, J.-Y. & Chung, H. S. Disordered proteins follow diverse transition paths as they fold and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bind to a partner. *Science* **368,** 1253–1257 (2020).

35. Varadi, M. *et al.* pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* **42,** D326-35 (2014).
36. Javier Zea, D., Miguel Monzon, A., Fornasari, M. S., Marino-Buslje, C. & Parisi, G. Protein conformational diversity correlates with evolutionary rate. *Mol. Biol. Evol.* **30,** 1500–1503 (2013).
37. Sharir-Ivry, A. & Xia, Y. The impact of native state switching on protein sequence evolution. *Mol. Biol. Evol.* **34,** 1378–1390 (2017).
38. Juritz, E., Palopoli, N., Fornasari, M. S., Fernandez-Alberti, S. & Parisi, G. Protein conformational diversity modulates sequence divergence. *Mol. Biol. Evol.* **30,** 79–87 (2013).
39. Saldaño, T. E., Monzon, A. M., Parisi, G. & Fernandez-Alberti, S. Evolutionary conserved positions define protein conformational diversity. *PLoS Comput. Biol.* **12,** e1004775 (2016).
40. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114,** 6589–6631 (2014).
41. Gouw, M. *et al.* The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res.* **46,** D428–D434 (2018).
42. Mittal, A., Holehouse, A. S., Cohan, M. C. & Pappu, R. V. Sequence-to-Conformation Relationships of Disordered Regions Tethered to Folded Domains of Proteins. *J. Mol. Biol.* **430,** 2403–2421 (2018).
43. Protter, D. S. W. *et al.* Intrinsically disordered regions can contribute promiscuous interactions to RNP granule assembly. *Cell Rep.* **22,** 1401–1412 (2018).
44. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555,** 61–66 (2018).
45. Brodsky, S. *et al.* Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Mol. Cell* (2020). doi:10.1016/j.molcel.2020.05.032
46. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.29.227363; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci USA* **110,** 13392–13397 (2013).

47. Rose, P. W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **43,** D345-56 (2015).
48. Walsh, I., Martin, A. J. M., Di Domenico, T. & Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28,** 503–509 (2012).
49. Dodge, C., Schneider, R. & Sander, C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* **26,** 313–315 (1998).
50. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1,** S71-7 (2002).
51. Sydykova, D. K. & Wilke, C. O. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. *PeerJ* **5,** e3391 (2017).
52. Zea, D. J., Anfossi, D., Nielsen, M. & Marino-Buslje, C. MIToS.jl: mutual information tools for protein sequence analysis in the Julia language. *Bioinformatics* **33,** 564–565 (2017).