A structural entropy index to analyse local conformations in intrinsically disordered proteins

Akhila Melarkode Vattekatte, Tarun Jairaj Narwani, Aline Floch, Mirjana Maljković, Soubika Bisoo, Nicolas Shinada, Agata Kranjc, Jean-Christophe Gelly, Narayanaswamy Srinivasan, Nenad Mitić, et al.

Preprint J Struct Biol. (2020) 2020:107464.

A structural entropy index to analyse local conformations in Intrinsically Disordered Proteins.

Akhila Melarkode Vattekatte[1,2,3,4], Tarun Jairaj Narwani[1,2,4], Aline Floch[2,5,6,7], Mirjana Maljković[8], Soubika Bisoo[1,2,4], Nicolas K. Shinada[1,2,4,9], Agata Kranjc[1,2,4], Jean-Christophe Gelly[1,2,4,10], Narayanaswamy Srinivasan[11], Nenad Mitić[8] & Alexandre G. de Brevern[1,2,4,10,*]

[1] Biologie Intégrée du Globule Rouge UMR_S1134, Inserm, Univ. Paris, Univ. de la Réunion, Univ. des Antilles, F-75739 Paris, France
[2] Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.
[3] Faculté des Sciences et Technologies, Saint Denis Messag, F-97715 La Réunion, France
[4] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.
[5] Etablissement Français du Sang Ile de France, Créteil, France.
[6] IMRB - INSERM U955 Team 2 « Transfusion et maladies du globule rouge », Paris Est- Créteil Univ., Créteil, France.
[7] UPEC, Université Paris Est-Créteil, Créteil, France.
[8] University of Belgrade, Faculty of Mathematics, Belgrade, Serbia.
[9] Discngine, SAS, 75012, Paris France.
[10] IBL, F-75015 Paris, France.
[11] Molecular Biophysics Unit, IISc, Bangalore, India.

Short title: IDP PBs

* Corresponding author:
Mailing address: Dr. Alexandre G. de Brevern, INSERM UMR_S 1134, DSIMB, Université Paris, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
e-mail : alexandre.debrevern@univ-paris-diderot.fr

Abstract

Sequence – structure – function paradigm has been revolutionized by the discovery of disordered regions and disordered proteins more than two decades ago. While the definition of rigidity is simple with X-ray structures, the notion of flexibility is linked to high experimental B-factors. The definition of disordered regions is more complex as in these same X-ray structures; it is associated to the position of missing residues. Thus a continuum so seems to exist between rigidity, flexibility and disorder. However, it had not been precisely described. In this study, we used an ensemble of disordered proteins (or regions) and, we applied a structural alphabet to analyse their local conformation. This structural alphabet, namely Protein Blocks, had been efficiently used to highlight rigid local domains within flexible regions and so discriminates deformability and mobility concepts. Using an entropy index derived from this structural alphabet, we underlined its interest to measure these local dynamics, and to quantify, for the first time, continuum states from rigidity to flexibility and finally disorder. We also highlight non-disordered regions in the ensemble of disordered proteins in our study.

Key words: protein structures / structural alphabet / entropy / rigidity / flexibility.

Abbreviations:

IDP: Intrinsically Disorder Protein.
NMR: Nuclear Magnetic Resonance.
PBs: Protein Blocks.
PDB: Protein Data Bank.
PED: Protein Ensemble Database.
SAXS: Small-angle X-ray scattering.

Introduction

Proteins are essential biological macromolecules involved in most of the cellular functions. The amino acid residue sequence dictates the functional three-dimensional structure of a protein. The access to the 3D structure allows apprehending at an atomistic level the biological functions and thus defines the dogma of sequence to structure extended to function (Shenoy and Jayaram, 2010). Analyses of protein 3D structures have been done extensively using regular secondary structures, namely the α-helix (or 3.6_{13} helix) and the β-sheet (Richardson, 1981), initially proposed by Pauling and Corey (Eisenberg, 2003; Pauling and Corey, 1951; Pauling et al., 1951). Nonetheless, secondary structure definition cannot be seen without ambiguity (see (Colloc'h et al., 1993; Fourrier et al., 2004; Martin et al., 2005; Tyagi et al., 2009a; Tyagi et al., 2009b)). Therefore, proposition of local protein structure libraries independent of classical secondary structures are of much interest. Amongst, structural alphabets (de Brevern, 2005; Karchin et al., 2003; Offmann et al., 2007), Protein Blocks (PBs) have shown significant applications in structural bioinformatics (de Brevern et al., 2000; Joseph et al., 2010), e.g. threading approaches (Ghouzam et al., 2015; Ghouzam et al., 2016) or protein structure superimposition and comparison (Leonard et al., 2014; Tyagi et al., 2006).

The sequence – structure - function dogma had to be redefine when it was accepted that a non-negligible part of the protein structures are not ordered and are likely to be unfolded in solution, under native, functional conditions (Mitic et al., 2018; Pavlovic-Lazetic et al., 2011; Uversky, 2002a; Uversky, 2002b; Uversky et al., 2000; Wright and Dyson, 1999). These Intrinsically Disordered Proteins (IDPs) (Habchi et al., 2014), do not have a well-defined 3-D structure but rather adopt an ensemble of conformations that are functional in solution. IDPs exist as dynamic ensembles, within which atom positions and backbone angles exhibit random temporal fluctuations (Dunker et al., 2001; Tompa, 2002). To describe these

ensembles, different experimental and computational approaches are available. In this field, Nuclear Magnetic Resonance provides quantitative residue-level information on structure and dynamics of IDPs as structural ensembles (Kragelj et al., 2015). Small-angle X-ray Scattering (SAXS) and electron microscopy (EM) gives an *in-situ* ensemble model describing the conformational behaviour of the disordered region, e.g. intrinsically disordered C-terminal domain of nucleoprotein is essential for transcription and replication of the measles virus (Jensen et al., 2011). Finally, Molecular Dynamics (MDs) is used to refine and propose such ensembles (Robustelli et al., 2018). Hence, the sequence – structure – function dogma is challenged by the fact that structure is always in a dynamic state where (a) it may acquire or lose rigid, flexible and deformable states (Carugo, 2018; Thorpe et al., 2001) and (b) is composed of ordered and disordered regions (Uversky, 2017).

The structural alphabet PBs have been used to analyse structural flexibility in Integrins (Goguet et al., 2017; Jallu et al., 2012; Jallu et al., 2014), in Duffy Antigen Chemokine Receptor protein (named now ACKR1)(de Brevern et al., 2005), impact of Post-Translational Modifications (Craveur et al., 2019), KISS1R (Chevrier et al., 2013), and in NMDA Receptor Channel Gate (Ladislav et al., 2018). Besides, it has also been used with experimental data (Schneider et al., 2014a; Schneider et al., 2014b). PBs have shown to be crucial in identifying rigid positions encompassed in flexible regions, i.e. local protein conformations that are rigid but are enclosed in deformable zones (Craveur et al., 2015).

With the understanding of flexibility in globular proteins, we extend here such analyses to the cases of IDPs from PED[3]: Protein Ensemble Database (Varadi et al., 2014). This dataset was assembled by specialists from various groups in the aim to have an open access database for structural information on IDPs and denatured protein ensembles based on NMR, SAXS data and MD simulations. Thus, it perfectly fits to see if some regions of these IDPs are more rigid than expected, as seen in the case of the multisite phosphorylation of some kinases

(Xiang et al., 2013).

Materials and Methods

Data sets. All the IDP structures were taken from PED[3]: Protein Ensemble Database, the database of conformational ensembles describing flexible proteins (http://pedb.vib.be/index.php, accessed 15th May 2018) (Varadi et al., 2014). They consist of 24 entries describing different IDPs with different technique, i.e. SAXS and NMR, NMR alone, SAXS alone and Molecular Dynamics (Allison et al., 2014; Bacot-Davis et al., 2014; De Biasio et al., 2014; Marsh and Forman-Kay, 2009; Mertens et al., 2012; Mittag et al., 2010; Ozenne et al., 2012; Sanchez-Martinez and Crehuet, 2014; Sivakolundu et al., 2005; Sterckx et al., 2014; Weeks et al., 2014).

A dataset of 169 X-ray globular ordered protein structures was used to compare to these proteins. This dataset was curated from Protein Data Bank (PDB), (Berman et al., 2000) and was extracted using ASTRAL 2.03 at 40% sequence identity (Fox et al., 2014). Their structure resolutions were better than 1.5 Å, without presence of heteroatoms, missing residues, alternate or modified residues in the chain. Only globular proteins, with chain length ranging between 50 and 250 residues, were selected. Three molecular dynamic (MD) simulations were performed for each protein structure with GROMACS 4.5.7 software (Pronk et al., 2013), using AMBER99sb force field (van Gunsteren et al., 1996). Each protein structure was put in a periodic dodecahedron box, using TIP3P water molecules (Jorgensen and Madura, 1983), and neutralised with Na^+ or Cl^- counter ions. The system was then energetically minimised with a steepest-descent algorithm for 2000 steps. The MD simulations were performed in isothermal-isobaric thermodynamics ensemble (NPT), with temperature fixed at 300 K and pressure at 1 bar. A short run of 1ns was performed to equilibrate the system, using Berendsen algorithm for temperature and pressure control

(Berendsen et al., 1984). The coupling time constants were equal to 0.1 ps for each physical parameter. Then, a production step of 50 ns was done using Parrinello-Rahman algorithm (Parrinello and Rahman, 1981) for temperature and pressure control, with coupling constants of T=0.1 ps and P=4 ps. All bond lengths were constrained with LINCS algorithm (Hess et al., 1997), which allowed an integration step of 2 fs. The PME algorithm (Darden et al., 1999) was used for long-range electrostatic interactions using a cut-off of 1 nm for non-bonded interactions.

This protocol was applied on each of the 169 protein chains. Conformations were saved every 10 ps. For each MD simulation, the secondary structures were analysed and the structural deviation of each snapshot from the initial structure was measured. Trajectory analyses were done with the GROMACS software, in-house Python and R scripts. Root mean square deviations (RMSD) and root mean square fluctuations (RMSf) were computed on Cα atoms. Normalized RMSfs and normalized B-factors were computed as in Bornot et al. study (Bornot et al., 2011).

Protein Blocks. Protein Blocks is the most widely used structural alphabet composed of 16 local prototypes (de Brevern et al., 2000), it is employed to analyse local conformations of protein structures from the Protein Data Bank (PDB) (Berman et al., 2000). Each PB is characterized by the φ and ψ dihedral angles of five consecutive residues. PBs give a reasonable approximation of all local protein 3D structures (Joseph et al., 2010) and are very efficient in tasks such as protein superimpositions (Gelly et al., 2011; Leonard et al., 2014) and Molecular Dynamics (MDs) analyses (Goguet et al., 2017; Jallu et al., 2014; Ladislav et al., 2018). PB assignment was carried out for every residue from every structure / structural model extracted from PED[3] using our PBxplore tool (Barnoud et al., 2017), available at GitHub (https://github.com/pierrepo/PBxplore). From this description, we have used a

recognized measure that helps in quantifying the flexibility of each amino acid called N_{eq} (N_{eq} means for equivalent Number of PBs) (de Brevern et al., 2000). N_{eq} is a statistical measurement similar to entropy, and represents the average number of PBs a residue may adopt at a given position. N_{eq} is calculated as follows (de Brevern et al., 2000):

$$Neq = exp(-\sum_{x=1}^{16} f_x \ln f_x) \qquad (1)$$

Where, f_x is the frequency of PB x at the position of interest. N_{eq} value can vary between 1 and 16. A N_{eq} value of 1 indicates that only one type of PB is observed, while a value of 16 indicates an equal probability for each of the 16 PBs, *i.e.* random distribution. It must be noticed that N_{eq} was originally proposed for prediction purpose, for which random frequencies of 1/16 for each PB can be seen. As the observed distribution of PBs (see (de Brevern, 2005)) ranges between 0.8% to more than 30%, the expected maximum N_{eq} is around 10 and not 16 in ordered proteins. IDPs do not follow the same expected frequency distribution and so N_{eq} higher than 10 can be expected. PBxplore tool allows the assignment of PBs, the calculation of N_{eq} and the creation of sequence logos of PBs (WebLogo) (Crooks et al., 2004).

Interest of Protein Blocks to distinguish mobility from flexibility. Analysis of flexibility can be done using experimental B-factors taken from X-ray structures or Root Mean Square fluctuation (RMSf) extracted from Molecular Dynamics simulations. These measures provide a continuum from rigid (low RMSf or B-factors) to flexible (high RMSf or B-factors). However, it has a limitation when a mobile element (i.e. rigid) is encompassed in a flexible region. The B-factors and RMSf values are high but the protein region is not flexible at all. RMSf is a Euclidean distance of all different conformations of the proteins observed during the dynamics. It is based on (i) the selection of a representative median conformation,

and then (ii) the superimposition of all snapshots on this representative conformation. A high RMSf reflects a large average distance to this barycentre conformation, and so not always a high flexibility. With PBs, the N_{eq} values are at these positions low, underlying the fact that the region is rigid (i.e. in this case even can be noted as mobile). This approach allows a simple view of these cases. This approach allows straightforward analyses and interpretations of these cases. Moreover, N_{eq} values are comparable between proteins when it is slightly more complex for RMSf and B-factors.

Analyses. The analyses were done using Python programming language v.2.7.10 (Foundation), and R software v.3.3.3 (Team, 2017) while 3D visualisation was done using MacPyMOL software v.1.7.2.2 (DeLano, 2002; Schrodinger, 2015).

Results

Analyses of local conformations in the IDP dataset vs Ordered dataset. As mentioned on the website, PED[3] contains 25,473 protein structures of 60 ensembles in 24 entries. Out of these, 6 entries have data from both SAXS and NMR, 7 from only SAXS, 10 from only NMR and one from Molecular Dynamics.

The entries were analysed separately as single chains and Protein Blocks were assigned to each protein structure (de Brevern et al., 2000) using PBxplore tool (Barnoud et al., 2017). As mentioned in the Methods section, it is possible to compute an entropy index, namely N_{eq}, from the PB distribution; N_{eq} ranges between 1 (only one PB seen at a defined position) and 16 (all 16 PBs are seen with equal frequency, depicting random distribution). A similar analysis was performed on 169 globular proteins for which 3 independent molecular dynamics simulations of 150 ns were performed (Narwani et al., 2018). Figure 1 shows the two magnified N_{eq} distributions of $N_{eq} > 1.0$ with an inset distribution plot showing whole

statistics (see also Supplementary Figure 1). Indeed, the observations are mainly directed by rigid local conformations, i.e. N_{eq} value of 1.0, that correspond to 60% of the residues from ordered globular protein dataset (Narwani et al., 2018) and 58% from IDP dataset (see Sup Table 2). The tendencies change drastically for $N_{eq} > 2.0$ as only 8% of the residues of globular protein dataset (see Figure 1A) and 36% of the residues of IDP dataset fall into this group (see Figure 1A); almost none of the residues in globular protein dataset (i.e. 0.82%) have $N_{eq} > 4.0$, and none have $N_{eq} > 8.0$, while 31% of IDP residues have $N_{eq} > 4.0$ and 15% have $N_{eq} > 8.0$. From this it is evident that the entropy index N_{eq} is appropriate related to the question of proteins' order and disorder as hypothesized previously (Craveur et al., 2015).

We would like to stress that N_{eq} is a particular measure, a local one. Indeed, as we have shown in (Narwani et al., 2019), on the ordered dataset the correlation between normalized B-factor and normalized RMSf is of 0.43, a similar value is also as seen in our previous study and many others (Bornot et al., 2009). Correlation with N_{eq}, is only of 0.24 and 0.14. These low values are also expected since N_{eq} does a local conformation analyses with more than 60% of residues having a N_{eq} of 1.0, meaning that single local protein conformation remains unchanged during MD simulations time. A very low N_{eq} of 1.0 can be found associated with high B-factor and RMSf as seen in (Craveur et al., 2015; Goguet et al., 2017). It defines then a mobile element, i.e. a rigid region encompassed between two deformable regions, which act as hinges.

Influence of data on IDP dataset. As expected, there is a huge variation among the N_{eq} values of IDPs from PED[3]. Average N_{eq} for entry is equal to 4.6, where 11 entries have N_{eq} higher than 6.2 and 11 entries have N_{eq} lower than 2.2 (see Figure 2A). The protein chains in the disordered dataset have an average length of 110 residues and does not influence the entropy values (see Figure 2B), as the longer protein sequences have an average N_{eq} of 4.1

while the shortest having 4.9. The number of models for each PED[3] entries has a slight impact on the N_{eq} values as seen in Figure 2C. It is thus implied that the number of occurrences influences N_{eq} values. Nonetheless, the relationship is not so straightforward, as with 130 models, the unbound p27 KID domain (PED2AA) has an average N_{eq} of 1.7 (Sivakolundu et al., 2005), while with 32 occurrences, the structural ensemble of phosphorylated Sic1 (PED1AAA) has an average N_{eq} of 7.6 (Mittag et al., 2010). An enlarged inset plot of Fig 2.C is shown in the Supplementary Figure 2, as the number of models can be substantial (e.g. 13718 for Sendai virus phosphoprotein ensemble (PED4AAB) (Sanchez-Martinez and Crehuet, 2014)) and in this case, N_{eq} values are high. The 8 entries with more than 500 models have an average N_{eq} value of 7.8 while the others have an average N_{eq} of 3.1.

Does the IDP dataset contain only IDPs? The above analyses might seem to categorize the proteins from PED[3] database. However, it is in fact more complex: as seen in the Figure 3 (see also Supplementary Figure 3), some entries are completely disordered, while others are rigid and third are intermediate cases.

While it is simple to say that a N_{eq} of 1 is associated to a rigid residue and a N_{eq} of 12 is in a disordered region, it is more difficult to classify the precise delineation of flexibility and disorder as these terms are too definitive for the characterization of a continuous value. We have so decided through a visual inspection of every entry (see Sup Data 1), that a N_{eq} higher than 8 must to be considered as completely disordered, while by looking at our multiple previous molecular dynamic simulations (Craveur et al., 2015; Goguet et al., 2017; Jallu et al., 2012; Jallu et al., 2014; Narwani et al., 2018; Narwani et al., 2019), we found that N_{eq} close to 4 can be considered as flexible. By extending this simple approach, a N_{eq} of 6 can be an intermediate state between highly flexible and disorder. The difficulty to precisely define without ambiguity a defined threshold was always critical in analysis of flexibility. For

instance, Schlessinger and Rost decided to use a strict threshold between rigid and flexible region, but also have proposed a non-strict threshold with a different value (Schlessinger et al., 2006). Our threshold values are so provided as them, but can be discussed in regards to the analysed proteins. Coarsely, we can classify the 24 entries into 4 groups (see Table 1): (i) completely disordered (8 entries), (ii) partially disordered (7 entries), (iii) mainly/largely flexible (5 entries), and (iv) composed of some rigid regions (4 entries).

Some completely disordered proteins have an impressive N_{eq} values of > 7.5, such as structural ensemble of phosphorylated Sic1 (N_{eq} 7.6) (Mittag et al., 2010) (see Figure 3A) and Sendai virus phosphoprotein ensemble (N_{eq} 8.6) (Sanchez-Martinez and Crehuet, 2014) (see Figure 3B). Sometimes, some positions are observed to be more rigid than expected, see position 102 of the measles nucleoprotein (Ozenne et al., 2012) with N_{eq} of 1.2, i.e. rigid, while average N_{eq} is characteristic of disorder about 8.4 (see Figure 3C). The latter example corresponds to a disorder protein with a rigid region. This mobile region is helical (see Sup Data 1). A similar observation was found for a solution-state ensemble of a β-synuclein (Allison et al., 2014). The average N_{eq} is of 6.3, leading to its characterization as a completely disordered protein. However, some regions in the C-terminus are below 4, especially near position 120 with a N_{eq} of 1, a highly rigid position for a helical structure (see Figure 3D), which is quite interesting for a terminal region. It is more striking, in case of a punctual variant fragment of heat shock protein β6 (Weeks et al., 2014), which is characterized by a rigid central region formed by residue position 20 to 95 enclosed by disordered termini (see Figure 3E). This example reflects most of the members of the second group of partial disordered proteins (see Table 1). It also corroborates with other studies in the field. Indeed, half of the protein structures lack the electron density data at their extremities due to high thermal displacement (Faure et al., 2008; Faure et al., 2009; Jacob and Unger, 2007).

The third group can be considered as flexible and not disordered as they have a lower

average N_{eq} representing some rigid regions. The phosphorylated Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase (entry PED5AAD (Bacot-Davis et al., 2014)) is a perfect example with an average N_{eq} of only 1.8, a maximum N_{eq} of only 4 and with five regions entirely rigid (see Figure 3F). The last group encompasses two rigid protein entries both very interesting cases: (i) singly phosphorylated mengovirus leader protein (PED4AAD, see Figure 3G (Bacot-Davis et al., 2014)), and (ii) the unphosphorylated mengovirus leader protein (PED3AAD, see Figure 3H (Bacot-Davis et al., 2014)). They have some rigid regions with low N_{eq} but also flexible regions and disordered regions with higher N_{eq}.

Discussion

IDPs and IDRs are dynamic ensembles (Dunker et al., 2001; Habchi et al., 2014; Tompa, 2002; Uversky et al., 2000). They have been widely analysed for the notions of rigidity – flexibility and order – disorder. The transition can be seen as a continuum from rigidity to flexibility to disorder. However, this continuum may mask interspersed rigid regions entrapped in highly flexible or disordered region, namely mobile regions as shown by (Craveur et al., 2015; Goguet et al., 2017) in/for the case of Calf-1 domain of integrin $\alpha_{IIb}\beta_3$.

In this study, we have shown that the use of our structural alphabet (de Brevern et al., 2000; Offmann et al., 2007) and an entropy index allows a precise analysis of flexibility and can be extended to identify disorder. PB analysis of ensembles from PED[3] database (see Sup Table 3) enables us to propose gradient between disorder (a very large number of local backbone conformations being sampled, i.e. N_{eq} of 8) and flexibility (fairly limited number of local protein conformations, i.e. N_{eq} of 4, see Table 1 and Sup Data 1). Moreover, it is possible to efficiently localize a mobile region in IDPs, as for example PED7AAC (see Figure 3C (Ozenne et al., 2012)) or PED6AAC, an ensemble of conformations of tau protein that has some extended structure property and one position with N_{eq} of 2 when the average N_{eq} of the

protein is 8.1.

Even with the limited number of examples; N_{eq} entropy is able to clearly distinguish backbone conformations, allowing the identification of mobile regions inside disordered regions. Thus, providing a metric not only to differentiate between flexible and disordered region but also to detect rigid regions encompassed by flexible regions. This work could be extended further owing to the physiological significance of the mobile regions in the IDPs and the biological implications of the different groups. In perspectives, we would like to go further on larger datasets of IDPS and IDRs, using NMR data, and combine it with molecular simulations. A very interesting approach would be to add information from NMR data using IDR assignment as proposed in (Ota et al., 2013). Another question will be to select proper IDRs in X-ray structures. Indeed, Godzik's group (Zhang et al., 2007) underline the existence of Dual Personality Fragments (DPFs), i.e. fragments that are found both in disordered IDRs and ordered regions. The selection of IDRs must be done with certainty and should not be biased by DPFs.

Acknowledgments

We would like to thank Charlotte Périn and Professor Catherine Etchebest as the organizers and participants of Belbi'2016 and Belbi'2018 for fruitful discussions. This work was supported by grants from the Ministry of Research (France), University de Paris, University Paris Diderot, Sorbonne, Paris Cité (France), National Institute for Blood Transfusion (INTS, France), National Institute for Health and Medical Research (INSERM, France), IdEx ANR-18-IDEX-0001 and labex GR-Ex. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02. TJN, NS and AdB acknowledge to Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grant (number 5302-2). NSh acknowledges support from ANRT. AMV is supported by Allocation de Recherche Réunion granted by the Conseil Régional de la Réunion and the European Social Fund EU (ESF). MM and NM acknowledge to project grants No. 174021 and 44006 from Ministry of Education, Science and Technological Development, Republic of Serbia

The authors were granted access to high performance computing (HPC) resources at the French National Computing Centre CINES under grant no. c2013037147, no. A0010707621 and A0040710426 funded by the GENCI (Grand Equipement National de Calcul Intensif). Calculations were also performed on an SGI cluster granted by Conseil Régional Ile de France and INTS (SESAME Grant).

Conflict of interest

The authors have no conflict of interest to declare. JCG and ADB are associated with IBL, Paris, France.

Legends

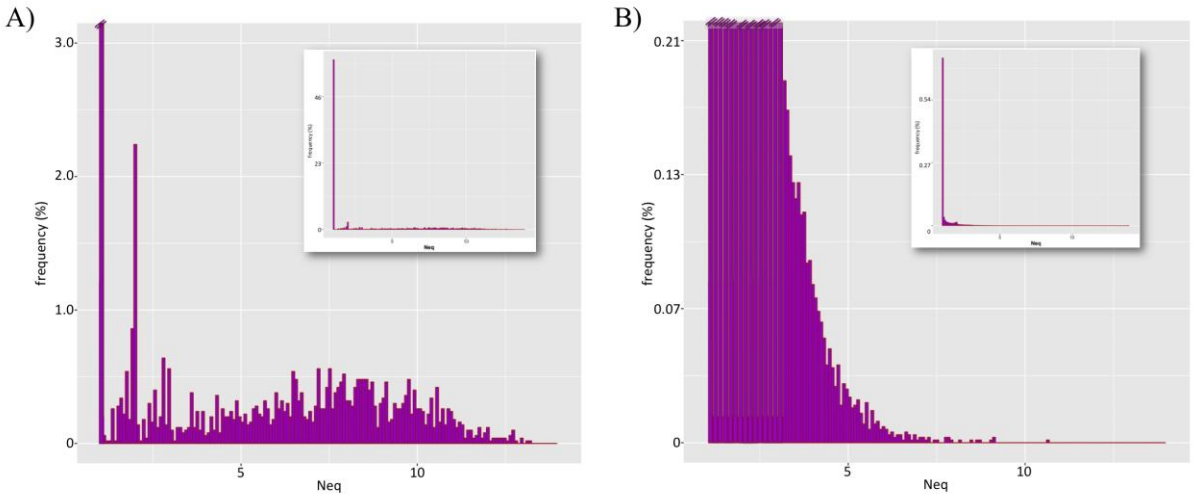

Figure 1. *Global probability distribution*. N_{eq} values for A) IDP proteins (from PED[3] database) and B) globular ordered proteins (from the ordered dataset). For clarity the main plots represent frequencies less than 3% for IDP and 0.2% for ordered dataset. The inset plot in both represents the global distribution for all N_{eq} values.

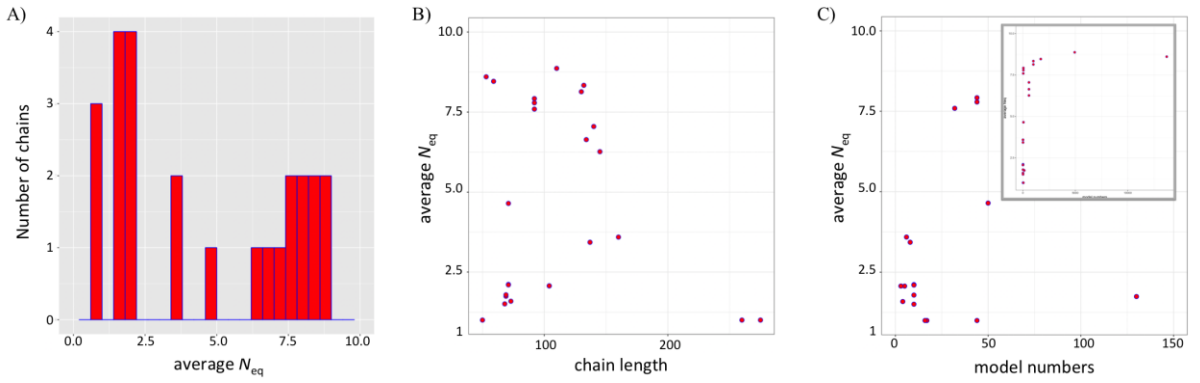

Figure 2. *Conformational entropy analysis*. A) average N_{eq} per entry, B) average N_{eq} versus chain length and C) average N_{eq} versus number of models (for clarity, the main plot represents number less than 200, Sup Fig 2 show all the distribution).

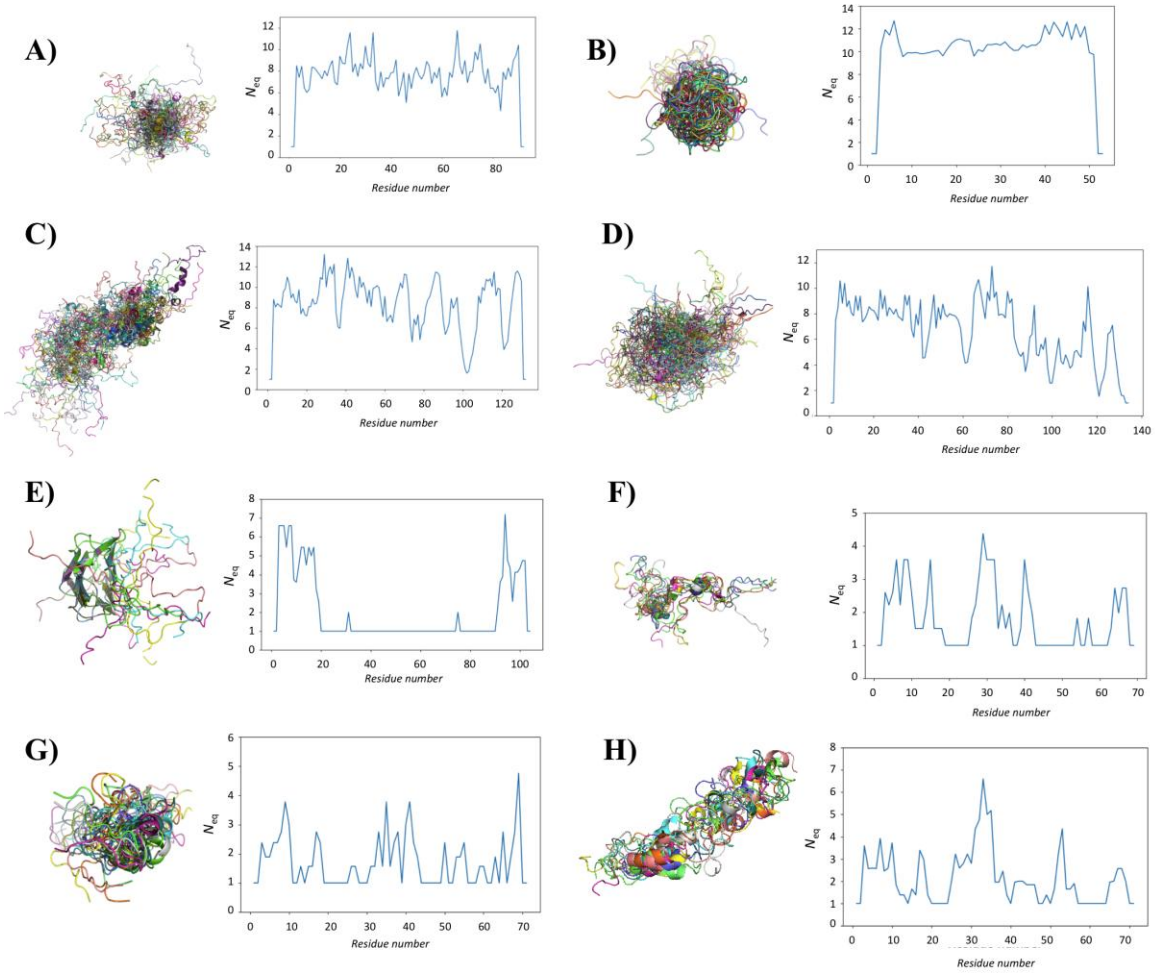

Figure 3. *3D visualisation and N_{eq} distribution of some characteristic entries*. A) PED1AAA - structural ensemble of phosphorylated Sic1 (Mittag et al., 2010), B) PED4AAB - Sendai virus phosphoprotein ensemble (Sanchez-Martinez and Crehuet, 2014), C) PED7AAC - Ensemble of the free form measles nucleoprotein (Ozenne et al., 2012), D) PED1AAD - β-synuclein: solution-state ensemble (Allison et al., 2014), E) PED1AAB - punctual variant fragment of Heat shock protein β-6 (Weeks et al., 2014), F) PED5AAD - phosphorylated Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase (Bacot-Davis et al., 2014), G) PED4AAD - singly phosphorylated mengovirus leader protein (Bacot-Davis et al., 2014), and H) PED3AAD - unphosphorylated mengovirus leader protein (Bacot-Davis et al., 2014).

Completely disordered	PED1AAA, PED4AAA, PED6AAA, PED6AAC[1], PED7AAC[1], PED8AAC, PED9AAA, PED9AAC[1]
partially disordered	PED1AAB[2], PED1AAD, PED2AAA[2], PED2AAB[2], PED2AAD, PED7AAA[2], PED8AAA[2]
flexible	PED3AAB[3], PED4AAB[3], PED5AAC, PED5AAD[3], PED6AAD[3]
rigid	PED3AAA, PED3AAD[4], PED4AAA, PED5AAA[4]

[1]: few constrained positions
[2]: mainly disorder Nt & Ct
[3]: some rigid regions
[4]: rigid, flexible and disordered regions

Table 1. *Classification of PED*[3] *entries*. Classified into 4 groups based on conformational entropy index N_{eq}.

References

Allison, J.R., Rivers, R.C., Christodoulou, J.C., Vendruscolo, M., Dobson, C.M., 2014. A relationship between the transient structure in the monomeric state and the aggregation propensities of alpha-synuclein and beta-synuclein. Biochemistry 53, 7170-7183.

Bacot-Davis, V.R., Ciomperlik, J.J., Basta, H.A., Cornilescu, C.C., Palmenberg, A.C., 2014. Solution structures of Mengovirus Leader protein, its phosphorylated derivatives, and in complex with nuclear transport regulatory protein, RanGTPase. Proc Natl Acad Sci U S A 111, 15792-15797.

Barnoud, J., Santuz, H., Craveur, P., Joseph, A.P., Jallu, V., de Brevern, A.G., Poulain, P., 2017. PBxplore: a tool to analyze local protein structure and deformability with Protein Blocks. PeerJ 5, e4013.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., Haak, J.R., 1984. Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 81, 3684-3690.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235-242.

Bornot, A., Etchebest, C., de Brevern, A.G., 2009. A new prediction strategy for long local protein structures using an original description. Proteins 76, 570-587.

Bornot, A., Etchebest, C., de Brevern, A.G., 2011. Predicting protein flexibility through the prediction of local structures. Proteins 79, 839-852.

Carugo, O., 2018. Atomic displacement parameters in structural biology. Amino Acids 50, 775-786.

Chevrier, L., de Brevern, A., Hernandez, E., Leprince, J., Vaudry, H., Guedj, A.M., de Roux, N., 2013. PRR repeats in the intracellular domain of KISS1R are important for its export to cell membrane. Mol Endocrinol 27, 1004-1014.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.P., 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 6, 377-382.

Craveur, P., Narwani, T.J., Rebehmed, J., de Brevern, A.G., 2019. Investigation of the impact of PTMs on the protein backbone conformation. Amino Acids.

Craveur, P., Joseph, A.P., Esque, J., Narwani, T.J., Noel, F., Shinada, N., Goguet, M., Leonard, S., Poulain, P., Bertrand, O., Faure, G., Rebehmed, J., Ghozlane, A., Swapna, L.S., Bhaskara, R.M., Barnoud, J., Teletchea, S., Jallu, V., Cerny, J., Schneider, B., Etchebest, C., Srinivasan, N., Gelly, J.C., de Brevern, A.G., 2015. Protein flexibility in the light of structural alphabets. Front Mol Biosci 2, 20.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res 14, 1188-1190.

Darden, T., Perera, L., Li, L., Pedersen, L., 1999. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure 7, R55-60.

De Biasio, A., Ibanez de Opakua, A., Cordeiro, T.N., Villate, M., Merino, N., Sibille, N., Lelli, M., Diercks, T., Bernado, P., Blanco, F.J., 2014. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. Biophys J 106, 865-874.

de Brevern, A.G., 2005. New assessment of a structural alphabet. In Silico Biol 5, 283-289.

de Brevern, A.G., Etchebest, C., Hazout, S., 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41, 271-287.

de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C., Etchebest, C., 2005. A structural model of a seven-transmembrane helix receptor: the Duffy

antigen/receptor for chemokine (DARC). Biochim Biophys Acta 1724, 288-306.
DeLano, W.L.T., 2002. The PyMOL Molecular Graphics System DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.
Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., 2001. Intrinsically disordered protein. J Mol Graph Model 19, 26-59.
Eisenberg, D., 2003. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. Proc Natl Acad Sci U S A 100, 11207-11210.
Faure, G., Bornot, A., de Brevern, A.G., 2008. Protein contacts, inter-residue interactions and side-chain modelling. Biochimie 90, 626-639.
Faure, G., Bornot, A., de Brevern, A.G., 2009. Analysis of protein contacts into Protein Units. Biochimie 91, 876-887.
Foundation, P.S., https://www.python.org/.
Fourrier, L., Benros, C., de Brevern, A.G., 2004. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 5, 58.
Fox, N.K., Brenner, S.E., Chandonia, J.M., 2014. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42, D304-309.
Gelly, J.C., Joseph, A.P., Srinivasan, N., de Brevern, A.G., 2011. iPBA: a tool for protein structure comparison using sequence alignment strategies. Nucleic Acids Res 39, W18-23.
Ghouzam, Y., Postic, G., de Brevern, A.G., Gelly, J.C., 2015. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. Bioinformatics 31, 3782-3789.
Ghouzam, Y., Postic, G., Guerin, P.E., de Brevern, A.G., Gelly, J.C., 2016. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. Sci Rep 6, 28268.
Goguet, M., Narwani, T.J., Petermann, R., Jallu, V., de Brevern, A.G., 2017. In silico analysis of Glanzmann variants of Calf-1 domain of alphaIIbbeta3 integrin revealed dynamic allosteric effect. Sci Rep 7, 8001.
Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., 2014. Introducing protein intrinsic disorder. Chem Rev 114, 6561-6588.
Hess, B., Bekker, H., Berendsen, H.J.C., Fraaije, J.G.E.M., 1997. LINCS: a linear constraint solver for molecular simulations. J. Comp. Chem. 18, 1463-1472.
Jacob, E., Unger, R., 2007. A tale of two tails: why are terminal residues of proteins exposed? Bioinformatics 23, e225-230.
Jallu, V., Poulain, P., Fuchs, P.F., Kaplan, C., de Brevern, A.G., 2012. Modeling and molecular dynamics of HPA-1a and -1b polymorphisms: effects on the structure of the beta3 subunit of the alphaIIbbeta3 integrin. PLoS One 7, e47304.
Jallu, V., Poulain, P., Fuchs, P.F., Kaplan, C., de Brevern, A.G., 2014. Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit beta3: Structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants. Biochimie 105, 84-90.
Jensen, M.R., Communie, G., Ribeiro, E.A., Jr., Martinez, N., Desfosses, A., Salmon, L., Mollica, L., Gabel, F., Jamin, M., Longhi, S., Ruigrok, R.W., Blackledge, M., 2011. Intrinsic disorder in measles virus nucleocapsids. Proc Natl Acad Sci U S A 108, 9839-9844.
Jorgensen, W.L., Madura, J.D., 1983. Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water. J. Am. Chem. Soc., 105, 1407-

1413.

Joseph, A.P., Agarwal, G., Mahajan, S., Gelly, J.-C., Swapna, L.S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadié, H., Schneider, B., Cadet, F., Srinivasan, N., de Brevern, A.G., 2010. A short survey on Protein Blocks. Biophysical Reviews 2, 137-145.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., Karplus, K., 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51, 504-514.

Kragelj, J., Blackledge, M., Jensen, M.R., 2015. Ensemble Calculation for Intrinsically Disordered Proteins Using NMR Parameters. Adv Exp Med Biol 870, 123-147.

Ladislav, M., Cerny, J., Krusek, J., Horak, M., Balik, A., Vyklicky, L., 2018. The LILI Motif of M3-S2 Linkers Is a Component of the NMDA Receptor Channel Gate. Front. Mol. Neurosci., 11, 113.

Leonard, S., Joseph, A.P., Srinivasan, N., Gelly, J.C., de Brevern, A.G., 2014. mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet. J Biomol Struct Dyn 32, 661-668.

Marsh, J.A., Forman-Kay, J.D., 2009. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. J Mol Biol 391, 359-374.

Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A.G., Gibrat, J.F., 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 5, 17.

Mertens, H.D., Piljic, A., Schultz, C., Svergun, D.I., 2012. Conformational analysis of a genetically encoded FRET biosensor by SAXS. Biophys J 102, 2866-2875.

Mitic, N.S., Malkov, S.N., Kovacevic, J.J., Pavlovic-Lazetic, G.M., Beljanski, M.V., 2018. Structural disorder of plasmid-encoded proteins in Bacteria and Archaea. BMC Bioinformatics 19, 158.

Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., Tyers, M., Forman-Kay, J.D., 2010. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 18, 494-506.

Narwani, T.J., Craveur, P., Shinada, N.K., Santuz, H., Rebehmed, J., Etchebest, C., de Brevern, A.G., 2018. Dynamics and deformability of α-, 310- and π-helices. Archives of Biological Sciences 70, 21-31.

Narwani, T.J., Craveur, P., Shinada, N.K., Floch, A., Santuz, H., Vattekatte, A.M., Srinivasan, N., Rebehmed, J., Gelly, J.C., Etchebest, C., de Brevern, A.G., 2019. Discrete analyses of protein dynamics. J Biomol Struct Dyn, 1-15.

Offmann, B., Tyagi, M., de Brevern, A.G., 2007. Local Protein Structures. Current Bioinformatics 3, 165-202.

Ota, M., Koike, R., Amemiya, T., Tenno, T., Romero, P.R., Hiroaki, H., Dunker, A.K., Fukuchi, S., 2013. An assignment of intrinsically disordered regions of proteins based on NMR structures. J Struct Biol 181, 29-36.

Ozenne, V., Schneider, R., Yao, M., Huang, J.R., Salmon, L., Zweckstetter, M., Jensen, M.R., Blackledge, M., 2012. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. J Am Chem Soc 134, 15138-15148.

Parrinello, M., Rahman, A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics 52, 7182-7190.

Pauling, L., Corey, R.B., 1951. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci U S A 37, 251-256.

Pauling, L., Corey, R.B., Branson, H.R., 1951. The structure of proteins; two hydrogen-

bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37, 205-211.

Pavlovic-Lazetic, G.M., Mitic, N.S., Kovacevic, J.J., Obradovic, Z., Malkov, S.N., Beljanski, M.V., 2011. Bioinformatics analysis of disordered proteins in prokaryotes. BMC Bioinformatics 12, 66.

Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D., Hess, B., Lindahl, E., 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29, 845-854.

Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv Protein Chem 34, 167-339.

Robustelli, P., Piana, S., Shaw, D.E., 2018. Developing a molecular dynamics force field for both folded and disordered protein states. Proc Natl Acad Sci U S A.

Sanchez-Martinez, M., Crehuet, R., 2014. Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings. Phys Chem Chem Phys 16, 26030-26039.

Schlessinger, A., Yachdav, G., Rost, B., 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22, 891-893.

Schneider, B., Gelly, J.C., de Brevern, A.G., Cerny, J., 2014a. Local dynamics of proteins and DNA evaluated from crystallographic B factors. Acta Crystallogr D Biol Crystallogr 70, 2413-2419.

Schneider, B., Cerny, J., Svozil, D., Cech, P., Gelly, J.C., de Brevern, A.G., 2014b. Bioinformatic analysis of the protein/DNA interface. Nucleic Acids Res 42, 3381-3394.

Schrodinger, LLC. 2015. The PyMOL Molecular Graphics System, Version 1.7.2.2.

Shenoy, S.R., Jayaram, B., 2010. Proteins: sequence to structure and function--current status. Curr Protein Pept Sci 11, 498-514.

Sivakolundu, S.G., Bashford, D., Kriwacki, R.W., 2005. Disordered p27Kip1 exhibits intrinsic structure resembling the Cdk2/cyclin A-bound conformation. J Mol Biol 353, 1118-1128.

Sterckx, Y.G., Volkov, A.N., Vranken, W.F., Kragelj, J., Jensen, M.R., Buts, L., Garcia-Pino, A., Jove, T., Van Melderen, L., Blackledge, M., van Nuland, N.A., Loris, R., 2014. Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. Structure 22, 854-865.

Team, R.C., 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Thorpe, M.F., Lei, M., Rader, A.J., Jacobs, D.J., Kuhn, L.A., 2001. Protein flexibility and dynamics using constraint theory. J Mol Graph Model 19, 60-69.

Tompa, P., 2002. Intrinsically unstructured proteins. Trends Biochem Sci 27, 527-533.

Tyagi, M., Bornot, A., Offmann, B., de Brevern, A.G., 2009a. Protein short loop prediction in terms of a structural alphabet. Comput Biol Chem 33, 329-333.

Tyagi, M., Bornot, A., Offmann, B., de Brevern, A.G., 2009b. Analysis of loop boundaries using different local structure assignment methods. Protein Sci 18, 1869-1881.

Tyagi, M., Gowri, V.S., Srinivasan, N., de Brevern, A.G., Offmann, B., 2006. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 65, 32-39.

Uversky, V.N., 2002a. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11, 739-756.

Uversky, V.N., 2002b. Cracking the folding code. Why do some proteins adopt partially folded conformations, whereas other don't? FEBS Lett 514, 181-183.

Uversky, V.N., 2017. Intrinsic disorder here, there, and everywhere, and nowhere to escape from it. Cell Mol Life Sci 74, 3065-3067.

Uversky, V.N., Gillespie, J.R., Fink, A.L., 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41, 415-427.

van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., Tironi, I.G., 1996. Biomolecular Simulation: The GROMOS96 Manual and User Guide, 1042.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., Sussman, J., Svergun, D.I., Uversky, V.N., Vendruscolo, M., Wishart, D., Wright, P.E., Tompa, P., 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res 42, D326-335.

Weeks, S.D., Baranova, E.V., Heirbaut, M., Beelen, S., Shkumatov, A.V., Gusev, N.B., Strelkov, S.V., 2014. Molecular structure and dynamics of the dimeric human small heat shock protein HSPB6. J Struct Biol 185, 342-354.

Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293, 321-331.

Xiang, S., Gapsys, V., Kim, H.Y., Bessonov, S., Hsiao, H.H., Mohlmann, S., Klaukien, V., Ficner, R., Becker, S., Urlaub, H., Luhrmann, R., de Groot, B., Zweckstetter, M., 2013. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. Structure 21, 2162-2174.

Zhang, Y., Stec, B., Godzik, A., 2007. Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. Structure 15, 1141-1147.