Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution

Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution

Julia Marchetti [1], Alexander Miguel Monzon [1,2], Silvio C.E. Tosatto [2], Gustavo Parisi [1] and María Silvina Fornasari [1]

*[1] - **Departamento de Ciencia y Tecnología,** CONICET, Universidad Nacional de Quilmes, Roque Sáenz Peña 352, B1876BXD, Bernal, Provincia de Buenos Aires, Argentina*
*[2] - **Department of Biomedical Sciences,** University of Padua, Padua, Italy*

Correspondence to Gustavo Parisi: *gusparisi@gmail.com*

https://doi.org/10.1016/j.jmb.2019.01.031
Edited by Monika Fuxreiter

Abstract

The conformations accessible to proteins are determined by the inter-residue interactions between amino acid residues. During evolution, structural constraints that are required for protein function providing biologically relevant information can exist. Here, we studied the proportion of sites evolving under structural constraints in two very different types of ensembles, those coming from ordered and disordered proteins. Using a structurally constrained model of protein evolution, we found that both types of ensembles show comparable, near 40%, number of positions evolving under structural constraints. Among these sites, ~68% are in disordered regions and ~57% of them show long-range inter-residue contacts. Also, we found that disordered ensembles are redundant in reference to their structurally constrained evolutionary information and could be described on average with ~11 conformers. Despite the different complexity of the studied ensembles and proteins, the similar constraints reveal a comparable level of selective pressure to maintain their biological functions. These results highlight the importance of the evolutionary information to recover meaningful biological information to further characterize conformational ensembles.

Introduction

The protein native state is described by a collection of the different conformers which a given sequence could adopt. This collection is also called a conformational ensemble and is an essential concept to understand protein biology [1,2]. The existence of conformational ensembles is known since the crystallization of hemoglobin with its two conformational states T and R (deoxy and oxygenated forms) in the early 1960. The growth of Protein Data Bank (PDB) redundancy, refinement and development of techniques such as NMR, small-angle X-ray scattering, and single-molecule spectroscopy over the last years have allowed the experimental characterization of a large number of protein ensembles [2,3]. Structural differences between conformers could result from the relative movements of large domains as rigid bodies [4], secondary and tertiary element rearrangements [5], and loop movements [6]. Apparently, most globular proteins have very few conformers describing their native state to achieve their functions [7]. Proteins with low flexibility at the backbone level, called rigids, have only one conformer in their ensembles [7] like the cellulase from *Clostridium cellulolyticum* [8]. Hemoglobin, as mentioned previously, is the paradigm for proteins with two conformers [9], while the dimeric catabolite activator protein [10] and the human glucokinase have three [11]. Complex proteins composed of several different chains, like mitochondrial ATP synthase, could have at least seven conformers [12]. As protein flexibility increases, the number of conformers in the ensemble increases as well, giving rise to very complex ensembles as in the case of intrinsically disordered proteins (IDPs) or regions (IDRs). IDPs are characterized by the lack of tertiary structure under physiological conditions [13,14]. IDP ensembles are composed by a large number of interconverting conformers given their low free-energy barriers among them [15]. Far

0022-2836/© 2019 Published by Elsevier Ltd.

Please cite this article as: J. Marchetti, A. M. Monzon, S. C.E. Tosatto, et al., Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution, Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2019.01.031

from being random polymers or random-coiled ensembles, it is becoming evident that IDP ensembles are not fully disordered, showing transient short and long-range structural organization [16]. Order–disorder transitions are frequently observed in IDPs or IDRs, sometimes associated with ligand binding [17] but in other cases just reflecting the heterogeneous composition of the ensembles [7,18].

Here, we studied the level of structural constraints in IDPs ensembles compared with those found in globular proteins. Structural constraints could be studied using direct methods such as the measurements of contacts between residues in a given conformer and some derived parameters such as the contact density (mean number of residue–residue contacts per residue) or their interaction networks [19]. However, inter-residue contacts could be artifacts or simply be irrelevant in very complex ensembles such as those found in IDPs, making it difficult to detect biologically relevant conformers [20]. For these reasons, in this work, we evaluated the amount of structural constraints using an evolutionary approach. It is a well-established concept that conservation of protein structures during evolution constrains sequence divergence modulating in this way the amino acid substitution pattern of certain positions [21,22]. These structural constraints are evidenced in sequence alignments as differentially conserved positions, showing a given physicochemical bias or subject to coevolutionary processes due to their relative importance to maintain protein fold and dynamics (i.e., conservation of given interactions to increase stability, sustain protein movements). This structurally constrained (SC) substitution pattern has been exploited to improve models of molecular evolution [23–25], explain rate heterogeneity [26], make functional predictions [27], and compare the substitution process in ordered and disordered proteins [28] and in the inference of given tertiary folds [29] to mention just a few examples of their many applications. Furthermore, evolutionary information could be used to predict native contacts and structural models of globular domains [30–32]. More recently, these methods were adapted to successfully predict globular states in disordered proteins and to show the evolutionary constraints in protein interfaces between disordered and ordered proteins again showing the importance of SC information during evolution [33,34].

Substitution patterns observed in sequence alignments can be described by evolutionary models [35]. Alternative models, making different assumptions about the amino acid substitution pattern, can be compared using maximum likelihood (ML) estimations to decide which assumptions better describe the evolutionary process in a given family. In particular, in this work, a model of protein evolution using protein structure to derive an SC site-specific substitution pattern was used [24]. As this model is structure-specific, each protein conformation represents different evolutionary models. Using ML estimations, we then compared how the SC substitution pattern outperforms models of evolution lacking structural information (e.g., JTT [36], Dayhoff [37], WAG [38]) in its ability to explain the observed site-specific substitution pattern in a set of homologous proteins for each studied protein. Interestingly, considering all conformers in the ensembles of globular and IDP proteins, we found that the number of SC positions is similar for both kinds of proteins.

Results

Description of the data sets

In the last years, an emerging picture evidences that increasing structural differences between conformers, connected by very different dynamical behaviors, produces a continuum in protein space [39]. One extreme feature of this continuum is the presence of rigids proteins with almost no backbone differences among their conformers and just displaying only conformational diversity at the residue level [7]. Increasing conformational diversity at the backbone level could evidence the presence of disorder, where the appearance of short-time dynamical behavior allows for the sampling of a large conformational space [40]. Figure 1 shows different types of ensembles as protein conformational diversity increases. In one extreme of the distribution (left-side panel in Fig. 1), typical globular or ordered proteins are shown. These proteins generally show large proportions of secondary structure where their spatial arrangement defines a single tertiary structure and hydrophobic core. The higher density of inter-residue interactions of this core constrains evolutionary rates when compared to exposed residues [41] and also contains enough information to define a global tertiary arrangement [42]. As mentioned before, ordered proteins could also contain different conformers to achieve their biological functions (Fig. 1, middle panel), giving place to additional restrictions in the protein substitution pattern [43]. Middle-panel examples of Fig. 1 also display proteins with ordered or globular regions as well as with very flexible regions showing different dynamical behavior and possibly originating disordered regions of different lengths. Right panel in Fig. 1 shows a typical ensemble of IDPs showing a collection of conformers determined by NMR. These ensembles show highly flexible chains and eventually small and transient segments of secondary or tertiary structure [44]. Consequently, IDPs have a large degree of conformational entropy that can be limited by inter-residue interactions originating a complex mixture of conformers in the ensemble [15,20]. As described in Materials and Methods, two hand-curated data sets were analyzed.

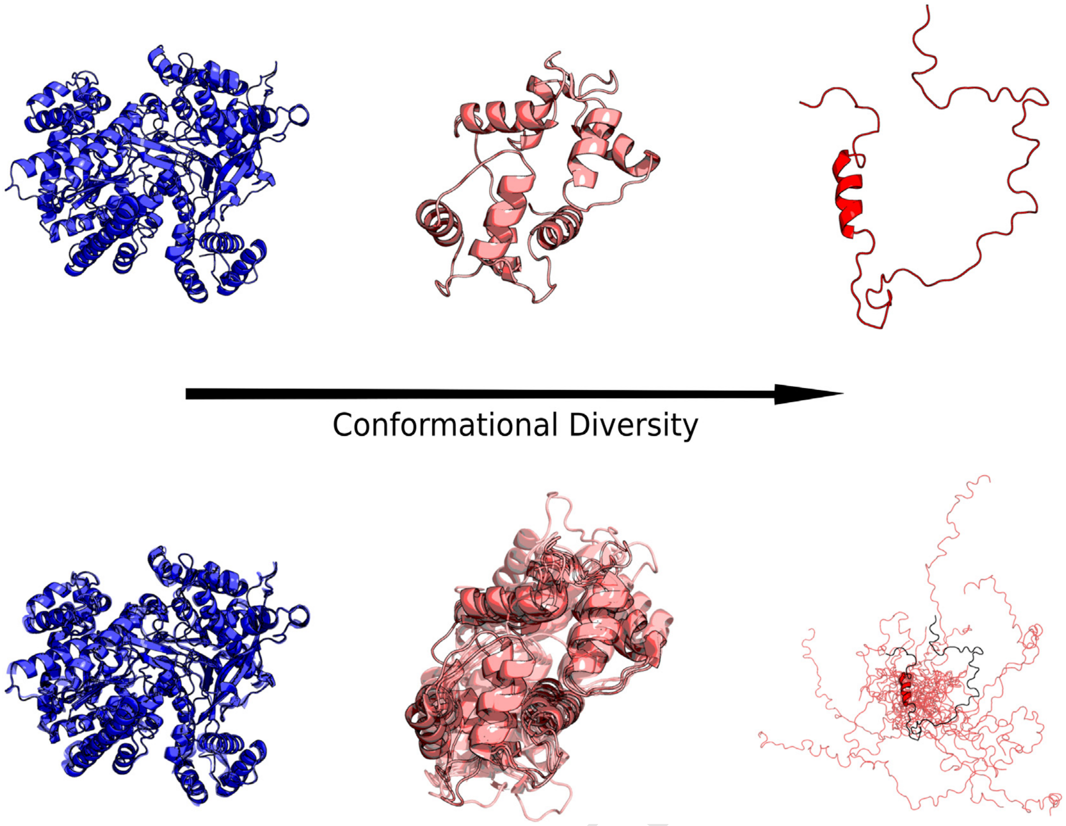

Fig. 1. Different protein ensembles as a function of flexibility increment. Top panel shows a given conformer, while the bottom panel shows all the available conformers in the ensemble. Left, maltodextrin phosphorylase, (PDB codes = 1AHP_A, 1AHP_B, 1L5V_B) showed as a rigid protein with 6.53% disordered and taken as a representative of ordered proteins. Calmodulin (PDB codes = 2FOT_A, 1LIN_A,1NIW_E, 3G43_A, 2BE6_A, 1CDL_A, 3GP2_A, 4L79_B, 1CLL_A) shows 10.64% of disorder. Thylakoid soluble phosphoprotein, (PDB ID = 2FFT_A) is a typical IDP ensemble with 100% of estimated disorder. The percentages of disorder were estimated with ESpritz.

The ordered data set is composed of 183 proteins with known crystallographic structure containing non-missing residues, and a disordered data set contains 93 NMR ensembles of different proteins. Disorder has been estimated in both data sets using ESpritz and Mobi 2.0 for the disordered and ordered data sets, respectively (see Materials and Methods). As is it shown in Fig. 2, ordered proteins show a low predicted content of disordered residues, while the disordered data set shows a distribution of disordered residues. The median of these distribution is 58% of disordered positions (minimum 40% and up to 98%). It is then expected that the disordered data set contains small globular regions and more than the half of the protein in a disordered state. Sequence alignments for each protein in each data set were extracted from HSSP database (see Materials and Methods), and to avoid high occurrence of indels, sequences above 30% identity with the protein with known structure were only considered. Additional information about protein alignments could be found in Fig. S1.

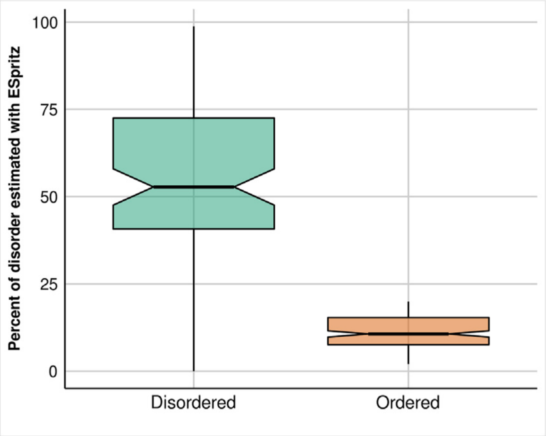

Fig. 2. Estimation of disorder content using NMR-ESpritz in the disordered set and ESpritz in the ordered set. It is shown that the ordered set has a low proportion of disorder well below the reported error in the estimation [45].

Please cite this article as: J. Marchetti, A. M. Monzon, S. C.E. Tosatto, et al., Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution, Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2019.01.031

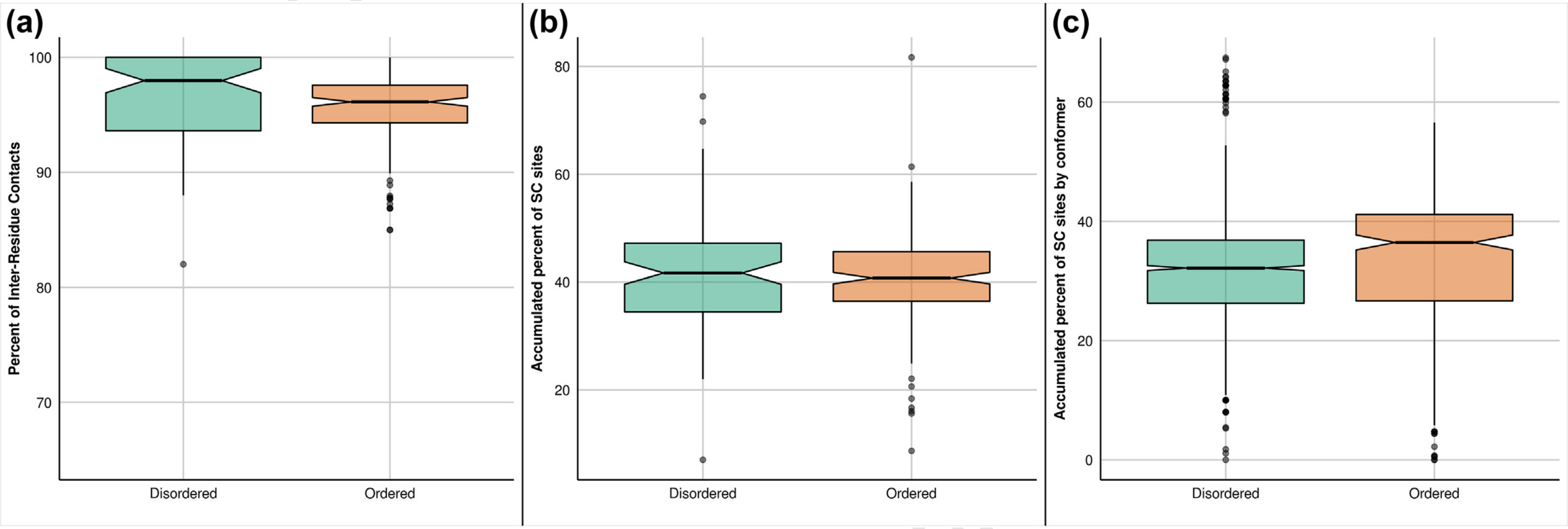

Fig. 3. (a) Percentage of inter-residue contacts for the disordered and ordered data sets (average median of 96.1%). (b) Distribution of the accumulated number of SCs for both data sets showing 41.6% and 40.5% of the positions. The distributions are statistically similar using a Kolmogorov–Smirnov test with *p* value = 0.39 and Mann–Whitney–Wilcoxon test with *p* value = 0.45. (c) Distribution of SCs per conformer per protein showing a median of 32.1% and 36.1% of their sites constrained.

Physical contacts *versus* structural constraints during evolution

To assess the structural constraints in ordered and disordered ensembles, we quantified the inter-residue interactions accumulating the contact information for each site through all the available conformers in each corresponding ensemble (Fig. S2, panel A). Accumulation is a reasonable idea sustained by the particular contributions each conformer makes to the biological function [2]. As a result, we obtained that the great majority of residues are involved in inter-residues contacts as it is shown in Fig. 3a. Permanent secondary and tertiary contacts in ordered proteins define their levels of structural constraints, while the contribution of transient contacts along the entire ensemble of IDPs produces almost the same amount of accumulated inter-residues contacts (third quartile is 100% and 97% for IDPs and ordered sets, respectively). According to this result, the vast majority of positions in IDPs are constrained by structural restrictions as well as those for ordered proteins. However, it is well established that the pattern of amino acid substitutions in IDPs is different from the one observed in ordered proteins. IDPs show also a highly conserved composition of amino acids [46] instead of the well-defined site-specific substitution pattern observed in ordered proteins [47]. In addition, IDPs and IDRs show higher evolutionary rates as well as higher rates of insertions and deletions compared with their ordered counterpart [13,44,48]. To elucidate the influence of such high levels of structural constraints (Fig. 3a), we turned to study the substitution pattern observed in the homologous family of each protein in both data sets. Using ML comparisons (Fig. S2, panel B), we assessed if the observed substitution pattern is better explained by an evolutionary model containing structural information (like SCPE, see Materials and Methods) or by other models not containing this information (JTT, Dayhoff and WAG models, see Materials and Methods). For every position showing a SCPE site-specific substitution matrix that outperforms each one of the other three models, it is inferred as a site evolving under structural constraints. Considering the different nature of ordered and disordered ensembles, unexpectedly, we found that the percentages of SCs are almost the same in both types of ensembles (41.6% and 40.5% for disordered and ordered data sets; Fig. 3b) and much lower than estimations made using the accumulated account of inter-residue contacts. Interestingly, the individual conformers show slightly less percentages of SC sites (Fig. 3c) showing 32.1% and 36.1% in average for the disordered and ordered data sets.

SC sites

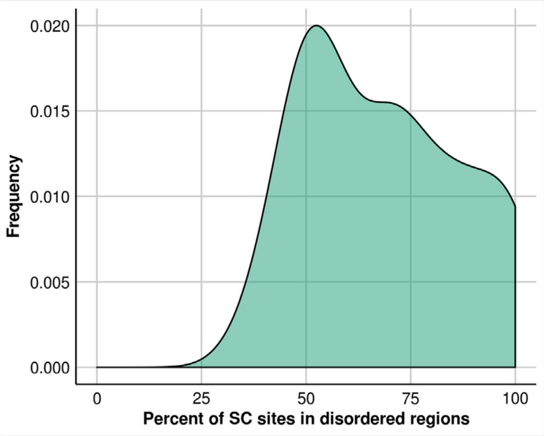

Fig. 4. Distribution of the accumulated number of SCs along all the ensembles. On average, 68.3% of the SC sites belong to predicted disordered regions.

SC sites are then sites that at least have one physical inter-residue contact in at least one conformer but also, and more importantly, modulates sequence divergence in that specific position. To further investigate these structural constraints, we studied the distribution of SC sites. We found that ~68% of the SCs are located in the disordered regions of the proteins belonging to the disordered data set (Fig. 4). As we mentioned before, disordered proteins could have permanent or transient globular regions that could increase the structural constraints of the protein as a whole. However, the number of SC sites in the globular or ordered regions of the disordered proteins is ~32%. These results indicate that globular regions of disordered proteins are less constrained than the corresponding one observed in the ordered data set (see Fig. 3b). Also, following our definition of inter-residue contacts (see Materials and Methods), all estimated contacts are tertiary and in ~57% the SCs are classified as long-range inter-residue contacts (see Fig. 5). This

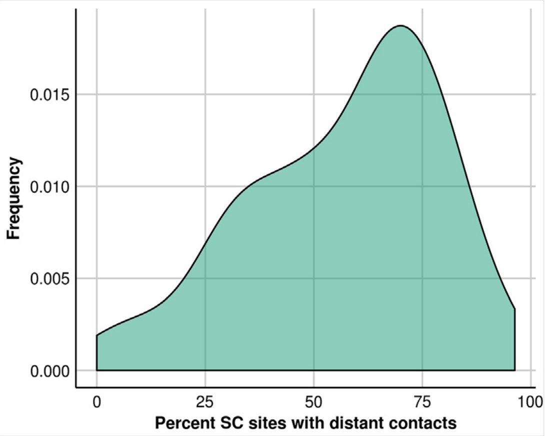

Fig. 5. Distribution of the accumulated number of SCs along all the ensembles, with long-distance contacts (at least five residues away). In average, 56.8% of the SC sites have long-range inter-residue contacts.

Please cite this article as: J. Marchetti, A. M. Monzon, S. C.E. Tosatto, et al., Ensembles from Ordered and Disordered Proteins Reveal Similar Structural Constraints during Evolution, Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2019.01.031

finding can explain how SC sites could appear in disordered regions. As we can see in Fig. 6, disordered proteins could have large conformational diversity. However, among the representative conformers of the ensembles, we can find some of them collapsing over the globular part of the protein or just adopting close conformations increasing in this way the number of contacts per site. As it is shown in Fig. 7, 51% of the positions have contacts that are present in the 100% of the conformers of the ensemble. However, there is still a tail in the distribution showing that single conformers could have SC sites; in other words, single conformers could have inter-residue contacts that modulate the substitution pattern of those positions.

Ensemble redundancy

How many conformers are required to fully describe evolutionary structural constraints contained in sequence alignments? When we calculated the minimum number of conformers per ensemble to reach the accumulated SC percentage per protein, we found that on average ~11 conformers are required for the proteins in the disordered data set (see Fig. 8), while

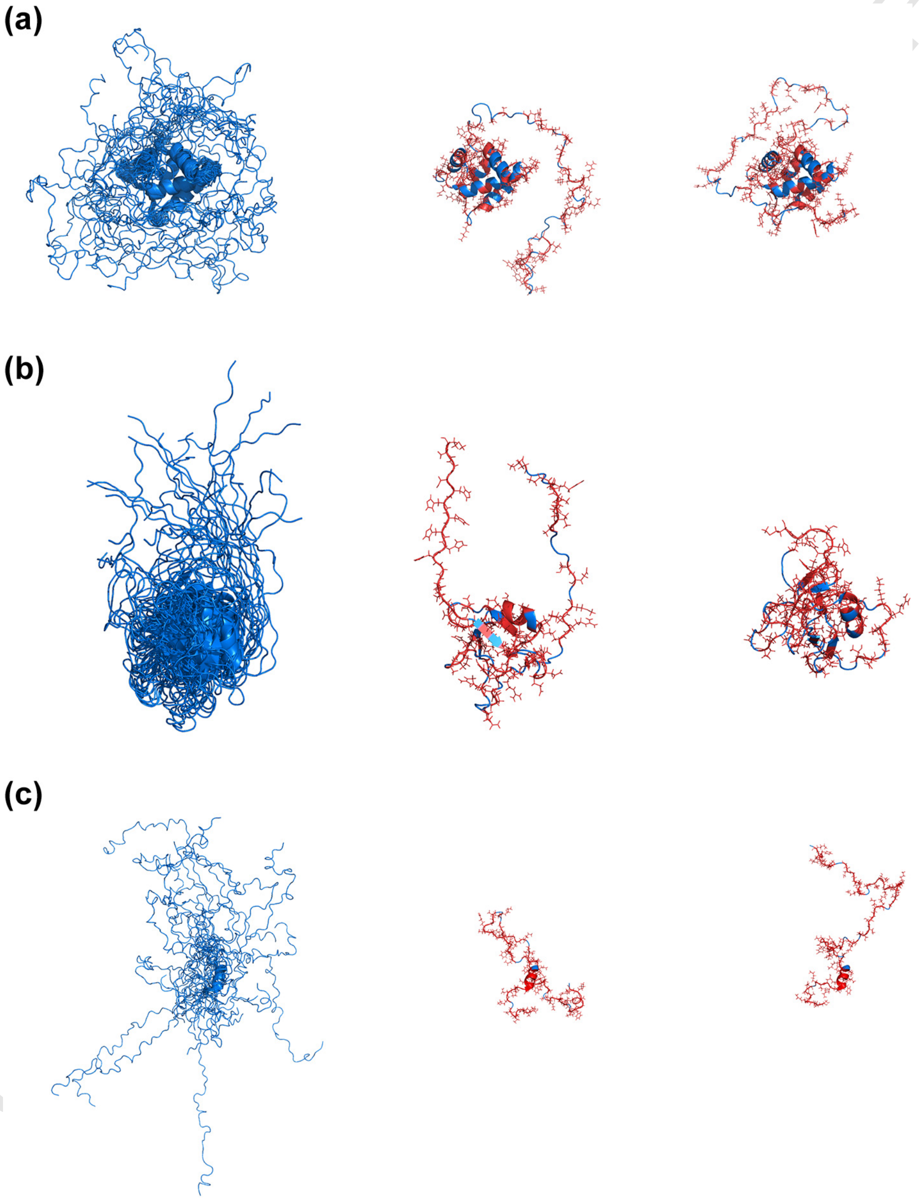

Fig. 6. Examples showing SC sites distribution in different conformers. The three panels (top, middle, and bottom) contain disordered proteins showing in the left the available ensemble, while in the middle and in the right, different conformers are shown. Proteins are shown. Cartoon representation was used. iSC sites are shown in red sticks, and the rest in blue. 2JRF_A, 2ADZ_A and 5MRG_A are the corresponding PDB codes for the top, middle, and bottom panels.

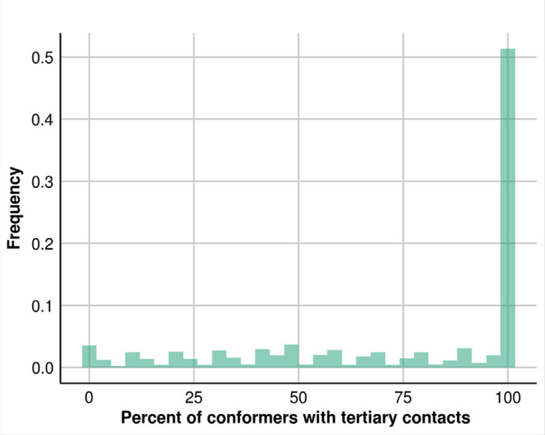

Fig. 7. Approximately ~51% of SC sites present contacts in 100% of the conformers, and only ~3% of SC sites present contacts in 50% of the conformers.

in the ordered data set, it is ~1.5. The value for the ordered data set is consistent with the available experimental evidence. Most ordered proteins show low conformational diversity, and then are called "rigid" [7], or could show very few conformers, mostly two, referring to the bound and unbound forms of the protein [49–51]. Due to the complexity of disordered ensembles, the number of conformers is difficult if not impossible to estimate. However, our measure of the number of conformers required to explain the evolutionary SC information in sequence alignments could offer a proxy to the number of conformers. Since the average of conformers in the NMR ensembles in our data set is ~20, our results indicate that they are Q8 mostly redundant.

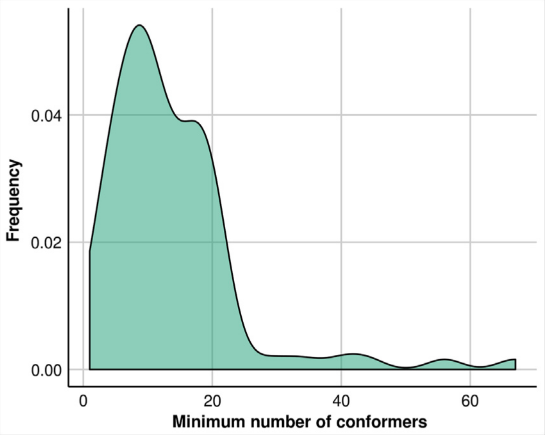

Fig. 8. Distribution of the minimum number conformers to reach the accumulated percentage of SC sites per protein for the 93 disordered proteins corresponding to the set obtained with Mobi 2.0 and ESpritz (NMR). Minimum = 1, average ~11, and maximum ~64.

Discussion

Two main findings emerge from the present work. First, the number of positions having inter-residue contacts accumulated along all available conformers in each ensemble approaches almost 100% of the positions (Fig. 3a). However, as we have shown, the average percentage of positions evolving under structural constraints is much lower, ~40% (Fig. 3b). Part of this reduction is expected, given that not all intramolecular non-covalent contacts could be equally relevant, for example, in structure stabilization [52]. Inaccurate models and atomic coordinate uncertainties could also play a role to explain the observed difference between the amount of physical contacts and the observed evolutionary derived structural constraints [53–55]. In addition, the reduction could be also attributed to the lack of structure/conformer-specific information contained in sequence alignments. This effect operates over SCPE substitution matrices, which are site and conformer specific but are evaluated using sequence alignments from corresponding homologous families. Thus, evolutionary information contained in those alignments reflects constraints of several sorts, such as structural divergence [41] or dynamical adaptations [56,57], which could certainly modify the contact pattern in the homologous proteins. It is then expected that this ~40% of SCs on average obtained for both ensembles does not capture subtle inter-residue contacts originated in functional adaptations for individual proteins. In line with this observation, it has been recently shown that the use of sequence alignments recovers the most conserved pattern of inter-residues contacts when co-evolutionary and evolutionary coupling methods are used [57]. The other important result is related with the comparable structural constraints on sequence divergence in ordered and disordered proteins (Fig. 3b). Our results suggest that individual contributions of each conformer in the disordered ensemble are required to sustain biological function as is well established for ordered proteins, and more recently suggested for disordered ones [2,13,48]. These small contributions from each disordered conformer give overall the same proportion of structural constraints as found in ordered proteins, possibly with different weights according to their biological role.

Interestingly, the number of conformers in the IDPs ensembles to reach the corresponding level of global constraints per protein is ~11 (Fig. 8). This means that IDP ensembles are redundant in terms of conformations and that possibly the number of biologically relevant conformers in IDP ensembles would not be so large as expected due to their high flexibility. These results are in agreement with the idea that different members of the ensemble could be directly involved in protein function, but also, they could be important as a local minimum representatives in the interconversion of biologically relevant conformations [58].

Our results highlight the importance of the evolutionary analysis in the discrimination of inter-residue contacts to detect meaningful biological information as well as the estimation of the number of conformers and structural constraints in such complex ensembles as those belonging to IDPs.

Materials and Methods

Data set collection

Globular or ordered protein ensembles were obtained from the CoDNas database [59]. Considering the presence of missing residues as a primary indicator of IDRs in proteins [60], we selected 183 proteins having no missing residues in any of their available conformers. These selected protein ensembles have at least five conformers in the database to assure a good estimation of the conformational variability [61]. Only the pair of conformers showing the maximum RMSD along all the ensemble was considered in this set. To obtain the IDPs data set, we predicted and estimated disorder in all the available NMR protein structures in PDB (available May 2018) using NMR-ESpritz [45] and Mobi 2.0 [62]. After a hand-curated revision considering length and protein biology, we finally obtained 93 protein NMR ensembles with more than 40% of disordered positions. Ordered set of proteins showed negligible levels of disorder predicted with ESpritz X-ray (see Figs. 3 and S3).

SC substitution pattern estimation

In Fig. S2, we resumed the workflow to analyze SCs and physical contacts. For each conformer and each protein in both data sets (for the disordered data set, we considered all the NMR available conformers, and for the ordered data set, we used those corresponding for the maximum RMSD according to CoDNaS), the SCPE model of protein evolution was run [24]. SCPE derives site-specific substitution matrices using evolutionary simulations under neutral conditions for protein fold conservation [47,63] (please see Fig. S4). Briefly, it uses energetic calculations to evaluate the structural perturbation introduced by non-synonymous substitutions in the simulation process. Using ML estimations, it is possible to compare SCPE matrices with models lacking structural information such as JTT [36], Dayhoff [64], and WAG [38]. Site-specific ML calculations were performed with the HYPHY package [65]. The alignments used for the ML analysis were obtained from HSSP [66] database. Neighbor-joining distance phylogenetic trees were obtained with the Phylip [67] package. To define whether a site was SC, Akaike information criteria (AIC) coefficient was used [68] and a ranking for the estimated models was made using ΔAIC [69] in which models having ΔAIC ≤ 2 have a substantial support, those where ΔAIC is between 4 and 7 have an intermediate support, and those with ΔAIC > 10 have no support. Tertiary contacts were estimated considering the distance between two non-contiguous residues having the van der Waals spheres of each residue side chain heavy atoms below 1.0 Å. Long-range inter-residues contacts were estimated using same definition but considering ±5 residues of a given residue.

Acknowledgments

G.P. and M.S.F. are CONICET researchers, and J.M. and A.M.M. are PhD and postdoctoral fellows of the same institution.

This work was supported by Universidad Nacional de Quilmes (PUNQ 1004/11) (G.P.), Agencia de Ciencia y Tecnología (PICT-2014-3430) (G.P.), and COST Action (BM1405) NGP-net (S.C.E.T.). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 778247 (IDPfun). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Q9

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.01.031.

Received 5 November 2018;
Received in revised form 23 January 2019;
Accepted 24 January 2019
Available online xxxx

Keywords:
protein evolution;
protein ensemble;
conformational diversity;
disordered proteins

Abbreviations used: Q6
PDB, Protein Data Bank; IDP, intrinsically disordered protein; IDR, intrinsically disordered region; SC, structurally constrained site; ML, maximum likelihood; AIC, Akaike information criteria.

References Q11

[1] C.J. Tsai, S. Kumar, B. Ma, R. Nussinov, Folding funnels, binding funnels, and protein function, Protein Sci. 8 (1999) 1181–1190.

[2] G. Wei, W. Xi, R. Nussinov, B. Ma, Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell, Chem. Rev. (2016) (acs.chemrev.5b00562). Q12

[3] C. Marino-Buslje, A.M. Monzon, D.J. Zea, S. Fornasari, G. Parisi, On the dynamical incompleteness of the Protein Data Bank, Brief. Bioinform. (2017) 1–4.

[4] M. Gerstein, N. Echols, Exploring the range of protein flexibility, from a structural proteomics perspective, Curr. Opin. Chem. Biol. 8 (2004) 14–19.

[5] M. Gerstein, A database of macromolecular motions, Nucleic Acids Res. 26 (1998) 4280–4290.

[6] Y. Gu, D.-W. Li, R. Brüschweiler, Decoding the mobility and time scales of protein loops, J. Chem. Theory Comput. 11 (2015) 1308–1314.

[7] A.M. Monzon, D.J. Zea, M.S. Fornasari, T.E. Saldaño, S. Fernandez-Alberti, S.C.E. Tosatto, G. Parisi, Conformational diversity analysis reveals three functional mechanisms in proteins, PLoS Comput. Biol. 13 (2017) 1–29.

[8] G. Parsiegla, C. Reverbel-Leroy, C. Tardif, J.P. Belaich, H. Driguez, R. Haser, Crystal structures of the cellulase Cel48F in complex with inhibitors and substrates give insights into its processive action, Biochemistry 39 (2000) 11238–11246.

[9] M.F. Perutz, W. Bolton, R. Diamond, H. Muirhead, H.C. Watson, Structure of haemoglobin. An X-ray examination of reduced horse haemoglobin, Nature 203 (1964) 687–690.

[10] N. Popovych, S. Sun, R.H. Ebright, C.G. Kalodimos, Dynamically driven protein allostery, Nat. Struct. Mol. Biol. 13 (2006) 831–838.

[11] K. Kamata, M. Mitsuya, T. Nishimura, J.-I. Eiki, Y. Nagata, Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase, System 12 (2004) 429–438.

[12] A. Zhou, A. Rohou, D.G. Schep, J.V. Bason, M.G. Montgomery, J.E. Walker, N. Grigorieff, J.L. Rubinstein, Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM, elife 4 (2015), e10180.

[13] J. Siltberg-Liberles, J.a. Grahnen, D.a. Liberles, The evolution of protein structures and structural ensembles under functional constraint, Genes 2 (2011) 748–762.

[14] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (2002) 527–533.

[15] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A.K. Dunker, I.C. Felli, J.D. Forman-Kay, R.W. Kriwacki, R. Pierattelli, J. Sussman, D.I. Svergun, V.N. Uversky, M. Vendruscolo, D. Wishart, P.E. Wright, P. Tompa, pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins, Nucleic Acids Res. 42 (2014) D326–D335.

[16] P. Tompa, Unstructural biology coming of age, Curr. Opin. Struct. Biol. 21 (2011) 419–425.

[17] D. Zea, A.M. Monzon, C. Gonzalez, M.S. Fornasari, S.C.E. Tosatto, G. Parisi, Disorder transitions and conformational diversity cooperatively modulate biological function in proteins, Protein Sci. 25 (2016) 1138–1146.

[18] S. DeForte, V.N. Uversky, Resolving the ambiguity: making sense of intrinsic disorder when PDB structures disagree, Protein Sci. 25 (2016) 676–688.

[19] D. Piovesan, G. Minervini, S.C. Tosatto, The RING 2.0 web server for high quality residue interaction networks, Nucleic Acids Res. 44 (W1) (2016) W367–W374.

[20] P. Sormanni, D. Piovesan, G.T. Heller, M. Bonomi, P. Kukic, C. Camilloni, M. Fuxreiter, Z. Dosztanyi, R.V. Pappu, M.M. Babu, S. Longhi, P. Tompa, A.K. Dunker, V.N. Uversky, S.C.E. Tosatto, M. Vendruscolo, Simultaneous quantification of protein order and disorder, Nat. Chem. Biol. 13 (2017) 339–342.

[21] J. Overington, M.S. Johnson, A. Sali, T.L. Blundell, Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction, Proc. Biol. Sci. 241 (1990) 132–145.

[22] C.L. Worth, S. Gong, T.L. Blundell, Structural and functional constraints in the evolution of protein families, Nat. Rev. Mol. Cell Biol. 10 (2009) 709–720.

[23] U. Bastolla, H.E. Roman, M. Vendruscolo, Neutral evolution of model proteins: diffusion in sequence space and overdispersion, J. Theor. Biol. 200 (1999) 49–64.

[24] G. Parisi, J. Echave, Structural constraints and emergence of sequence patterns in protein evolution, Mol. Biol. Evol. 18 (2001) 750–756.

[25] C.L. Kleinman, N. Rodrigue, N. Lartillot, H. Philippe, Statistical potentials for improved structurally constrained evolutionary models, Mol. Biol. Evol. 27 (2010) 1546–1560.

[26] J. Echave, S.J. Spielman, C.O. Wilke, Causes of evolutionary rate variation among protein sites, Nat. Rev. Genet. 17 (2016) 109–121.

[27] A.L. Simon, E.A. Stone, A. Sidow, Inference of functional regions in proteins by quantification of evolutionary constraints, Proc. Natl. Acad. Sci. U. S. A. (2001). Q13

[28] C.J. Brown, A.K. Johnson, G.W. Daughdrill, Comparing models of evolution for ordered and disordered proteins, Mol. Biol. Evol. 27 (2010) 609–621.

[29] J. Surkont, J.B. Pereira-Leal, Evolutionary patterns in coiled-coils, Genome Biol. Evol. 7 (2015) 545–556.

[30] T.A. Hopf, C.P.I. Schärfe, J.P.G.L.M. Rodrigues, A.G. Green, O. Kohlbacher, C. Sander, A.M.J.J. Bonvin, D.S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes, elife 3 (2014), https://doi.org/10.7554/eLife.03430.

[31] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) E1293–E1301.

[32] S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information, elife 3 (2014) https://doi.org/10.7554/elife.02030.

[33] A. Toth-Petroczy, P. Palmedo, J. Ingraham, T.A. Hopf, B. Berger, C. Sander, D.S. Marks, Structured states of disordered proteins from genomic sequences, Cell 167 (2016) 158–170.e12.

[34] R. Pancsa, F. Zsolyomi, P. Tompa, Co-evolution of intrinsically disordered proteins with folded partners witnessed by evolutionary couplings, Int. J. Mol. Sci. 19 (2018), https://doi.org/10.3390/ijms19113315.

[35] J.L. Thorne, Models of protein sequence evolution and their applications, Curr. Opin. Genet. Dev. 10 (2000) 602–605.

[36] D.T. Jones, W.R. Taylor, J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput. Appl. Biosci. 8 (1992) 275–282.

[37] M.O. Dayhoff, A model of evolutionary change in proteins, Atlas Protein Seq. Struct. 5 (1972) 89–99.

[38] S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, Mol. Biol. Evol. 18 (2001) 691–699.

[39] H.N. Motlagh, J.O. Wrabl, J. Li, V.J. Hilser, The ensemble nature of allostery, Nature 508 (2014) 331–339.

[40] R.B. Berlow, H. Jane Dyson, P.E. Wright, Functional advantages of dynamic protein disorder, FEBS Lett. 589 (2015) 2433–2440.

[41] K. Illergård, D.H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence—a study of structural response in protein cores, Proteins 77 (2009) 499–508.

[42] A.F. Pereira de Araujo, J.N. Onuchic, A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 19001–19004.

[43] D. Javier Zea, A. Miguel Monzon, M.S. Fornasari, C. Marino-Buslje, G. Parisi, Protein conformational diversity correlates with evolutionary rate, Mol. Biol. Evol. 30 (2013) 1500–1503.

[44] J.D. Forman-Kay, T. Mittag, From sequence and forces to structure, function, and evolution of intrinsically disordered proteins, Structure 21 (2013) 1492–1499.

[45] I. Walsh, A.J.M. Martin, T. Di Domenico, S.C.E. Tosatto, ESpritz: accurate and fast prediction of protein disorder, Bioinformatics 28 (2011) 503–509.

[46] H.A. Moesa, S. Wakabayashi, K. Nakai, A. Patil, Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification, Mol. BioSyst. 8 (2012) 3262–3273.

[47] M.S. Fornasari, G. Parisi, J. Echave, Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations, Mol. Biol. Evol. 19 (2002) 352–356.

[48] C.J. Brown, A.K. Johnson, A. Keith Dunker, G.W. Daughdrill, Evolution and disorder, Curr. Opin. Struct. Biol. 21 (2011) 441–446.

[49] A. Gutteridge, J. Thornton, Conformational changes observed in enzyme crystal structures upon substrate binding, J. Mol. Biol. 346 (2005) 21–28.

[50] A. Gutteridge, J. Thornton, Conformational change in substrate binding, catalysis and product release: an open and shut case, FEBS Lett. 567 (2004) 67–73.

[51] T. Amemiya, R. Koike, A. Kidera, M. Ota, PSCDB: a database for protein structural change upon ligand binding, Nucleic Acids Res. 40 (2012) D554–D558.

[52] R. Sathyapriya, J.M. Duarte, H. Stehr, I. Filippis, M. Lappe, Defining an essence of structure determining residue contacts in proteins, PLoS Comput. Biol. 5 (2009), e1000584.

[53] S.O. Garbuzynskiy, B.S. Melnik, M.Y. Lobanov, A.V. Finkelstein, O.V. Galzitskaya, Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins: Struct. Funct. Bioinf. 60 (2005) 139–147.

[54] C.A.E.M. Spronk, J.P. Linge, C.W. Hilbers, G.W. Vuister, Improving the quality of protein structures derived by NMR spectroscopy, J. Biomol. NMR 22 (2002) 281–289.

[55] C.A.E.M. Spronk, S.B. Nabuurs, A.M.J.J. Bonvin, E. Krieger, G.W. Vuister, G. Vriend, The precision of NMR structure ensembles revisited, J. Biomol. NMR 25 (2003) 225–234.

[56] A.M. Monzon, D.J. Zea, C. Marino-Buslje, G. Parisi, Homology modeling in a dynamical world, Protein Sci. (2017), https://doi.org/10.1002/pro.3274.

[57] D.J. Zea, A.M. Monzon, G. Parisi, C. Marino-Buslje, How is structural divergence related to evolutionary information? Mol. Phylogenet. Evol. 127 (2018) 859–866.

[58] K.K. Turoverov, I.M. Kuznetsova, V.N. Uversky, The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation, Prog. Biophys. Mol. Biol. 102 (2010) 73–84.

[59] A.M. Monzon, C.O. Rohr, M.S. Fornasari, G. Parisi, CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state, Database 2016 (2016), baw038.

[60] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.

[61] R.B. Best, K. Lindorff-Larsen, M.A. DePristo, M. Vendruscolo, Relation between native ensembles and experimental structures of proteins, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 10901–10906.

[62] D. Piovesan, S.C.E. Tosatto, Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures, Bioinformatics 34 (2017) 122–123.

[63] G. Parisi, J. Echave, Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes, Gene 345 (2005) 45–53.

[64] W.M. Fitch, M.O. Dayhoff, Atlas of protein sequence and structure, 1972, Syst. Zool. 22 (1973) 196.

[65] S.L. Pond, S.D. Frost, S.V. Muse, HyPhy: hypothesis testing using phylogenies, Bioinformatics 21 (2005) 676–679.

[66] W.G. Touw, C. Baakman, J. Black, T.A.H. te Beek, E. Krieger, R.P. Joosten, G. Vriend, A series of PDB-related databanks for everyday needs, Nucleic Acids Res. 43 (2015) D364–D368.

[67] J. Feisenstein, PHYLIP: Phylogeny Inference Package Version 3.2 Manual, 1989.

[68] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Contr. 19 (1974) 716–723.

[69] F.S. Guthery, K.P. Burnham, D.R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach, J. Wildl. Manag. 67 (2003) 655.