Natural protein sequences are more intrinsically disordered than random sequences

Cell. Mol. Life Sci. (2016) 73:2949–2957
DOI 10.1007/s00018-016-2138-9

Cellular and Molecular Life Sciences

ORIGINAL ARTICLE

Natural protein sequences are more intrinsically disordered than random sequences

Jia-Feng Yu[1] · Zanxia Cao[1,2] · Yuedong Yang[3] · Chun-Ling Wang[2] · Zhen-Dong Su[1] · Ya-Wei Zhao[1] · Ji-Hua Wang[1,2] · Yaoqi Zhou[1,3]

Received: 18 December 2015 / Revised: 10 January 2016 / Accepted: 11 January 2016 / Published online: 22 January 2016

Abstract Most natural protein sequences have resulted from millions or even billions of years of evolution. How they differ from random sequences is not fully understood. Previous computational and experimental studies of random proteins generated from noncoding regions yielded inclusive results due to species-dependent codon biases and GC contents. Here, we approach this problem by investigating 10,000 sequences randomized at the amino acid level. Using well-established predictors for protein intrinsic disorder, we found that natural sequences have more long disordered regions than random sequences, even when random and natural sequences have the same overall composition of amino acid residues. We also showed that random sequences are as structured as natural sequences according to contents and length distributions of predicted secondary structure, although the structures from random sequences may be in a molten globular-like state, according to molecular dynamics simulations. The bias of natural sequences toward more intrinsic disorder suggests that natural sequences are created and evolved to avoid protein aggregation and increase functional diversity.

✉ Yaoqi Zhou
yaoqi.zhou@griffith.edu.au

Ji-Hua Wang
jhw25336@126.com

[1] Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China

[2] College of Physics and Electronic Information, Dezhou University, Dezhou 253023, China

[3] Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr, Southport, QLD 4222, Australia

Keywords Random sequence · Protein intrinsic disorder · Secondary structure · Molten globule · Molecular dynamics simulation

Background

Proteins are linear polymeric chains made of a combination of 20 different types of amino acid residues. The total number of proteins explored by nature since the origin of life is estimated between 10^{21} and 10^{43} [1]. This number is infinitesimal compared to the number of possible protein sequences because the sizes of proteins can range from 2 to as long as 35,000 amino acid residues [2] and even for a small protein of 100 amino acid residues, the number of possible proteins with distinct sequences is 20^{100} or 10^{130}. The tiny sequence space explored by the nature raises an interesting question: if and how random-sequence proteins differ from natural proteins constrained by their functional and structural requirements? Investigating random sequences is also important because some proteins can arise suddenly from non-coding regions [3, 4]. Frame-shifting translation that produces random sequences after the insertion/deletion point was also proposed for the creation of novel proteins [5].

Artificial proteins with random sequences have been studied experimentally. Random co-polymerization of mixed amino-acid *N*-carboxyanhydrides was shown to produce compact structures similar to proteins [6, 7]. Random sequences of three residue types (Q, R, and L) of 70–90 amino acid residues were expressed in *E. coli* and shown to have secondary structures and cooperative unfolding [8]. Further studies indicate that random 120-amino-acid sequences of 20 residue types are aggregation-prone, and 12 residue-type sequences have better solubility [9]. Some soluble proteins are found to be

compact with some secondary structures. Chiarabelli et al. [10] showed that 20 % of 79 random 50-residue proteins are likely folded as they were protected from serine protease thrombin. Two of the selected proteins can reversibly fold and unfold. LaBean et al. [11] studied about 30 71-residue random-sequence proteins and found some with high secondary-structure contents with cooperative unfolding. These latest experimental studies suggested frequent appearance of native-like properties in random-sequence proteins. However, the sequences in these studies were obtained according to prescribed frequencies of DNA bases. They may not reflect natural usages of amino acid residues. In addition, codon usage bias in an expression system such as *E. coli* may provide additional biases toward proteins actually expressed. Furthermore, three-dimensional structures of these random-sequence proteins were not determined by either NMR or X-ray crystallography. In fact, other studies suggested the rare occurrence of stably folded or functional proteins. For example, only several functional proteins [12] resulted from initial 4×10^{12} random sequences followed by many iterations of in vitro selections and directed evolution [13]. No folded structures were yielded from in vitro random recombination of secondary structure elements (blocks) [14, 15].

Random-sequence proteins were also studied computationally, and two different views emerged. Some supported the view that natural sequences differ only slightly from random sequences [16]. For example, Weiss et al. [17] showed that random protein sequences have similar information content as non-redundant natural protein sequences. Crooks et al. [18] found that protein sequence-structure correlations based on mutual information in sequences of natural proteins can also be generated from random-sequence proteins. Lavelle and Pearson [19] investigated four- and five-amino-acid segments and found no significant biases between natural and random sequences. Angyan et al. [20] compared natural sequences to random protein sequences generated from random DNA sequences at various GC contents. They found that at 40–60 % GC contents, intrinsic disorder and aggregation propensity of translated random proteins are similar to those of natural proteins. By contrast, Pande et al. [21] showed that natural sequences have "pronounced deviations from pure randomness, directed toward minimization of the energy of the three-dimensional structure". Others supported significant difference between random and natural sequences by developing highly accurate two-state classifiers [22–24]. These computational studies, however, were limited mostly to comparing random-sequence proteins to either fully disordered proteins or fully structured proteins.

This paper presents a comparative study of structure and intrinsic disorder of natural and random-sequence proteins. We compared several structural properties of natural and random protein sequences: predicted intrinsic disorder by IUpred [25] and SPINE-D [26], predicted secondary structures by SPIDER 2 [27], and predicted tertiary structures by SPARKS-X [28]. A few selected model structures were simulated by molecular dynamics simulations. The comparison revealed that natural and random sequences have essentially the same structural properties except that the former have more long disordered regions, likely evolved to avoid detrimental aggregation.

Results

We constructed three databases of 10,000 protein sequences of 60 amino acid residues at 30 sequence identity cutoff (see Materials and methods). There are natural wild-type sequences (Pnat), random sequences generated according to natural occurrences of amino acid types (Prnd) and random sequences generated according to a fixed occurrence at 5 % for every amino acid type (Preq).

Figure 1 shows the number of protein sequences in number of disordered residues predicted by IUpred and SPINE-D for sequences in Pnat, Prnd, and Preq, respectively. Overall speaking, IUpred predicts more proteins with less number of disordered residues than SPINE-D, regardless of sequence datasets. This observation is

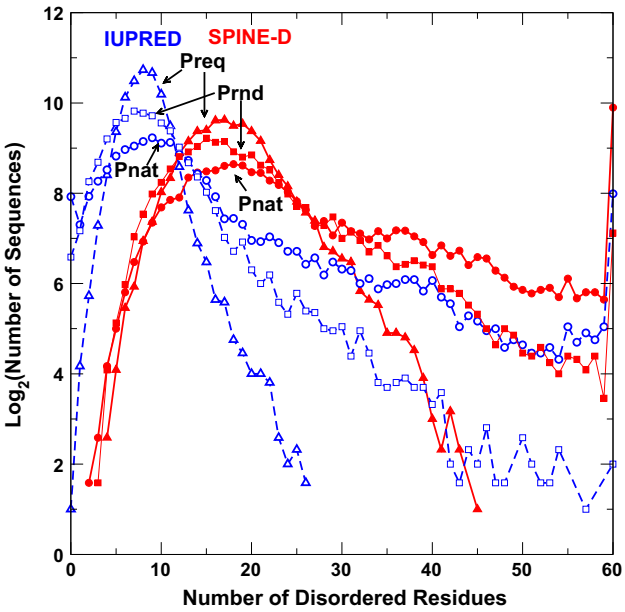

Fig. 1 The number of protein sequences (in log2) with a given number of disordered residues predicted by IUpred (in *blue*) and SPINE-D (in *red*) for three separate sequence datasets (natural sequences, Pnat in *circles*; random sequences with natural amino-acid frequencies, Prnd in *squares*; and random sequences with a fixed 5 % frequency for all residues, Preq in *triangles*). Natural sequences are more disordered than random sequences as predicted by either IUPRED or SPINE-D. Here all points with 0 occurrence are not shown

consistent with the fact that IUpred has a lower sensitivity than SPINE-D [26]. Nevertheless, IUpred and SPINE-D yield qualitatively similar trends for three sequence databases. That is, natural protein sequences contain less proteins having smaller number of disordered residues (5–26 for SPINE-D) but more proteins having higher number of disordered residues (27–60 for SPINE-D) than random sequences with or without fixing amino-acid compositions at 5 %. The distribution given by random sequences with natural occurrence of amino acid residues (Prnd) is closer to the distribution given by natural sequences (Pnat) rather than to that of random sequences with a fixed composition (Preq). It is of interest to note that natural sequences have more fully disordered proteins (60 residues long) and more fully structured proteins than random sequences although Pnat has only slight more nearly full-structured proteins (number of disordered residues $\leq$5). Based on SPINE-D, there are 59 natural sequences, 55 random sequences of natural compositions, and 24 random sequences of fixed compositions with $\geq$55 residues in structured regions. The same trend (more fully structured and more fully disordered proteins for natural sequences) is also observed by IUPRED.

To confirm that natural sequences have more nearly fully structured and fully disordered proteins, we re-examine the results based on largest continuous disordered or structured regions in Fig. 2. Here we randomly divided 10,000 sequences into five equal sets and obtained the average and standard deviations between five sets of sequences. For clarity, we showed the result from SPINE-D only as IUPRED gives the same trend. Figure 2a indicates that natural sequences have more long disordered regions than sequences in Prnd or Preq. The difference is larger than standard deviation. In particular, there are 954 fully disordered sequences for all 10,000 natural sequences but only 139 for random sequences with natural amino acid compositions and 0 for random sequences with fixed amino acid compositions.

While there is a large difference in three sequence databases for number of proteins with long disordered regions (Fig. 2a), the difference is not significant for number of proteins (within standard deviations) with long structured regions (>50 residues, Fig. 2b). Random sequences tend to have more sequences with structured regions between 40 and 50 residues. There are 58 proteins with $\geq$55 residues in a continuous structured region for Pnat, 55 for Prnd, and 24 for Preq. The small difference between 58 for natural sequences and 55 for random sequences with the same overall composition of amino acids suggests that natural sequences are only slightly or marginally more optimized than random sequences for full structured proteins.

To confirm the accuracy of predicted structured and disordered regions (defined by SPINE-D with a threshold at 0.5), we investigated composition bias (ΔP_i^o and ΔP_i^d) in structured

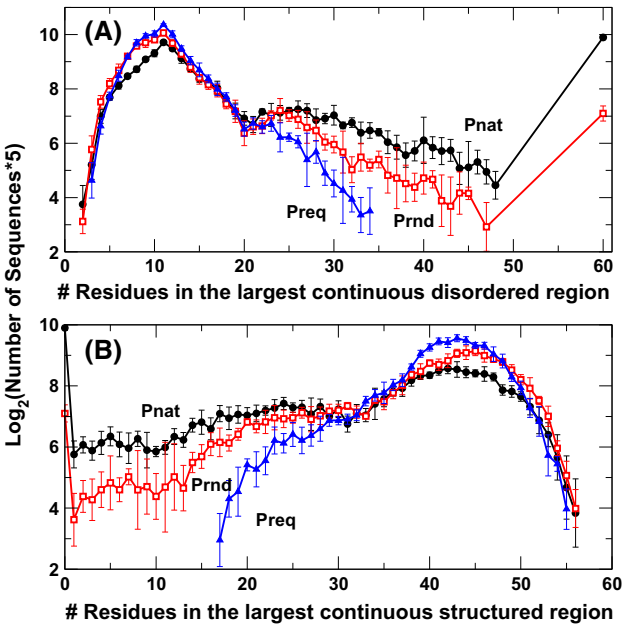

Fig. 2 The average number of protein sequences (times 5 in log2) as a function of the number of residues in the largest continuous disordered (**a**) or structured (**b**) regions for three separate sequence datasets (natural sequences, Pnat in *circles*; random sequences with native amino-acid frequencies, Prnd in *squares*; and random sequences with a fixed 5 % frequency for all residues, Preq in *triangles*) according to SPINE-D prediction. 10,000 sequences were randomly divided into five equal subsets. The averages and standard deviations are shown. Natural sequences have slightly more nearly fully structured proteins (>55 residues) than random sequences. Here, all points with 0 occurrence in any subsets are not shown

and intrinsically disordered regions and compared to annotated regions in the DisProt database [29]. Composition bias in predicted regions (ordered or disordered) by SPINE-D is highly similar to that in annotated regions for three separate sequence databases with high pairwise Pearson's correlation coefficients. For structured regions, the correlation coefficients to annotated regions are 0.75 for natural sequences, 0.91 for Prnd, and 0.90 for Preq, respectively. For intrinsically disordered regions, the correlation coefficients to annotated regions are 0.74 for natural sequences, 0.90 for Prnd, and 0.90 for Preq, respectively. Lower correlation coefficients of composition biases between natural sequences and annotated regions are likely because composition biases in random sequences play more important roles in disorder classification as a result of less informative sequence profiles from multiple sequence alignment than natural sequences.

Secondary structural contents predicted by SPIDER2 for three sequence datasets in structured and disordered regions are compared in Table 1. The difference is small but statistically significant (p value for unpaired t test <0.002 for all cases): Pnat has 3–7 % higher fraction of helical residues per protein (35.6 %) than Prnd (32.9 %) and Preq (28.3 %) but 7 % less sheet residues (22.6 %,

Table 1 The average helical and sheet contents in structured and disordered regions

Database	Structured		Disordered	
	Helix	Sheet	Helix	Sheet
Preq[a]	0.28 ± 0.20	0.29 ± 0.17	0.15 ± 0.17	0.10 ± 0.12
Prnd[a]	0.33 ± 0.24	0.30 ± 0.19	0.20 ± 0.19	0.12 ± 0.13
Pnat[a]	0.36 ± 0.29	0.23 ± 0.22	0.21 ± 0.19	0.09 ± 0.11
Pstruc[b]	0.37 ± 0.32	0.21 ± 0.18	–	–

[a] Based on predicted secondary structure by SPINE-D

[b] Based on actual secondary structure by DSSP

compared to 29.9 % for Prnd and 29.4 % for Preq) in the structured regions. All sequences in disordered regions have significantly (10 % or more) less helical and sheet residues than in structured regions. Table 1 also tabulated fractions of annotated helical and sheet residues in 110 non-redundant monomeric protein structures (Pstruc). Helical and sheet contents in Pstruc are similar to those in Pnat, confirming the overall accuracy of predicted secondary structures.

Figure 3 compares the length distribution of helices and sheets in Preq, Prnd, Pnat, and Pstruc in structured regions. The difference between Pnat and Prnd is small. This indicates that natural and random sequences (given the same overall compositions) have similar helical and sheet lengths. Similar distribution is observed for structured proteins (Pstruc) although the dataset is much smaller (110 vs. 10,000 sequences), suggesting that there is no evolutionary preference in lengths of helices and sheets in protein structures.

Figure 4 compares the length distribution of helices and sheets in Preq, Prnd, and Pnat in intrinsically disordered regions. Pnat has more long helices than Prnd and Preq. This is largely because Pnat has significantly more long continuously disordered regions (Fig. 2). However, the length distributions of sheets are much closer to each other, despite that Pnat has more proteins with long disordered regions.

Can random sequences have well-defined three-dimensional structures? We performed the fold recognition method SPARKS X [28] for all proteins with predicted structural regions of more than 54 residues (59 for Pnat, 55 for Prnd and 24 for Preq). SPARKS X is a method that

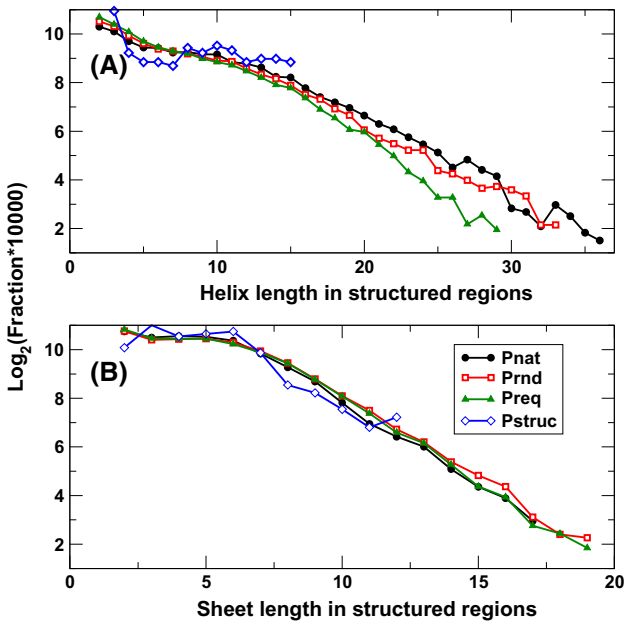

Fig. 3 The fraction of helices (**a**) and sheets (**b**) in a given length [$\log_2$ (fraction × 10,000)] in structured regions for four databases as labeled. To ensure statistics, the sizes of helices or sheets that appeared in less than five proteins in the dataset are not shown

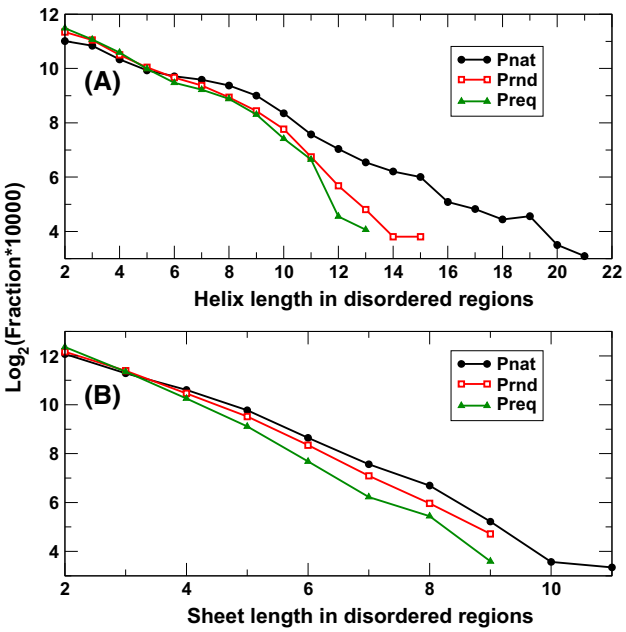

Fig. 4 The fraction of helices (**a**) and sheets (**b**) in a given length [$\log_2$ (fraction × 10,000)] in intrinsically disordered regions for three databases as labeled

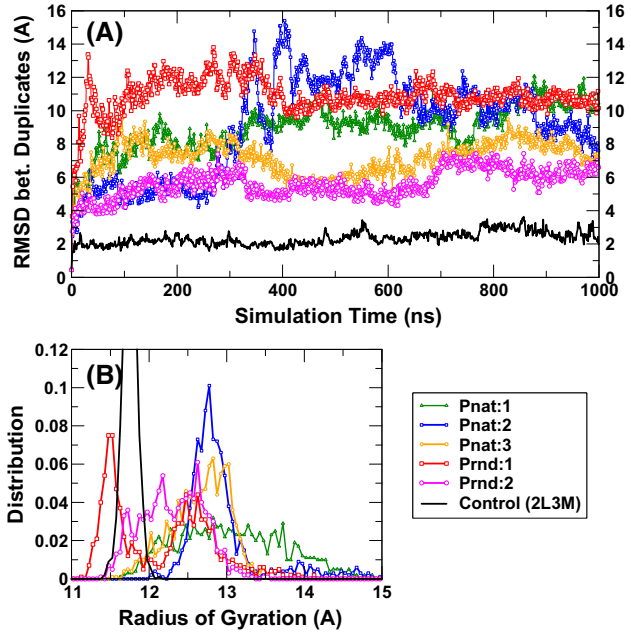

Fig. 5 **a** The RMSD between two conformations from two independent simulations as a function of simulation time for six proteins (Pnat: 1, 2, and 3 refer to UniRef50_M0WDE6, UniRef50_J9E0E9, and UniRef50_D6GUH9, respectively; Prnd: 1 and 2 refer to seq 08789 and seq 04514, respectively). **b** The distribution of radius of gyration for the last 50 ns of two duplicate simulations for each sequence

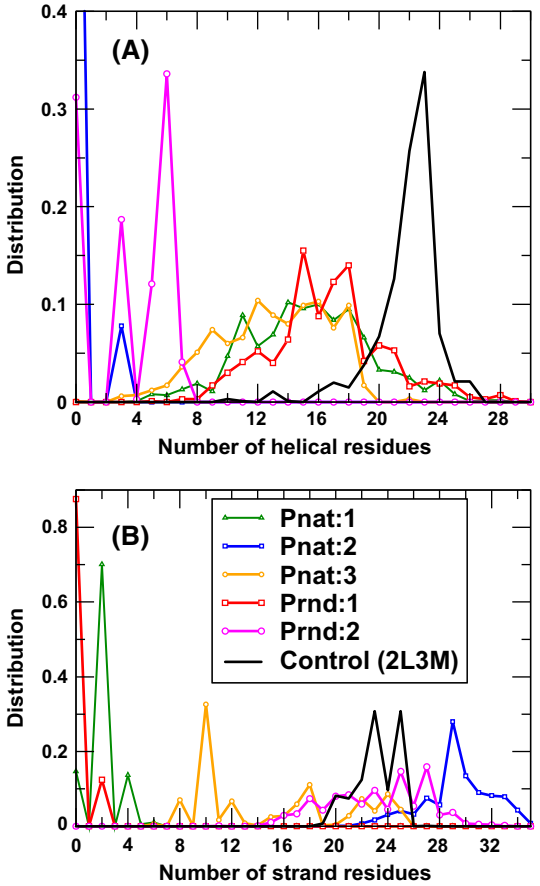

Fig. 6 **a** The distribution of the number of helical residues for the last 50 ns of two duplicate simulations for each of six proteins (Pnat: 1, 2, and 3 refer to UniRef50_M0WDE6, UniRef50_J9E0E9, and UniRef50_D6GUH9, respectively; Prnd: 1 and 2 refer to seq 08789 and seq 04514, respectively). **b** As in (a) but for the number of strand residues

attempts to map a query sequence of unknown structure to all known structures stored in the protein databank based on multi-dimensional matches of sequence and structural information. The significance of a match is measured by a *Z*-score with *Z*-score >7 suggesting a highly significant match. There are 25 out of 59 proteins with *Z*-score >7 for Pnat, two out of 55 for Prnd, and three out of 24 for Preq. Despite a similar number of proteins with predicted structural regions of more than 54 residues, Pnat has many more predicted proteins with quality predicted structures than Prnd. This is largely because natural sequences have more naturally occurring homologs or remote homologs.

We performed molecular dynamics simulations for two sequences from Prnd (Seq 08789 and Seq 04514 with *Z*-score = 7.41 and 7.07, respectively) and three sequences from Pnat (UniRef50_M0WDE6, UniRef50_J9E0E9, and UniRef50_D6GUH9 with *Z*-score = 9.24, 8.97, and 8.94, respectively). As a control, we also performed MD for one solution NMR structure of a putative copper-ion-binding protein from *Bacillus anthracis* str. Ames (PDB ID 2L3 M, 71 residues long). All models either from Pnat or from Prnd failed to have a stable structure after 100-ns simulations (6–10 Å RMSD from the starting conformations and 7–10 Å between two last conformations in duplicate simulations, Fig. 5a) while the PDB structure 2L3 M remains stable (2.8 and 2.6 Å RMSD, respectively, from the native conformation) after 100-ns simulation. Interestingly, only minor increases in radius of gyration were observed for Prnd (1 and 6 %) and Pnat sequences (−2, 3, and −6 %, respectively). The distributions of radius of gyration in the last 50 ns for all six pairs of simulations are shown in Fig. 5b. These results indicate that model structures are more flexible and slightly less compact than the native structure (2L3 M). Figure 6 further examines the distribution of amount of secondary structures (helical and sheet residues in Fig. 6a, b, respectively) in model structures as compared to the native structure (2L3 M). It is clear that the distributions of the numbers of helical and sheet residues are much narrower in native structures than in model structures. These results indicate that model structures are not accurate enough to confirm whether random sequences are capable of having unique structures by molecular dynamics simulations. Nevertheless, MD simulation results confirm that random sequences are capable of forming collapsed globule structures with some secondary structures.

Discussion

We have studied structure and disorder in 10,000 naturally occurring and random protein sequences by using current state-of-the-art techniques for prediction of protein intrinsic disorder, secondary structure, and tertiary structure. Based on intrinsic disorder prediction, natural sequences have many more disordered residues in long continuous regions but only marginally more nearly full-structured proteins than random sequences. In predicted structured regions, natural sequences have marginally higher helical residues but less sheet residues than random sequences with the same amino acid compositions. In predicted disordered regions, there is no significant difference in helical and sheet contents between natural and random sequences of the same amino acid compositions. The distributions of helical and sheet lengths for random and natural sequences follow essentially the same power-law distribution in the structured region. Although molecular dynamics simulations of a few selected model structures did not reveal stable conformations, these model structures remain highly compact, suggesting that these proteins (with random and natural sequences) at least are collapsed molten globules with some secondary structures.

Random protein sequences are nearly as structured or more structured than natural sequences. This finding, based on disorder prediction and prediction of secondary structure, is consistent with several experimental examinations of sequences from random co-polymerization of mixed amino-acid *N*-carboxyanhydrides [6, 7], random three residue types (Q, R, and L) of 70–90 amino acid residues [8], random 120-amino-acid sequences of 20 and 12 residue types [9], random 50-residue proteins [10], and random 71-residue proteins [11]. These experimental studies showed that random sequences have compact structures, cooperative unfolding, secondary structures, and/or protected from serine protease thrombin. The consistency between experimental and our computational studies occurs despite that experimental protein sequences were obtained at DNA levels, expressed in *E. coli* (i.e., subjected to codon optimization).

It should be noted, however, that SPINE-D [26] likely over-predicts structured regions because it cannot distinguish proteins in molten globule states (compact with some secondary structures [30]) from proteins in unique three-dimensional structures. This happens because only native structures and disordered regions were employed for training SPINE-D [26]. Indeed, long molecular dynamics simulations of predicted model structures of random sequences failed to produce a well-defined conformation. However, model structures of natural sequences also failed to have a well-defined conformation, suggesting that model inaccuracy is likely the main reason for unfolding of model structures in molecular dynamics simulations. If the majority of predicted structured regions are in a molten globule state, it explains the difficulty in producing folded structures from in vitro random recombination of secondary structure elements (blocks) [14, 15].

What is interesting is that natural sequences have more disordered residues and more long disordered regions with helical conformations. In a recent paper, we have shown that the fraction of order and semi-disorder (disorder probability <0.7) predicted by SPINE-D can be effectively employed to predict residues in aggregation prone regions with an accuracy comparable to several state-of-the-art techniques dedicated for aggregation prediction [31]. Thus, more disordered and long disordered regions for natural sequences indicate that natural sequences are created and evolved for solubility so as to avoid protein aggregation. This is consistent with the finding that random 120-amino-acid sequences of 20 residue types are aggregation-prone [9]. The existence of helical regions in long disordered regions indicates that nature may also employ disorder to enhance plasticity for function because helices in disordered regions are one of the widely utilized motifs in protein–protein interactions [32]. Disordered regions also provide accessibility of key residues for post-translational modifications, and serve as flexible linkers for separating functional domains or entropic bristles for keeping non-interacting molecules apart [33].

Materials and methods

Construction of protein sequence databases

Natural sequences in Pnat are obtained from the UniRef50 sequence database [34]. Its sequence redundancy was removed by using BLASTClust [35] with 30 % sequence identity cut-off. Sequence non-redundancy in Prnd and Preq was examined and confirmed by the program CD-HIT [36] with 30 % sequence identity cut-off. The natural occurrences of amino acid types were obtained from BLOSUM62 [37].

Intrinsic disorder prediction

The existence of intrinsic disorder in proteins (natural or artificial sequences) is probed by two different algorithms. One method is IUpred, which predicts disorder based on knowledge-based interaction strengths within sequentially neighboring amino acid residues [25]. IUpred is computationally fast because it does not require evolutionary information of protein sequences. Another method is

SPINE-D, which employs a neural network trained for disorder prediction [26]. SPINE-D provides a more accurate prediction of intrinsic disorder than IUpred and was independently assessed to be among the best-performing methods in the Critical Assessment of Structure Prediction techniques (CASP 9, 2010) [38]. It is more accurate because protein evolutionary information accounts for the fact that structured regions are more likely conserved than unstructured, intrinsically disordered regions. Comparing predictions between IUpred and SPINE-D will allow us to evaluate the consistency in computational predictions in the presence and absence of sequence evolution information.

Protein secondary-structure prediction

A recently developed method SPIDER2 [27] was employed to predict secondary structure by iterative deep learning of multiple structural properties (backbone torsion angle, solvent accessible surface area, and Calpha angles) in addition of secondary structure. It was chosen because it is one of the most accurate predictors of secondary structures.

Protein secondary structure analysis

For comparison, we also obtained structured proteins (Pstruc) with 3.5-Å resolution or better and sequence lengths between 50 and 70 amino acid residues from the protein databank. We further removed protein structures that are in complex with RNA, DNA, or proteins. The final dataset (Pstruc) contains 110 proteins after removing redundancy at 30 % sequence-identity cut-off. The secondary structures of these proteins were obtained from the PDBfinder database [39]. Eight-state annotations were merged into three states [H, G, and I for Helix (H), B and E for sheet (E), T, S, and D for Coil (C)].

Amino acid preferences

We evaluated the preferences of amino acid residues in ordered or intrinsically disordered regions by examining the difference of their occurrence in the region (P_i^o, P_i^d) from their occurrence in all sequences in the database (P_i^{all}) [40]. That is, $\Delta P_i^o = (P_i^o - P_i^{all})/P_i^{all}$ and $\Delta P_i^d = (P_i^d - P_i^{all})/P_i^{all}$ in addition to calculating amino acid preferences from predicted ordered and disordered regions, we also calculated amino acid preferences in annotated structured and disordered regions by using the DisProt database [29]. A total of 548 annotated sequences were obtained from the DisProt database after removing redundancy by using CD-HIT (30 % sequence identity cut-off). These sequences contain 911 intrinsically disordered regions and 978 structured regions.

Structure prediction and molecular dynamics simulations

For those random sequences predicted to be structured, we performed template-based structure prediction by SPARKS X with default parameters [28]. Selected model structures are then simulated in the presence of water molecules. Molecular dynamics (MD) simulation in the isothermal-isobaric (NPT) ensemble was performed using the GROMACS 4.6.2 software package [41]. We employed the amber99sb-ildn force field for proteins and TIP3P for water molecules [42]. The protein was solvated in a truncated octahedron box with the minimum solute-box boundary distance being set to 12 Å. The long-range electrostatic interaction was treated with the particle-mesh Ewald method with a grid spacing of 1.2 Å and a fourth-order interpolation [43, 44]. Protonation states of ionizable groups were chosen for pH = 7.0. For each protein, two independent simulations were performed for 100 ns with different initial velocities for pressure P at 1 bar and temperature T at 298K. The temperature of the system was kept constant by velocity rescaling with a stochastic term [45]. The pressure of the system was kept constant by using the Berendsen algorithm [46]. The simulation employed a temperature coupling time of 0.1 ps and pressure coupling time of 2 ps. The time step for the MD integrator was set to 2 fs and LINCS [47] was applied to constrain all bond lengths.

Availability of data and materials

All sequence datasets (Pnat, Prnd, and Preq) are made available at http://sparks-lab.org.

Acknowledgments This work is supported by National Natural Science Foundation of China (61271378) to J. W., J. Y., and Y. Y., National Natural Science Foundation of China (61302186) to J. Y., the Microsoft Azure for Research Awarded to Y. Y., the Taishan Scholars Program of Shandong province of China, the National Health and Medical Research Council (1059775 and 1083450) of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (Project Number LE150100161) to Y. Z. We also gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing financial interests.

References

1. Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on Earth? J R Soc Interface 5(25):953–956. doi:10.1098/rsif.2008.0085
2. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fedotov A, Foulger R, Garavelli J, Huntley R, Jacobsen J, Kleen M, Laiho K, Leinonen R, Legge D, Lin Q, Liu WD, Luo J, Orchard S, Patient S, Poggioli D, Pruess M, Corbett M, di Martino G, Donnelly M, van Rensburg P, Bairoch A, Bougueleret L, Xenarios I, Altairac S, Auchincloss A, Argoud-Puy G, Axelsen K, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Ciapina L, Coral D, Coudert E, Cusin I, Delbard G, Doche M, Dornevil D, Roggli PD, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Farriol-Mathis N, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kappler T, Keller G, Lachaize C, Lane-Guermonprez L, Langendijk-Genevaux P, Lara V, Lemercier P, Lieberherr D, Lima TD, Mangold V, Martin X, Masson P, Moinat M, Morgat A, Mottaz A, Paesano S, Pedruzzi I, Pilbout S, Pillet V, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Yip LN, Zuletta L, Wu C, Arighi C, Arminski L, Barker W, Chen CM, Chen YX, Hu ZZ, Huang HZ, Mazumder R, McGarvey P, Natale DA, Nchoutmboube J, Petrova N, Subramanian N, Suzek BE, Ugochukwu U, Vasudevan S, Vinayaka CR, Yeh LS, Zhang J, Consortium U (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142–D148. doi:10.1093/Nar/Gkp846
3. Neme R, Tautz D (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genom 14:117. doi:10.1186/1471-2164-14-117
4. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M (2012) Proto-genes and de novo gene birth. Nature 487(7407):370–374. doi:10.1038/nature11184
5. Okamura K, Feuk L, Marques-Bonet T, Navarro A, Scherer SW (2006) Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics 88(6):690–697. doi:10.1016/j.ygeno.2006.06.009
6. Rao SP, Carlstrom DE, Miller WG (1974) Collapsed structure polymers. A scattergun approach to amino acid copolymers. Biochemistry 13(5):943–952
7. Anufrieva EV, Bychkova VE, Krakovyak MG, Pautov VD, Ptitsyn OB (1975) A synthetic polypeptide with a compact structure and its self-organization. FEBS Lett 55(1):46–49
8. Davidson AR, Sauer RT (1994) Folded proteins occur frequently in libraries of random amino acid sequences. Proc Natl Acad Sci USA 91(6):2146–2150
9. Tanaka J, Doi N, Takashima H, Yanagawa H (2010) Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Protein Sci 19(4):786–795. doi:10.1002/pro.358
10. Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL (2006) Investigation of de novo totally random biosequences, part II: on the folding frequency in a totally random library of de novo proteins obtained by phage display. Chem Biodivers 3(8):840–859. doi:10.1002/cbdv.200690088
11. Labean TH, Butt TR, Kauffman SA, Schultes EA (2011) Protein folding absent selection. Genes 2(3):608–626. doi:10.3390/genes2030608
12. Lo Surdo P, Walsh MA, Sollazzo M (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. Nat Struct Mol Biol 11(4):382–383. doi:10.1038/Nsmb745
13. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410(6829):715–718
14. Graziano JJ, Liu WS, Perera R, Geierstanger BH, Lesley SA, Schultz PG (2008) Selecting folded proteins from a library of secondary structural elements. J Am Chem Soc 130(1):176–185. doi:10.1021/Ja074405w
15. Tsuji T, Onimaru M, Doi N, Miyamoto-Sato E, Takashima H, Yanagawa H (2009) In vitro selection of GTP-binding proteins by block shuffling of estrogen-receptor fragments. Biochem Bioph Res Commun 390(3):689–693. doi:10.1016/J.Bbrc.2009.10.029
16. Ptitsyn OB (1985) Random sequences and protein folding. Theochem J Mol Struc 24(1–2):45–65
17. Weiss O, Jimenez-Montano MA, Herzel H (2000) Information content of protein sequences. J Theor Biol 206(3):379–386. doi:10.1006/Jtbi.2000.2138
18. Crooks GE, Wolfe J, Brenner SE (2004) Measurements of protein sequence-structure correlations. Proteins 57(4):804–810. doi:10.1002/prot.20262
19. Lavelle DT, Pearson WR (2010) Globally, unrelated protein sequences appear random. Bioinformatics 26(3):310–318. doi:10.1093/bioinformatics/btp660
20. Angyan AF, Perczel A, Gaspari Z (2012) Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett 586(16):2468–2472. doi:10.1016/j.febslet.2012.06.007
21. Pande VS, Grosberg AY, Tanaka T (1994) Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? Proc Natl Acad Sci USA 91(26):12972–12975
22. Munteanu CR, Gonzalez-Diaz H, Borges F, de Magalhaes AL (2008) Natural/random protein classification models based on star network topological indices. J Theor Biol 254(4):775–783. doi:10.1016/j.jtbi.2008.07.018
23. Teraguchi S, Patil A, Standley DM (2010) Intrinsically disordered domains deviate significantly from random sequences in mammalian proteins. BMC Bioinformatics 11(Suppl 7):S7. doi:10.1186/1471-2105-11-S7-S7
24. De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G (2012) Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. PLoS One 7(5):e36634. doi:10.1371/journal.pone.0036634
25. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434. doi:10.1093/bioinformatics/bti541
26. Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y (2012) SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struc Dyan 28(4):799–813
27. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang JH, Sattar A, Yang YD, Zhou YQ (2015) Improving prediction of secondary structure, local backbone angles, and solvent

accessible surface area of proteins by iterative deep learning. Sci Rep 5:11476. doi:10.1038/srep11476
28. Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. Bioinformatics 27:2076–2082
29. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the database of disordered proteins. Nucleic Acids Res 35(Database issue):D786–D793. doi:10.1093/nar/gkl893
30. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11(4):739–756. doi:10.1110/Ps.4210102
31. Zhang T, Faraggi E, Li Z, Zhou Y (2013) Intrinsically semi-disordered state and its role in induced folding and protein aggregation. Cell Biochem Biophys 67(3):1193–1205. doi:10.1007/s12013-013-9638-0
32. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN (2006) Analysis of molecular recognition features (MoRFs). J Mol Biol 362(5):1043–1059. doi:10.1016/j.jmb.2006.07.087
33. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41(21):6573–6582
34. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932. doi:10.1093/bioinformatics/btu739
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2
36. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. doi:10.1093/bioinformatics/btl158
37. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
38. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A (2011) Evaluation of disorder predictions in CASP9. Proteins 79(S10):107–118. doi:10.1002/prot.23161
39. Hooft RW, Sander C, Scharf M, Vriend G (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. Comput Appl Biosci 12(6):525–529
40. Yu JF, Wu ES, Wang CL, Wang HM, Wang JH (2016) Classification of ordered/disordered regions of intrinsically disordered proteins based on comprehensive sequence analysis and Chou's pseudo amino acid composition method. MATCH Commun Math Computer Chem 75:417–430
41. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. J Comput Chem 26(16):1701–1718. doi:10.1002/jcc.20291
42. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78(8):1950–1958. doi:10.1002/prot.22711
43. Darden T, York D, Pedersen L (1993) Particle mesh Ewald—an N.Log(N) method for Ewald Sums in large systems. J Chem Phys 98(12):10089–10092. doi:10.1063/1.464397
44. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103(19):8577–8593. doi:10.1063/1.470117
45. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126(1):014101. doi:10.1063/1.2408420
46. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR (1984) Molecular-dynamics with coupling to an external bath. J Chem Phys 81(8):3684–3690. doi:10.1063/1.448118
47. Hess B (2008) P-LINCS: a parallel linear constraint solver for molecular simulation. J Chem Theory Comput 4(1):116–122. doi:10.1021/ct700200b