Biochimica et Biophysica Acta 1844 (2014) 1827–1834

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Does Lack of Secondary Structure Imply Intrinsic Disorder in Proteins? A Sequence Analysis

Pooja Rani, Anupaul Baruah, Parbati Biswas *

Department of Chemistry, University of Delhi, Delhi-110007

ARTICLE INFO

Article history:
Received 26 April 2014
Received in revised form 28 July 2014
Accepted 31 July 2014
Available online 8 August 2014

Keywords:
Intrinsically disordered proteins
Secondary structure
Sequence complexity
Sequence composition
Charge
Hydrophobicity

ABSTRACT

Intrinsically disordered proteins (IDPs)/protein regions (IDPRs) lack unique three-dimensional structure at the level of secondary and/or tertiary structure and are represented as an ensemble of interchanging conformations. To investigate the role of presence/absence of secondary structures in promoting intrinsic disorder in proteins, a comparative sequence analysis of IDPs, IDPRs and proteins with minimal secondary structures (less than 5%) is required. A sequence analysis reveals proteins with minimal secondary structure content have high mean net positive charge, low mean net hydrophobicity and low sequence complexity. Interestingly, analysis of the relative local electrostatic interactions reveal that an increase in the relative repulsive interactions between amino acids separated by three or four residues lead to either loss of secondary structure or intrinsic disorder. IDPRs show increase in both local negative-negative and positive-positive repulsive interactions. While IDPs show a marked increase in the local negative-negative interactions, proteins with minimal secondary structure depict an increase in the local positive-positive interactions. IDPs and IDPRs are enriched in *D*, *E* and *Q* residues, while proteins with minimal secondary structure are depleted of these residues. Proteins with minimal secondary structures have higher content of *G* and *C*, while IDPs and IDPRs are depleted of these residues. These results confirm that proteins with minimal secondary structure have a distinctly different propensity for charge, hydrophobicity, specific amino acids and local electrostatic interactions as compared to IDPs/IDPRs. Thus we conclude that lack of secondary structure may be a necessary but not a sufficient condition for intrinsic disorder in proteins.

1. Introduction

Intrinsically disordered proteins (IDPs) lack stable tertiary structures and may be represented as a dynamic ensemble resembling unfolded proteins under physiological conditions *in vitro* [1–3]. IDPs may be either extended like a random coil or collapsed as a molten globule [1,4]. Interestingly, thousands of the structures deposited in the RCSB PDB consists intrinsically disordered proteins out of which 70% of the disordered proteins belong to the terminal regions i.e. 30 residues from each end which consists of 30% of all residues in PDB [5]. In some cases, local regions of the sequence adopt fluctuating elements of secondary structures but absence of secondary structures may or may not necessarily imply presence of disorder in proteins. Disordered proteins exhibit a multitude of diverse functions, especially in transcriptional regulation, translational control, cell cycle control and cellular signal transduction [6–9]. The prevalence of this structural disorder is implicated in diseases like cancer, cardio-vascular diseases, neurodegenerative diseases etc. due to their propensity to form misfolded amyloid like aggregates [6,10–12]. Due to high structural plasticity and large surface accessibility, IDPs interact efficiently with different targets such as other proteins or nucleic acids as compared to the globular proteins [13,14].

Globular proteins are characterized by sufficient number of contacts to compensate the loss of conformational entropy [15]. The lack of ordered structure in "Natively unfolded proteins" may be related to the intrinsic properties of amino acid residues [16]. Recent studies show that IDPs are characterized by a low content of hydrophobic amino acids and a preponderance of polar and charged residues, low secondary structure content, low sequence complexity with a biased composition of amino acid residues and high predicted flexibility [7,17–20]. Typically, some of these proteins may undergo disorder-order transition and fold partially upon binding to specific target molecules like ligands, co-factors, different substrates and other proteins/nucleic acids. This implies that the disorder may be encoded in the specific amino acid sequence of IDPs, which are significantly different from those of globular proteins with well-defined native structures [7,21–23]. Efforts in this direction have culminated in the development of various types of sequence-based disorder predictors [24–28], which can identify local disorder, where certain regions of the protein lack stable tertiary structure, from the global one characterized by a complete lack of ordered structures throughout the entire protein.

Abbreviations: IDPs, Intrinsically disordered proteins; IDPRs, Intrinsically disordered protein regions; PDB, Protein data bank; KL, Kullback-Leibler

* Corresponding author.
E-mail address: pbiswas@chemistry.du.ac.in (P. Biswas).

http://dx.doi.org/10.1016/j.bbapap.2014.07.020
1570-9639/

In this article, the sequence analysis of four different data sets of proteins with varying degree of disorder is presented. Globular proteins have well-defined structures with least disorder. Proteins comprising of partially disordered regions (IDPRs), completely disordered (IDPs) proteins and proteins with minimal secondary structure content (less than 5%) constitute three different groups respectively. The globular proteins, partially disordered proteins and proteins with < 5% secondary structures are compiled from RCSB Protein Data Bank (PDB), while completely disordered proteins are selected from the DisProt database. The sequence characteristics of proteins with minimal secondary structures are compared to globular and disordered proteins. The sequence complexity, mean hydrophobicity, mean net charge, compositional bias and local electrostatic interactions for each group of proteins are examined. Specific roles of particular amino acid residues causing loss of secondary structures and/or imparting disorder in proteins are discussed. The results suggest that the amino acids *D*, *E* and *Q* are abundant in partially and completely disordered proteins, but are scarcely present in proteins with minimal secondary structures. Proteins with minimal secondary structures are abundant in *G* and *C*. An increase in the local repulsive interactions between amino acids separated by 3 or 4 residues along the sequence leads to intrinsic disorder or loss of secondary structure. It is found that IDPs may be characterized by higher local negative-negative interactions while increased local positive-positive interactions imply loss of secondary structure. IDPRs show increase in both local negative-negative and positive-positive repulsive interactions compared to globular proteins. The analysis of different physico-chemical characteristics reflects that the proteins with minimal secondary structure content comprise of an independent group of proteins which is distinctly different from IDPRs, IDPs and globular proteins. Hence, it is concluded that lack of secondary structure may be necessary but not a sufficient condition for protein disorder.

2. Materials and methods

2.1. Database selection

2.1.1. Globular proteins

A database of globular proteins with X-ray crystallographic structures comprising of monomeric chains without any ligand is compiled from the Protein Data Bank (PDB) [29]. Monomeric chains are chosen to avoid the possibility of disorder-order transition of some multimeric proteins due to complex formation [18]. All protein chains with missing residues are removed from the selected data set to ensure that the coordinates of all atoms, except the hydrogen atoms, are available. A sequence similarity cut-off value of ≤ 25% and length ≥ 40 residues is applied on 6721 proteins using PISCES server [30]. The resultant non-redundant data set of globular proteins consists of 1917 monomeric protein chains. Apart from the database of globular proteins, three additional data sets of proteins are chosen for sequence analysis:

2.1.2. Group I

A data set of protein chains containing intrinsically disordered protein regions (IDPRs) is selected from PDB. This data set consists of proteins with missing residues in the electron density map of their X-ray crystallographic structures, where the disordered regions comprises of ≥ 3 consecutive sequence residues are listed in the remark 465 in PDB. A non-redundant data set of 9508 proteins is obtained from the compiled database using PISCES server with selection criteria of sequence similarity ≤ 25% and length ≥ 40 residues. The resultant data set is classified into 5 bins depending upon the degree of structural disorder in the available crystal structures. The degree of structural disorder is calculated as the percentage of structural disorder with respect to the total length of the protein sequence. These 5 bins are as follows: (i) Upto 10%, (ii) 10 to 20%, (iii) 20 to 35%, (iv) 35 to 50% and (v) 50 to 98.5% disorder.

2.1.3. Group II

A set of 109 Intrinsically Disordered Proteins (IDPs) is selected from DisProt, i.e., Database of Protein Disorder [31] (Release 6.02) database. These proteins are completely disordered and have no available crystal/NMR structures. A sequence similarity cut-off value of ≤ 25% and length ≥ 40 residues is applied on these proteins using PISCES server. Final data set of IDPs comprises of 91 proteins.

2.1.4. Group III

A data set of 324 protein chains with < 5% secondary structure content having at least 40 residues whose coordinates are well-defined in their respective structures is selected from the PDB. These sequences are submitted to PISCES server with the selection criteria of sequence similarity ≤ 25% and length ≥ 40 residues. The resultant data set of Group III Proteins consists of 83 protein chains.

The PDB ID's for the globular proteins, Group I proteins and Group III proteins and DisProt ID's for Group II proteins are provided in the Supplementary material. The length distribution of all four groups of proteins is shown in Fig. S2 of the Supplementary material.

2.2. Charge-hydrophobicity

Intrinsically disordered proteins are characterized by a unique combination of low mean hydrophobicity and high net charge per residue and occupy a unique region in the charge-hydrophobicity space, separated from those of globular proteins by a linear boundary [7]. At a given pH, the net charge, *Q* for a given protein may be calculated as [32,33]

$$Q = \sum Q^- + \sum Q^+ \tag{1}$$

where Q^- (for - COOH, - SH, - PhOH functional groups) and Q^+ (for $-NH_3^+$, $= NH^+_2$, $\equiv NH^+$) are the negative and positive charges of a residue respectively, which are calculated as

$$Q^- = \frac{(-1)}{1 + 10^{-(pH - pK_a)}} \tag{2}$$

$$Q^+ = \frac{(+1)}{1 + 10^{+(pH - pK_a)}} \tag{3}$$

The respective pK_a values used in this work for the ionizable groups are given in Table 1 [32]. Although the pK_a values for the *N* and *C*-terminus vary from one peptide to another, here we have used averaged pK_a values of the *N* and *C*-termini for the given data set of proteins. Mean net charge per residue is obtained from the net charge divided by the total number of amino acid residues present in the protein sequence.

Kyte and Doolittle approximation [34] is used for the calculation of mean net hydrophobicity. Hydrophobicity of each residue is first normalized on a 0 to 1 scale [7] and the normalized hydrophobicity for each segment is calculated by using a sliding window of 5 residues. The values are averaged over all windows to obtain the mean net hydrophobicity.

Table 1
pK_a values for ionizable residues [32].

Residue	pK_a values
ASP	3.9
GLU	4.35
HIS	6.5
TYR	9.9
LYS	10.35
ARG	12.5
N-terminus	8.5
C-terminus	3.3