## Disorder in Proteins

**Aily Sanjeev**, Tezpur University, Tezpur, Assam, India  
**Sandeep Kaushik**, Headquarters of the European Institute of Excellence on Tissue Engineering and Regenerative Medicine, Guimaraes, Portugal  
**Biswa R Meher**, Berhampur University, Berhampur, Odisha, India  
**Venka SK Mattaparthi**, Tezpur University, Tezpur, Assam, India

© 2018 Elsevier Inc. All rights reserved.

| 1 | Background | 1 |
| 1.1 | Intrinsically Disordered Proteins | 1 |
| 1.2 | Intrinsically Disordered Protein Regions | 1 |
| 1.3 | Disordered Proteins and Disease | 2 |
| 2 | Short Motifs Found in Disordered Regions and the Sequence-Based Bioinformatics | 3 |
| 3 | Structure-Based Computational Approaches for Modeling and Simulation of Disordered Regions | 4 |
| 4 | Emerging Computational Tools and Techniques | 5 |
| 5 | Future Perspectives and Closing Remarks | 6 |
| Acknowledgement | 7 |
| References | 7 |

### 1 Background

#### 1.1 Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) are important cellular components that play significant roles in various processes such as cell signaling, translation, transcription, and cell cycle regulation. Disorder in a polypeptide structure allows the same polypeptide to assume diverse interactions with different consequences (Dunker et al., 2008; Dyson and Wright, 2005). IDPs are distinguished based on their low sequence complexity and biased amino acid composition: they typically contain a high proportion of charged and hydrophilic amino acids and a low proportion of bulky hydrophobic amino acids. These IDPs undertake transitions of disorder-to-order state on binding to the partners though they do not contain a stable tertiary structure (Uversky, 2013; Wright and Dyson, 2009). These proteins are not able to fold into well-defined, globular three-dimensional structures spontaneously even though they are perfectly functional. But they have the ability to rapidly fluctuate through a range of conformations that can cover a continuous conformational space that range from the extended statistical coils to collapsed globules and are dynamically disordered (Dyson and Wright, 2005).

IDPs also play a significant role in the ordered assembly of macromolecular machines such as ribosomes, chromatin organization, assembly and disassembly of microfilaments and microtubules, nuclear pore transportation, binding and small molecule transportation, and functional protein domain separation (Frey et al., 2006; Guharoy et al., 2013; Tompa, 2005, 2010). In addition to IDPs, some proteins contain some disordered sequences known as Intrinsically Disordered Regions (IDRs). IDRs are known to interact with and function as hubs in the various protein-interaction networks (Dunker et al., 2005; Kim et al., 2008). IDRs get subjected to various processes such as alternative splicing of their mRNA and combinatorial Post-Translational Modifications (PTM) that can add complexity to the regulatory networks and provide a tissue-specific signaling mechanism. Abundance of these IDRs are tightly being regulated that are related to the disease for ascertaining the signaling both in time and space and also for mutation in the IDPs (Ilbas et al., 2011; Gsponer et al., 2008; Vavouri et al., 2009). In the same manner, various folded proteins can control the transitions of order-to-disorder state that can mediate the various biological functions (Mitrea and Kriwacki, 2013; Schultz and Natarajan, 2013). IDPs are also known to take part in the signaling complexes assembly and in the self-assembly of dynamic cytoplasmic and membrane-less nuclear organelles. However the PTMs can control the IDPs function leading to structural changes (folding). The role of PTM is to stabilize or destabilize the individual secondary structural elements in the various IDPs (Espinoza-Fonseca et al., 2008; Pufall et al., 2005; Theillet et al., 2012). IDPs also have the ability to bind to multiple interaction partners transiently in the dynamic regulatory networks that can accurately respond to cellular signals and possess potential for complex information processing when performing the signaling functions (Stein et al., 2009). The interactions of IDPs with their binding partners are controlled by the PTMs allowing them to work as switches and rheostats (Borg et al., 2007; Dyson and Wright, 2005; Gsponer and Babu, 2009; Lee et al., 2010; Van Roey et al., 2012, 2013). IDPs are known to exchange their binding partners for which they compete for the central hub of the protein that is present in very limited amount. The disordered regions in signaling and regulatory proteins however consist of many conserved sequence motifs which can interact with nucleic acids or other proteins (Dyson and Wright, 2002, 2005).

#### 1.2 Intrinsically Disordered Protein Regions

Various neurodegenerative diseases such as Parkinson’s disease and prion diseases have been linked with intrinsically disordered proteins. Also many glutamine-rich motifs transcription factors are involved in polyglutamine diseases (McCampbell and Fischbeck, 2001).
Even after binding to the targets (Baker et al., 2007; Mittag et al., 2010), some IDPs remain disordered and form fuzzy complexes (Tompa and Fuxreiter, 2008). In addition disordered dynamic sites on distinct, non-overlapping surfaces of the target protein have been observed (Ferreon et al., 2009; Ishiyama et al., 2010). These binding interactions mediate pathway crosstalk via the formation of ternary complexes and also enhance the target-binding affinity with various binding partners that can modulate allosteric interactions (Ferreon et al., 2013). These proteins consist of multiple interaction motifs that can mediate binding to a diversity of targets which may act as 'central hubs' in signaling networks (Dunker et al., 2005; Kim et al., 2008). They can facilitate the assembly of both ternary and higher-order complexes through multiple sites and also integrate diverse signaling pathways. These sites allow allosteric responses in biological signaling, and the large entropy variations between the free and bound states can act to tune the energetics of the binding process (Flock et al., 2014; Hilser and Thompson, 2007; Motlagh et al., 2014; Nussinov and Tsai, 2013). In addition to fully disordered proteins, a large number of proteins present in various organisms contain polypeptide segments that do not form a stable 3D structure but are however functional (Daughdrill et al., 1997; Dunker et al., 2008; Dyson and Wright, 2005; Fink, 2005; Gsponer and Babu, 2009; Kriwacki et al., 1996; Tompa, 2002; Uversky et al., 2000; Uversky and Dunker, 2010; Ward et al., 2004; Wright and Dyson, 1999). In comparison to the original view of the globular and classical structure-function method, the functionality of the IDRs results in a very unique approach (Li et al., 2001, 2012a,b; Romero et al., 2001; Weatheritt et al., 2014). For example, these regions are required for the initiation of programmed necrosis through the formation of heterodimeric amyloid fibril (Li et al., 2012a,b). Recent reports suggest that disordered sequences of the RHIM (RIP homotypic interaction motif) are buried in the inactive state but on kinase activation they become exposed to amyloid, leading to necroosome formation (Li et al., 2012a,b).

IDPs have a greater impact on the constitutive pre-mRNA splicing and alternative splicing catalyzed by a dynamic ribonucleoprotein machine called the spliceosome. The components of this body are highly enriched in intrinsically disordered regions (Korneta and Bujnicki, 2012) that are considered as the spliced segments of the protein substrates alternatively (Romero et al., 2006) (‘alternatively spliced’).

### 1.3 Disordered Proteins and Disease

IDPs and IDRPs were considered as the two main variants of protein consisting of varying physico-chemical and functional properties. This arises mainly due to the varying percentage of disordered regions they contain. The IDRPs are the protein sequence that consists of regions which are locally disordered and they are featured by the existence of disordered domains in structure that are folded globally. These IDPs also contain some disordered regions in the relevant percentage of their residues.

The representation of the differential distribution of IDRPs in the Homo sapiens and in proteins annotated as disease-related, as a function of the minimal length of the disordered regions is depicted in Fig. 1. From the figure, the overall percentages of IDRPs present in the various diseases were depicted as a function of their disordered protein length in the case of Homo sapiens. We noticed among all the various diseases, the percentage of proteins in the IDRPs were higher in cases of calcium homeostasis and cancer. In cases of calcium homeostasis the range of IDRPs exceed by 46% while in the whole human proteome it is about 26%. Although further investigation is required for the remarkable case of IDRPs enrichment that is penetrated by the calcium homeostasis.

![Fig. 1](image-url) The varying distribution of the overall IDRPs percentage in the Homo sapiens represented in diseases as a function of disordered protein length.
The occurrence of IDPs in the human proteome and genetic diseases is almost same. As the IDPs prevail in various diseases, it can be suggested that there is a lack of three-dimensional structure due to which the diseases occur reported earlier (Uversky et al., 2008, 2009). But the development of diabetes is not influenced by the prevalence of IDPs as studied earlier (Uversky et al., 2008, 2009).

The occurrence of IDPs in the various diseases is more when compared to the human proteome. But this fact cannot illustrate that IDPs are necessary for the occurrence of diseases as the maximum proteins are not IDPs connected to the various diseases. However IDPs are more common than the other proteomes. Hence we can infer that in the development of cancer and neuro-degenerative diseases, the absence of a three dimensional structure play a significant role while it have no such effective role on the other diseases. In case of calcium homeostasis, though the protein is well structured with flexible loops but this flexibility is necessary for the development of the diseases (Lu et al., 2015; Vacic et al., 2012; Vacic and Iakoucheva, 2012).

In regard to the presence of mutations, there exists a vast difference between the ordered and disordered regions (Lu et al., 2015; Vacic et al., 2012; Vacic and Iakoucheva, 2012). The illustration of the effect of diseased mutations on the ordered and disordered regions referred from Vacic and Iakoucheva is depicted in Fig. 2.

The IDPs can contribute to the signaling complexes assembly and also membrane-less nuclear and cytoplasmic organelles self-assembly. To identify and characterize the various disordered protein regions several experimental, bioinformatics and computational analysis are being performed that result in higher evolution of the biological processes (Wright, 2015).

A typical example of IDP is the α-synuclein protein which is the principal component for the occurrence of debilitating disease, Parkinson’s disease (PD). A schematic representation of the 3D structure of α-synuclein protein is depicted in Fig. 3.

2 Short Motifs Found in Disordered Regions and the Sequence-Based Bioinformatics

Currently the Protein Data Bank (PDB) consists of about 100,000 deposited structures that are noticed to be accumulated for the past decades (Berman et al., 2013). But almost all the structures have been discovered till now. About 33% proteome present in the eukaryotes are noticed to possess putative long disordered segments (Ward et al., 2004). Numerous peptide motifs known as Short Linear Motifs (SLiMs) exist in the proteome of the disordered regions that are already discovered and also this SLiMs have been demonstrated experimentally (Tompa et al., 2014). However significant importance has been given to sequence motifs that impact protein function. Sequence motifs are often observed at protein functional sites such as cleavage sites, binding sites, post-translational modification sites and sub-cellular targeting sites, and are involved in specific protein-protein interactions, regulatory functions and signal transduction (Van Roey et al., 2014).

The SLiMs are recognized as the well-defined sequence patterns that represent graphically via the machine-readable regular expression (REs) or via the sequence logos (Schneider and Stephens, 1990). This however consist the definition for the positionspecific of the allowed residue types and also for the ambiguous positions. The Molecular Recognition Features (MoRFs) segments are the specific class of intrinsically disordered regions (IDRs) consisting of specific molecular recognition and binding functions. They are usually short segments which are interaction-prone and are found within the long IDRs (Vacic et al., 2007). They exhibit disorder-to-order transitions when they bind to their partners (Mohan et al., 2006). They are classified on the basis of their bound state structures into α-MoRFs, β-MoRFs and r-MoRFs. They are not characterized based on the sequence pattern (RE) like SLiMs. Rather they are defined as an interaction-prone disordered segments which form ordered secondary structures when bind to the protein partner. As IDRs significantly helped in the protein interactions, the understanding of protein-protein interactions has effectively changed (Dinkel et al., 2014; Edwards et al., 2012; Petsalaki and Russell, 2008). The interactions via the SLiMs represent various functions in the diverse process which includes control of cell cycle progression, substrate selection for proteasomal degradation, targeting proteins to specific sub-cellular locations and for stabilizing scaffolding complexes.
ScanSite is another database that is known to stores data in 12 different groups for 65 motifs \cite{Obenauer2003}. Apart from ScanSite, Prosite is another database consisting of various data for about 1308 patterns or regular expression even though it comprises of domain signatures \cite{Sigrist2013}. There are very less information related to the SLiMs although much information is there about the eukaryotic cell regulation \cite{Tompa2014}. Hence from these, the need for developing computational methods for predicting novel SLiMs in the protein sequences have been urged effectively.

3 Structure-Based Computational Approaches for Modeling and Simulation of Disordered Regions

For describing the IDP structure and dynamics accurately, the best method is the all-atom molecular simulations consisting of explicit protein and solvent \cite{Ithuralde2016,Knott2012,Stanley2015}. Molecular dynamics (MD) is a method that computationally simulates the time evolution of set of interacting atoms using Newton’s equations of motion. These techniques are dependent on how the molecules will interact- a force field and are popular in materials chemistry, biochemistry and biophysics. It is a technique for producing dynamical trajectory for systems that are composed of N particles through integrating the Newton’s equations of motion. For this, we need a set of initial conditions (positions and velocities of each particle), a good model to represent the forces acting between the particles and to define the boundary conditions to be employed \cite{Hospital2015}. A force field is defined as a mathematical expression that describes the dependence of the energy of a system on the coordinates of its particles. Many force fields are available in the literature that have different degrees of complexity, and oriented to treat different kinds of systems. But there also exists two main drawbacks for the recent force fields on investigating the IDP properties. The first drawback is over-populating a $\beta$-sheet or $\alpha$-helical structure from the older force fields that are eventually

![Fig. 3](image-url) A three-dimensional schematic view of $\alpha$-synuclein protein visualized in PyMol software.
noticed in the disordered chains not forming a long-range tertiary structure (Best et al., 2008). Through the explicit optimization of the backbone torsion parameters opposite to the experimental NMR data, the problems related to the recent force field generation have been clarified. The updated versions of the force-fields are CHARMM 22* (Piana et al., 2011), CHARMM 36 (Best et al., 2012; Huang et al., 2016), Amber ff03* (Best and Hummer, 2009), Amber f99SB*-ILDN (Best and Hummer, 2009; Lindorff-Larsen et al., 2010) and Amber f99SBnmr (Li and Buschweiler, 2011). These marked updated force fields have been reported in the literature by Piana et al. (2014).

The second drawback for the all-atom force fields was evident due to the direct comparison of all-atom simulations with single-molecule FRET and SAXS experiments which showed that the simulated unfolded states were much too collapsed (Piana et al., 2014; Nettels et al., 2009) (being close to maximally compact globules). As such the secondary structures were also over-populated (Skinner et al., 2014). Till the recent two years, the various issues related to the force fields were considered as same. However various modification related to the force field have already addressed this issue (Best et al., 2014; Huang et al., 2016; Nerenberg et al., 2012; Piana et al., 2015; Ploetz et al., 2010; Yoo and Aksimentiev, 2016). From all these force fields, the first approach that was selected was the direct strengthening of protein-water interactions by upgrading the Lennard-Jones mixing rule (Best et al., 2014; Huang et al., 2016; Nerenberg et al., 2012) for investigating the protein-water interactions, as stated by the Amber ff03 ws force field (Best et al., 2014). The second approach evaluated the increasing the well-depth of Lennard-Jones from the novel water model of TIP4P-D (Piana et al., 2015). But the overall impact for the disordered proteins dimension via using Amber ff03ws and TIP4P-D was noticed to be same. This is due to the fact that for both the models, the rise in the interaction for protein-water dispersion is same. Whereas for both models, the average bulk water-water interactions remain same as both the models match the experimental heat of water vaporization. These two models have been shown to play a significant role with the data from SAXS for the other related IDPs(Henriques et al., 2015; Henriques and Skepo, 2016; Zheng et al., 2016). The peptide reconfiguration dynamics is also precisely captured due to the less level of collapsed chain (Zerze et al., 2016).

Apart from these two, the third promising is defined as the Kirkwood-Buff force field (KBFF) designed for proteins for emphasizing on the properties of solution (Ploetz et al., 2010). The KBFF gives rise to results for FG-nucleoporins that are considered consistent with SAXS and FRET experiments (Mercadante et al., 2015). Even after designing new force fields for IDPs, there exists some drawbacks such as folded peptides and proteins are destabilized (Mercadante et al., 2015; Nerenberg et al., 2012; Piana et al., 2015) due to the favourable interactions with water in the denatured state (Best et al., 2014). This effect is compensated presumably by specific stabilizing interactions that are absent in the current force fields (Nerenberg et al., 2012). But however the improvements in the force field may not be considered universal as the presence of some molecules such as disordered arginine/serine (RS) peptide are comparatively more expanded than the experiment (Rauscher et al., 2015). However the atomistic simulations are almost unaffordable and possibly unnecessary due to the larger length spanned by many IDPs in their functional states.

4 Emerging Computational Tools and Techniques

At present, there exist numerous resources which can help us to gain an insight for the function of IDR. The illustrations of the various resources available are summarized in Table 1. There are several areas where efforts are being upgraded and made to help us understand the protein disorder function. We however notice the combination of multiple existing classification schemes that can help us to retain the function for predicting the IDRs. This perhaps can help us to have an extensive knowledge of the function of protein and also to obtain enhanced coverage of function.

On the basis of sequence analysis, various web servers have been designed to predict the disordered regions in protein structures. Recently a database has been launched that is coded by the human genome which consists of consensus disorder cation schemes that can generate the structural ensembles of disordered proteins. Based on the NMR, SAXS data and pE-DB, the modeling of ensembles have been made (Marsh and Forman-Kay, 2012; Salmon et al., 2010; Sibille and Bernado, 2012).

Various approaches have been initiated for the design of disordered proteins and to measure the degree of disordered regions. PrDOS is one such software that was created by Ishida and Kinoshita at the University of Tokyo (Ishida and Kinoshita, 2007). In this software, a hybrid method is used along with a machine learning approach that depends on both the SVM model with a template-based model. The input is used as an evolutionary profile of the input sequence by the machine learning model. Whereas the template-based model explore the homologues in the PDB. Both the predictions made by the machine learning and template-based models are based on a weighted average of the results.

Another approach is the MFDp software developed by Kurgan’s group at the University of Alberta (currently at the Virginia Commonwealth University) (Mizianty et al., 2010). In this meta-predictor, three SVM models are combined in order to calculate long, short, and all-size IDRs. A variety of different inputs are being utilized by SVM that include information taken from the amino acid sequence. The putative disorders are predicted from three predictors that include: evolutionary profile, putative B-factors, putative secondary structure and backbone dihedral torsion angles, putative solvent accessibility and putative annotation of globular domains. MFDp was recently upgraded to MFDp2, in 2013 (Mizianty et al., 2013, 2014) that combines predictions generated by MFDp with predictions that are based on computed alignment against disordered proteins database that are being extracted from DisProt. The predictions generated are corrected in a way that predicted disordered residues number matches putative disordered residues output number by DisCon method (Mizianty et al., 2011).
A last approach was being developed named as DISOPRED that was released by Jones’s group at the University College London (Jones and Cozzetto, 2015). The first version of this method, DISOPRED was published in 2003 (Jones and Ward, 2003), the second version, DISOPRED2, in 2004 (Ward et al., 2004a) and the latest third version, DISOPRED3, in 2015 (Jones and Cozzetto, 2015). The latest version i.e. DISOPRED3 is basically a machine learning method that is implemented as a two-stage neural network. It uses the predictions from three main predictors: DISOPRED2 which is a specialized predictor for the long IDRs and a nearest neighbor-based model with similarities uses to a set of proteins by the IDRs. The design of this method is somewhat similar to MFDp2. It is from the fact this model includes a module that predicts long IDRs and an alignment-based module. However the main difference lies in the input method consisting of an evolutionary profile and the second stage using a neural network mixes these three predictions. The DISOPRED3 software is also known to predict the protein binding site that is defined as the protein binding regions located mainly inside the IDRs.

5 Future Perspectives and Closing Remarks

Still many works are there to characterize the unfolded functional proteins. However various computational techniques for screening the protein sequences, entire genome sequences for the intrinsic disorder proteins will render more proteins belonging to this class. Our understanding will advance for the functions of these disordered regions of proteins in parallel progress of the functional genomics. Although researches in this field have reached the peak point of maturity but recently it has been switched to the prediction of the disordered regions functions. The low complexity sequences situated between the structured domains is however started to be addressed. But the mobility of polypeptide degree consistent with the function of protein is still not clear. Hence proteins and their functions will however progress to a continuum from a fully unstructured protein that folds on binding to the target to a relatively rigid protein that contain varied mobile functional regions. Hence the functional proteins should progress from a static state to a dynamic state where several conformations are consistent with their function.

Hence by understanding the functionality and unique features of IDPs, we can gain an insight onto the biologically active conformations that might play an important role for the diagnosis of several neurodegenerative diseases. Also from the newly developed databases for the intrinsically disordered protein, the determination of structure and their properties can be well understood.
Acknowledgement

The authors thank Central Libraries of Tezpur University and Central University of Jharkhand (CUJ) for providing us resources to complete the article. BRM is thankful to DBT-BUILDER programme (BT/PR-9028/INF/22/193/2013) for financial support.

References


