INTRINSICALLY DISORDERED PROTEINS: ANALYSIS, PREDICTION, SIMULATION, AND BIOLOGY

JIANHAN CHEN
Department of Biochemistry
Kansas State University, Manhattan, KS 66506, USA
jianhanc@ksu.edu

JIANLIN CHENG
Department of Computer Science
Informatics Institute
University of Missouri, Columbia, MO 65211, USA
chengji@missouri.edu

A. KEITH DUNKER
Center for Computational Biology and Bioinformatics
School of Informatics
Indiana University, Indianapolis, IN 46202-5122, USA
kedunker@iupui.edu

1. Introduction

Intrinsically disordered proteins (IDPs) are an important newly recognized class of proteins that rely on the lack of stable structures, both in vitro and in vivo, for function. Before the computational studies of the late 1990s, the tendency was to believe that IDPs were artifacts, resulting for example from denaturation during isolation or from the loss of critical factors such as metal ions or small molecules that were needed for the formation or maintenance of structure. However, the bioinformatics work of the late 1990s showed that most of the IDPs have amino acid compositions distinctly different from the compositions of structured proteins. These studies suggested that, in addition to the concept of "sequence → structure → function" there is another concept of "sequence → disordered ensemble → to function."

Furthermore, the functions carried out by structured proteins and disordered ensembles are very different. The main functions for structured proteins are catalysis, membrane transport, and reversible binding of small molecules, whereas the main functions of IDPs are reversible binding of large molecules, assembly, regulation, and control. Finally, from the amino acid compositional differences between structured and disordered proteins, predictors of disorder were developed. Application of these predictors to various proteomes shows that eukaryotic cells contain a much higher fraction of IDPs compared to bacteria and archaea. For instance, recent predictions suggest that 35% – 45% of human proteins contain long regions that have IDP features, and that, overall, about 45% of the amino acids in humans are located in regions with IDP features.

Thanks to the important and perhaps pivotal role played by computation in the development of the IDP field, it is now recognized by many but not yet all that IDPs play fundamental roles in crucial biological functions and are extensively involved in human diseases. The study of IDPs is rapidly evolving into a field of growing prominence. Nonetheless, the dynamic and heterogeneous nature of IDPs has remained a significant challenge. There are still many open questions in the field, including how to further improve

the accuracy of computational IDP predictions, how to integrate IDP predictions with the study and prediction of other protein structural features, how to better predict IDP sequence-to-function relationships, how to better compute and simulate different states of IDPs, what is the most effective means for integrating structural data from simulation and experiment, how to computationally probe mechanistic aspects of coupled binding and folding and to understand how IDP recognition is regulated, and how to develop rational strategy for modulating IDP function in human diseases. Thus, there is a great demand for further development of computational IDP approaches that are more efficient and more accurate, and at the same time, can improve our understanding of biology from the molecular to the system level. This represents exciting opportunities for computational approaches to make crucial contributions, from aspects ranging from prediction and analysis to simulation.

One of the main goals of this Session is thus to introduce and discuss: (1) important advances in all frontiers of computational "IDPology", (2) available computational capabilities in prediction, analysis and simulation of IDPs, and more importantly, (3) outstanding challenges, further directions and key biological questions to be addressed. The other goal is to promote communication and collaboration between scientists working in different areas of IDP computation and between experiment and computation.

2. Papers in this Session

This session includes 11 high-quality papers addressing a wide range of important problems in prediction, simulation and analysis of IDPs.

Analysis of IDPs' function and evolution

Elucidating biological processes and pathways that IDPs take part in and the consequent evolutionary pressures on IDPs is an important topic. **Huang et al.** developed a method to classify proteins into four categories - structured, mixed, disordered and rare, based on charge-hydropathy and cumulative distribution function. Their investigation of functions of proteins in different categories discovered that disordered class is highly active in mitosis-related processes and the mixed class is highly associated with signaling pathways. **Patil et al.** reported a method and a web tool to group intrinsically disordered domains using a similarity metric based on amino acid composition. The groups identified from a large number of disordered domains by their method were enriched with specific Gene Ontology (GO) and Pfam function annotations, suggesting that this method may be applied to functional annotation of IDPs. **Hsu et al.** designed an experiment to study the sequence conservation of binding regions of multiple partner proteins that interact with the same disordered segment of another protein. The experiment showed that residues interacting with disordered segments were substantially more conserved that other buried or exposed residues, suggesting that these disordered segments may play important function roles (e.g. signaling) in cells. **Jeong and Kim** provided a comprehensive study of co-evolution of intrinsically disordered regions. The study revealed that the degree of residue co-evolution significantly decreased in disordered regions and differd in function categories.

Studying how disordered regions modulate molecular binding and recognition of IDPs is an active research topic. **Vuzman et al.** discovered an interesting mechanism through which disordered tails in DNA-binding proteins were used by post-translational modification to modulate DNA-protein binding affinity. The study demonstrated that post-translational modifications such as acetylation and phosphorylation, which often took place at disordered sites, could modulate the interactions of proteins with DNA by changing the local and global properties of the disordered tails. The binding affinity may be

tuned by adjusting the number of modifications and the cross-links between them. **Guo et al.** used a combination of bioinformatics and statistical methods to study the distribution of disordered residues in human transcription factors (TFs) and non TFs. It was found that the regions flanking DNA Binding Domains (DBD) in TFs were significantly enriched with disordered residues. This specific identification of the locations of enriched disordered regions in TFs may shed light on the functions of these disordered regions. **Gao and Xu** studied the correlation between correlation between posttranslational modification (PTM) and intrinsic disorder in proteins using both functional and structural data. Their investigation showed that most PTMs (e.g. phosphor-serine/-threonine/-tyrosine) preferentially occurred in disordered regions, while a few PTMS occurred in ordered regions more often. The work also suggests that disorder-order transitions may be introduced by PTMs.

Simulation of IDPs' conformation

Generating an ensemble of realistic conformations for a disordered protein is critical for studying its structural and functional properties. **Fisher et al.** introduced a computationally efficient algorithm called Variational Bayesian Weighting with Structure Selection (VBWSS). VBWSS aims to construct an ensemble of an IDP using a minimal number of conformations and provides estimates for the uncertainties in various properties of the ensemble. The authors successfully validated the algorithm against reference ensembles and applied it to construct an ensemble for the 140-residue IDP, α-synuclein. **Burger et al.** carried out long timescale explicit solvent molecular dyanamics (MD) simulations to quantify the ligand-free state of nuclear receptor co-activator binding domain (NCBD) of CREB binding protein. While NCBD has been shown to be a molten globule, the intrinsic disordered nature has prevented detailed structural characterization using traditional methods. Their simulation and analysis was able to capture certain rare states at the tails of the conformational distribution, which, intriguingly, included three of four ligand bound states that were mutually accessible via a complex network of pathways. **Mitrea et al.** employed a series of structure-based and sequence-based methods to elucidate the mechanism of how the N-terminal oligomerization domain of nucleophosmin protein (Npm-N) was transformed from ordered states into disordered states. Their study led to an interesting hypothesis that Npm-N had evolved energetic switches within its structure to enable transformation to a disordered state and the transformation was triggered by sequential phosphorylation of solvent exposed hot spots.

Prediction of IDPs

Predicting disordered regions from protein sequence serves as the basis for many kinds of computational IDP studies. Developing more accurate disorder prediction methods or integrating existing base predictors is important for structural and functional analysis of IDPs. **Peng et al.** designed a new regression-based model to quantify quality of the majority-vote consensus of a given triplet of disorder predictors based on their individual performance and their complementarity measured at the disorder residue and segment levels. The method can effectively integrate individual disorder prediction methods to improve prediction accuracy.

Acknowledgements

We would like to thank all the authors who submitted manuscripts to this session and all the anonymous reviewers who rigorously and constructively reviewed the manuscripts. We greatly appreciate the IDP community's enthusiastic support for this session.