Intrinsically Disordered Proteins – A Tutorial

Jianhan Chen[1], Jianlin Cheng[2], A. Keith Dunker[3]

[1]*Department of Biochemistry, Kansas State University, Manhattan, KS 66506, USA*
[2]*Department of Computer Science, Informatics Institute, University of Missouri, Columbia, MO 65211, USA*
[3]*Center for Computational Biology and Bioinformatics, School of Informatics, Indiana University, Indianapolis, IN 46202-5122, USA*

Introduction

Recognition of the importance of intrinsically disordered proteins (IDPs) has gone through a long process [1-6]. In the 1950s and 1960, a few proteins were suggested to be similar to proteins unfolded by 8M urea, e.g. to be IDPs, using optical rotation or optical rotator dispersion, but much larger numbers of IDP regions were observed as missing electron density in X-ray crystal structures in the 1970s and 1980s. With a few exceptions for which such missing density was regarded as indicating functional disorder, in most cases such missing density was treated as arising from technical difficulties in X-ray diffraction experiments or was just ignored. More recently, IDPs are being considered to be a major group of proteins that stand in contrast to proteins having stable tertiary structures [1-6]. Simply put, IDPs are dynamic and flexible without defined, equilibrium coordinates and bond angles. Although structured proteins exhibit substantial and important dynamic motions, these motions tend to be normal mode oscillations and random thermal fluctuations about equilibrium conformations.

A major factor raising recent interest in the IDPs has been the realization that IDPs and structured proteins carry out complementary sets of biological functions. Almost all structured proteins are enzymes, or transport proteins, or receptors that bind a variety of ligands. IDPs and IDP regions, on the other hand, are involved in regulating enzyme active sites, in providing sites for posttranslational modification, or in providing sites for signaling interactions with nucleic acids or with other proteins. Such signaling interactions typically involve disorder-to-order transitions of IDP regions rather than lock and key interactions of preformed structures,

During the process of gaining recognition of IDPs and IDP regions as an important group of proteins, starting in the 1990s, bioinformatics investigations and increased use of NMR in protein structural studies played important and complementary roles in helping to establish IDPs as a major class of protein that carry out essential functions in all three kingdoms of life, particularly in eukaryotes. Since then, computational research has exploded in all the aspects of IDPs, ranging from prediction of disorder, simulation of disordered ensembles, analysis of disorder function, to biological interpretation of disordered proteins in biological systems. In order to introduce this exciting research field to the broad scientific community, this tutorial provides an overview of major developments and new research frontiers in experimental determination, computational prediction, simulation, and analysis of intrinsically disordered proteins.

Despite the long history and all of the advances briefly outlined above, there is no discussion of IDPs nor of their functional importance in any of the major biochemistry textbooks. This continuing omission

suggests that, despite the mounting evidence, IDPs are still not accepted by the entire community and that more work is still needed.

Prediction of IDPs

Computational prediction of disordered regions in IDPs from protein sequence is often the first important step of studying IDPs. To date, the bioinformatics community has developed more than 50 tools to predict disorder [7], which can be roughly organized into five categories [8], including *ab initio* methods [9-28], clustering methods [29-31], template-based methods [32], hybrid methods [32], and consensus methods [33-35]. *Ab inito* methods, the largest group of all, use only sequence information as input to predict the probability that a residue is disordered. Clustering methods simulate a number of tertiary structure models for a target protein, and then superpose the models together using structural alignment tools in order to identify the degree of structural variation of each residue in the models, which is then used to calculate the probability that the residue is disordered. Template-based methods search a target protein against known protein structures in order to identify homologous structural templates, and the regions of the target protein sequence that can be aligned with the templates are considered ordered and otherwise more likely disordered. Template-based methods are often used with *ab initio* methods to form hybrid methods in order to handle different kinds of protein targets. Hybrid methods choose either template-based or ab initio method to make prediction according to their suitability for the target. Consensus methods apply more than one method to predict disordered regions of a protein and use the consensus of the methods as final predictions.

Recognition and determination of IDPs

Predicted disordered regions or domains can be detected and validated using a range of biophysical techniques including nuclear magnetic resonance (NMR), circular dichroism (CD), small angle X-ray scattering (SAXS), hydrodynamic characterizations, single molecular Förster resonance energy transfer (FRET), and others. Applications and practical considerations of these methods have been extensively discussed in a number of reviews [36-37] (and references therein). NMR is by far the most comprehensive method that can characterize the structural and dynamic unbound IDPs [38]. Many observables can be measured for multiple sites throughout the protein to infer (transient) organizations at the secondary and tertiary levels, including: chemical shift, coupling constant, nuclear Overhauser effect (NOE), paramagnetic resonance enhancement (PRE), residual dipolar coupling (RDC), and spin relaxation. Relaxation dispersion experiments can be also used to derive unparalleled mechanistic insights on binding-induced folding of IDPs [39].

Simulation and modeling of IDPs

Detailed understanding of the conformational properties of unbound IDPs is a key starting point for establishing the physical basis of how intrinsic disorder might support function. The heterogeneous and dynamic nature of the disordered protein states, however, poses significant challenges for obtaining such understanding. Frequently, only ensemble-averaged properties can be measured for disordered proteins except with single-molecule techniques (which have their own limitations in spatial resolution, labeling need and protein size). Recovering the underlying structural heterogeneity using averaged properties is a severely underdetermined problem [40-43]. Great care must be taken to avoid over-interpreting the experimental data and to establish the uniqueness of derived structural ensembles consistent with the

available data. Substantial challenges in experimental characterization of IDPs arguably represent a unique opportunity for molecular modeling to make unique contributions [44]. At the same time, simulation of IDPs also pushes the limit in the accuracy of the current protein force fields as well as our ability to sufficiently sample relevant conformational space. So-called implicit solvent approaches have emerged as an effective approach that can provide a necessary balance between accuracy and computational efficiency [45].

Analysis of IDPs

IDPs play a variety of functional roles in cell, such as signaling, regulation, recognition and control [46-49]. Analysis of IDPs' function, evolution, structural transition, and pharmaceutical applications, has become one of the most active research fields in IDPology [50]. Recent work in this direction includes identification of function sites of IDPs (e.g. short linear motifs) [51], function and structural classification of IDPs [52-53], evolution of IDPs [54-55], protein post-translational modifications of IDPs [56-58], structural disorder of disease related proteins [59-65], and drug potency of disordered proteins such as p53, BRCA1, CTFR and a-synuclein [66-68].

References

Instead of identifying the most important papers in this field, here we will describe a collection of reviews that very recently appeared in Current Opinions in Structural Biology. These reviews are from leaders in the study of intrinsically disordered proteins, and each of these reviewers in turn identifies the recent papers that the authors feel are the most important from their various perspectives. Full citations will be given for this set of papers. Subsequent papers will be cited in a more abbreviated format.

1. A.K. Dunker, J. Gough, Sequences and Topology: Intrinsic disorder in the evolving university of protein structure. Curr Opin Struct Biol., 2011, 21(3):379-81.

   This is the overview editorial for the following set of reviews. This overview frames the entire set of reviews.

2. A. Schlessinger, C. Schaefer, E. Vicedo, M. Schmidberger, M. Punta, B. Rost. Protein disorder – a breakthrough invention of evolution? Curr Opin Struct Biol., 2011, 21(3):412-8.

   Schlessinger et al. argue that disorder should be viewed as distinct from extreme flexibility, that disorder is important for organisms to adapt to different environments, and that use of intrinsic disorder was a crucial step in the evolution from simple bacteria to complex eukaryotes. This paper also emphasizes the need for new advanced computational methods to study these proteins.

3. P. Tompa, Unstructural biology coming of age. Curr Opin Struct Biol., 2011, 21(3):419-25.

   This article suggests that synergy involving developments in 4 research areas is leading to the rapid maturation of "unstructural" biology. These 4 research areas are (i) more advanced bioinformatics tools; (ii) improved ensemble description in both the free and bound states; (iii) application of in-cell approaches for characterizing their structure in vivo; and (iv) generation of small molecule inhibitors.

4. C.K. Fisher, C.M. Stultz, Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol., 2011, 21(3):426-31.

Since IDPs exhibit relatively flat energy landscapes, modeling IDP ensembles requires extra care. This review describes the problems associated with building such ensembles and of using computational techniques to interpret experimental data in terms of such ensembles. Importantly, this review critically assess the advantages and limitations of current techniques and discusses new methods for validating such ensembles.

5. M.M. Babu, R. van der Lee, N.S. de Groot, and J. Gsponer, Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol., 2011, 21(3):432-40.

   These authors review the evidence that altered expression of intrinsically disordered proteins (IDPs) is associated with numerous diseases. Perhaps for these reasons, IDPs are tightly regulated, and dosage-sensitive genes often encode proteins that are rich in IDP segments.

6. C.J. Brown, A.K. Johnson, A.K. Dunker, G.W. Daughdrill, Evolution and Disorder, Curr Opin Struct Biol., 2011, 21(3):441-6.

   This review points out that structured and IDP regions evolve differently due to differences in sequence composition, intramolecular contacts, and function. IDPs have a different pattern of accepted point mutations, exhibit higher rates of insertions and deletions, and, generally, but not always, evolve more rapidly than structured proteins.

7. B. He, et al., Cell Res., 2009, 19, 929–949.
8. X. Deng, et al., Molecular Biosystems, 2011, DOI:10.1039/C1MB05207A.
9. L. Iakouchev et al., Protein Sci, 2001, 3:561-571.
10. J. Cheng, et al., Data Min. Knowl. Discovery, 2005, 11:213–222.
11. J.J. Ward, et al., Bioinformatics, 2004, 20:2138–2139.
12. X. Deng, et al., BMC Bioinformatics, 2009, 10:436.
13. S. Hirose, et al., Bioinformatics, 2007, 23:2046–2053.
14. K. Shimizu, et al., BMC Bioinformatics, 2007, 8:78.
15. Y. Zhou and E. Faraggi, in Introduction to Protein Structure Prediction: Methods and Algorithms, ed. H. Rangwala and G. Karypis, Wiley, 2010.
16. Z. Dosztanyi, et al., J. Mol. Biol., 2005, 347:827–839.
17. Z. Dosztanyi, et al., Bioinformatics, 2005, 21:3433–3434.
18. P. Romero, et al., Genome. Inf. Ser., 1997, 8:110–124.
19. P. Romero, et al., Proteins: Struct., Funct., Genet., 2001, 42:38–48.
20. K. Peng, et al.. BMC Bioinformatics, 2006, 7:208.
21. R. Linding, R. B. Russell, V. Neduva and T. J. Gibson, Nucleic Acids Res., 2003, 31, 3701–3708.
22. J. Prilusky, et al., Bioinformatics, 2005, 21:3435–3438.
23. Z.R.Yang, et al., Bioinformatics, 2005, 21:3369–3376.
24. R. Thomson, et al., Bioinformatics, 2003, 19:1741–1747.
25. J. Liu, B. Rost, Nucleic Acids Res., 2003, 31:3833–3835.
26. K. Coeytaux and A. Poupon, Bioinformatics, 2005, 21:1891–1900.
27. S. O. Garbuzynskiy, et al., Protein Sci., 2004, 13:2871–2877.
28. A. Vullo, et al., Nucleic Acids Res., 2006, 34:W164–W168.
29. L.J. McGuffin, Bioinformatics, 2008, 24:1798–1804.
30. L.J. McGuffin, Proteins: Struct., Funct., Bioinf., 2009, 77:185–190.
31. D.B. Roche, et al., Nucleic Acids Res., 2011, 39(suppl. 2):W171–W176.
32. T. Ishida and K. Kinoshita, Nucleic Acids Res., 2007, 35:W460–W464.
33. T. Ishida and K. Kinoshita, Bioinformatics, 2008, 24:1344–1348.

34. M. J. Mizianty, et al., Bioinformatics, 2010, 26:i489–i496.
35. Z. Peng and L. Kurgan. Pacific Symposium on Biocomputing, 2012, 17:176-187.
36. V.N. Uversky, et al., J of Mol Recogn, 2005, 18:343-384.
37. D. Eliezer, Current Opinion in Structural Biology, 2009, 19:23-30.
38. P.E Wright and H.J. Dyson, Current Opinion in Structural Biology, 2009, 19:31-38.
39. K. Sugase, et al., Nature 2007, 447:1021-1025.
40. D. Ganguly, J. Chen, J Mol Biol 2009, 390:467-477.
41. C.K. Fisher, et al., J Am Chem Soc 2010, 132:14919-14927.
42. T. Mittag, J.D. Forman-Kay, Curr Opin Struct Biol 2007, 17: 3-14.
43. C.K. Fisher, C.M. Stultz, Curr Opin Struc Biol 2011, 21: 426-431.
44. T.H. Click, et al., Int J Mol Sci 2010, 11:5293-.
45. A.H. Mao, et al., Proc Natl Acad Sci U S A, 2010, 107: 8183-.
46. A.K. Dunker, et al., Biochemistry, 2002, 41:6573-6582.
47. A.K. Dunker, et al., Adv. Prot. Chem, 2002, 62:25-49.
48. H. Xie, et al., Proteome Res., 2007, 6:1882-1932.
49. V. Vacic, et al., Proteome Res, 2007, 6:2351-2366.
50. P. Tompa. Current Opinion in Structural Biology, 2011, 21(3):419-25.
51. C.M. Gould, et al., Nucleic Acids Res, 2009, 38:D167-180.
52. F. Huang, et al., Pacific Symposium on Biocomputing, 2012, 17:128-139.
53. A. Patil, et al., Pacific Symposium on Biocomputing, 2012, 17:164-175.
54. W.L. Hsu, et al., Pacific Symposium on Biocomputing, 2012, 17:116-127.
55. C.S. Jeong, D. Kim, Pacific Symposium on Biocomputing, 2012, 17:140-151.
56. D. Vuzman, et al., Pacific Symposium on Biocomputing, 2012, 17:188-199.
57. X. Guo, et al., Pacific Symposium on Biocomputing, 2012, 17:104-115.
58. J. Gao and D. Xu, Pacific Symposium on Biocomputing, 2012, 17:94-103.
59. J. Gsponer et al., Science, 2008, 322:1365-1368.
60. Y. Cheng, et al., Trends Biotechnol, 2006, 24:435-442.
61. V.N. Uversky, et al., Annu Rev Biophys, 2008, 37:215-246.
62. F. Chiti, C.M. Dobson, Annu Rev Biochem, 2006, 75:333-366.
63. H. Hegyi, PLoS Comput Biol, 2009, 5:e1000552.
64. T. Vavouri, et al., Cell, 2009,138:198-208.
65. N.E. Davey, et al., Trends Biochem Sci, 2010, 36:159-169.
66. L.T. Vassilev, et al., Science, 2004, 303:844-848.
67. S.J. Metallo, Curr Opin Chem Biol, 2010, 14:481-488.
68. D.I. Hammoudeh, et al., J Am Chem Soc 2009, 131:7390-7401.