PostDoc Journal
Vol.2, No.1, January 2014

Journal of Postdoctoral Research
www.postdocjournal.com

Protein Structure and Function: Methods for Prediction and Analysis

Ravi Ramesh Pathak
Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
Email: rpathak@health.usf.edu

Abstract

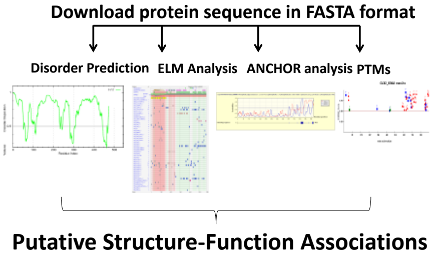
Protein structure and functions that are associated with it have been studied extensively. Traditional methods of studying protein and structure have largely relied on biochemical and biophysical characterization methods. However, these methods rely on the assumption that proteins are largely structured and adopt energetically stable three-dimensional conformations with minimum free energy. The discovery of naturally unfolded proteins or disordered proteins that are characterized by lack of stable tertiary structure paved the way to study hitherto unexplained facets of protein structure and function. There are a number of computational methods that exploit protein sequence information to predict whether a protein is disordered. These are complemented by a number of other tools that allow users to predict protein function based on the occurrence of post translational modifications, short linear motifs and other disorder associated regions driving functional plasticity. These easy to apply and interpret tools are presented in the form of a user-friendly workflow for bench-scientists in the current communication.

Keywords: intrinsic disorder, IDP, protein, protein structure, PTM

Introduction

The traditional paradigm of protein structure which states that a well-defined structure dictates the function of a protein has been challenged due to the relatively recent discovery of intrinsically unfolded proteins(Mirsky and Pauling 1936; Wu 1995; Romero, Obradovic et al. 1998; Dunker, Silman et al. 2008). These proteins that are naturally flexible with intrinsically disordered regions (IDRs) are called intrinsically disordered proteins (IDPs)(Wright and Dyson 1999; Dunker and Obradovic 2001). IDPs and IDRs are highly ubiquitous and have been shown to mediate a number of biological functions, participating in recognition and in various signaling and regulatory pathways, via specific protein-protein, protein-nucleic acid and protein-ligand interactions(Tompa 2005; Vucetic, Xie et al. 2007; Xie, Vucetic et al. 2007; Xie, Vucetic et al. 2007). Additionally, sites of various post-translational modifications (PTMs) are frequently associated with regions of intrinsic disorder(Xie, Vucetic et al. 2007). IDPs have been implicated in a number of diseases that include human neurodegenerative disorders and cancers(Uversky 2003; Oldfield, Meng et al. 2008; Uversky, Oldfield et al. 2008). This makes the study of intrinsic disorder very critical for understanding the role of structural plasticity with disease associated functions. The availability of a number of computational tools to predict disorder opens new avenues for researchers (http://www.disprot.org)(Romero, Obradovic et al. 2001; Linding, Jensen et al. 2003; Obradovic, Peng et al. 2003; Vucetic, Brown et al. 2003; Obradovic, Peng et al. 2005; Xue, Dunbrack et al. 2010). Researchers can apply a combination of tools outlined in the following section to study intrinsic disorder in proteins and form testable hypothesis to relate protein structure with function.

Work Flow:


Download protein sequence in FASTA format
Disorder Prediction
ELM Analysis
ANCHOR analysis
PTMs
Putative Structure-Function Associations

Disorder Prediction

Researchers can retrieve protein sequences in FASTA format from protein databases in FASTA format. The FASTA sequences can be used as input for any one of the online disorder predictors that are available at disport.org. The choice of which predictor to use depends on which disorder feature of the proteins is being analyzed. These have been reviewed in greater detail elsewhere. The predictors analyze the sequences and generate graphical as well as tabular outputs that can be downloaded by the users and interpreted for a given hypothesis. Users can also apply a composition profiler tool http://www.cprofiler.org/ to analyze amino-acid compositional bias characterized by a low content of so-called order-promoting residues such as Cys, Trp, Phe, Tyr, Val, Leu, and Ile, and a high content of so-called disorder-promoting residues, Glu, Lys, Arg, Asp, Gln, Ser, Pro, and Thr(Vacic, Uversky et al. 2007).

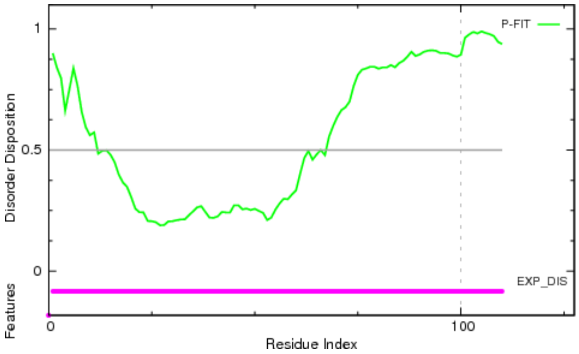


Figure 1: Disorder prediction for 60S acidic ribosomal protein P2-beta using PONDR-FIT predictor(Xue, Dunbrack et al. 2010). Residue scores above 0.5 represent disorder.

ELM Prediction

Short linear motifs (SLiMs), are functional modules found in intrinsically disordered regions(Diella, Haslam et al. 2008; Davey, Van Roey et al. 2012; Dinkel, Van Roey et al. 2013). Interactions mediated by SLiMsare known to direct many diverse processes including cell cycle progression, tagging proteins for proteasomal degradation targeting proteins to specific sub-cellular localizations and stabilizing scaffolding complexes(Diella, Haslam et al. 2008). Researchers can use FASTA sequences as input on the eukaryotic linear motif (ELM http://elm.eu.org) resource, which is a hub for collecting, classifying and curating information about short linear motifs (SLiMs)(Puntervoll, Linding et al. 2003). The tool provides users with a detailed summary of current literature on the motif, information about biological context, taxonomic distribution, a set of representative class instances, interacting protein domain(s), as well as links to primary literature and additional resources. The output is the form of an easy to interpret graphical table that highlights all the associated features and raw data can be downloaded as an excel file.

ANCHOR Prediction

Disordered proteins contain binding regions that function by undergoing a disorder-to-order transition upon binding to a globular protein partner(Meszaros, Simon et al. 2009). Users can apply the ANCHOR tool (http://anchor.enzim.hu/) to identify segments in a generally disordered region that cannot form enough favorable intrachain interactions, however possess the capability to energetically gain by interacting with a globular partner protein(Dosztanyi, Meszaros et al. 2009). Users can also input more than one sequence using the multiple sequence version of ANCHOR. FASTA sequences or list of UniProt IDs/ACs can be used as the input and the output is a text file. Graphical plots are also provided and users can choose to retrieve the results as such. The ANCHOR tool can also be used to identify specific motifs in the disordered regions.

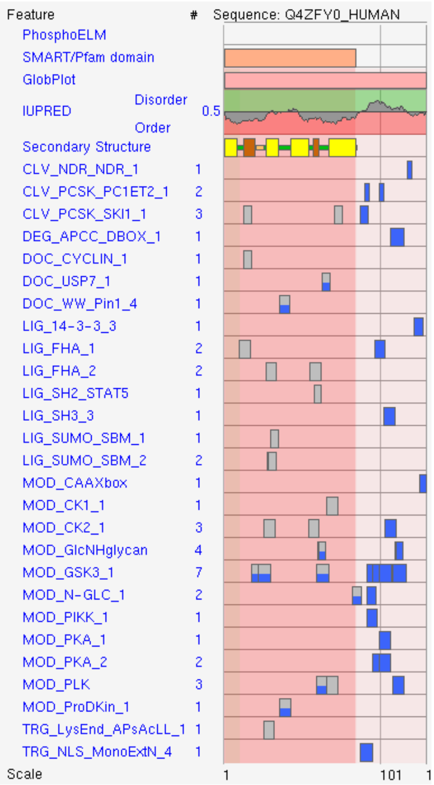


Figure 2: ELM prediction for Putative uncharacterized protein ARHE protein. Different features are represented in the graphical output that includes disordered regions, domain information on the protein and different classes of ELMs. Dark blue represents highly conserved ELMs in the protein.

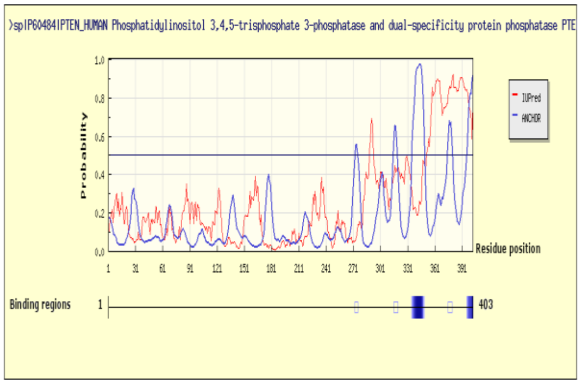


Figure 3: ANCHOR analysis of Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase (PTEN). Disorder regions are predicted using IUPRED in red while the anchor regions are denoted in blue. Binding regions are represented with amino acid positions.

Post Translational Modification (PTM) Prediction

In order to predict PTMs researchers can use Musite (http://musite.net/), a tool specifically designed for large scale predictions of both general and kinase-specific phosphorylation sites(Gao, Thelen et al. 2010). In addition to phosphorylation, it can also predict a range of other PTMs that include methylation, acetylation, sumoylation and sulfation to name a few. The input can be accession IDs or FASTA sequence of a given protein. Users can save the output as a graphical plot and raw data as excel sheets. In addition to the Musite tool, users can apply Disphos 1.3 (http://www.dabi.temple.edu/disphos/pred.html) to computationally predict serine, threonine and tyrosine phosphorylation sites in proteins(Iakoucheva, Radivojac et al. 2004). DISPHOS applies disorder information to discriminate between phosphorylation and non-phosphorylation sites thereby providing a strong support for the hypothesis that protein phosphorylation predominantly occurs in regions of intrinsic disorder.

Combining multiple predictions for structure and function analysis

Once these multiple tools are applied to predict the different features associated with protein disorder the researcher can overlay all the information to arrive at a composite view of the protein. For instance overlapping ELMs, ANCHOR motifs and PTMs on the disordered regions/residues of a given protein can allow the user to predict the co-occurrence of one or more of these features; which in turn can be correlated with structural plasticity based on intrinsic disorder.

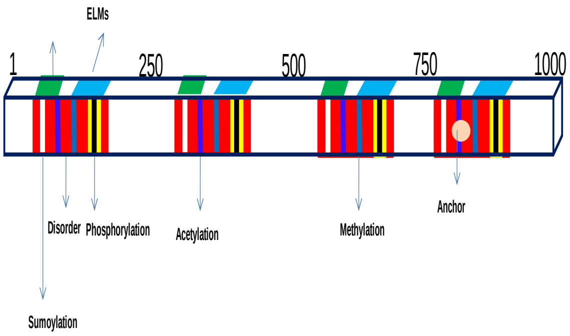


Figure 4: Combination of structural features on full length protein sequence. Users can overlay information about the different structural attributes like disorder and PTMs onto the protein to decipher structure-function associations. Figure represents amino acids (1-1000), colors denote individual structural features.

Conclusion

Application of the flow chart described in this communication can allow researchers to correlate protein structure with function, thereby allowing formulation of testable hypothesis to study protein-protein interactions.

References

1. Davey, N. E., K. Van Roey, et al. (2012). "Attributes of short linear motifs." Molecular BioSystems **8**(1): 268-281.
2. Diella, F., N. Haslam, et al. (2008). "Understanding eukaryotic linear motifs and their role in cell signaling and regulation." Frontiers in bioscience : a journal and virtual library **13**: 6580-6603.
3. Dinkel, H., K. Van Roey, et al. (2013). "The eukaryotic linear motif resource ELM: 10 years and counting." Nucleic acids research.
4. Dosztanyi, Z., B. Meszaros, et al. (2009). "ANCHOR: web server for predicting protein binding regions in disordered proteins." Bioinformatics **25**(20): 2745-2746.
5. Dunker, A. K. and Z. Obradovic (2001). "The protein trinity--linking function and disorder." Nature Biotechnology **19**(9): 805-806.
6. Dunker, A. K., I. Silman, et al. (2008). "Function and structure of inherently disordered proteins." Current opinion in structural biology **18**(6): 756-764.
7. Gao, J., J. J. Thelen, et al. (2010). "Musite, a tool for global prediction of general and kinase-specific phosphorylation sites." Molecular & cellular proteomics : MCP **9**(12): 2586-2600.
8. Iakoucheva, L. M., P. Radivojac, et al. (2004). "The importance of intrinsic disorder for protein phosphorylation." Nucleic acids research **32**(3): 1037-1049.
9. Linding, R., L. J. Jensen, et al. (2003). "Protein disorder prediction: implications for structural proteomics." Structure **11**(11): 1453-1459.
10. Meszaros, B., I. Simon, et al. (2009). "Prediction of protein binding regions in disordered proteins." PLoS computational biology **5**(5): e1000376.
11. Mirsky, A. E. and L. Pauling (1936). "On the Structure of Native, Denatured, and Coagulated Proteins." Proceedings of the National Academy of Sciences of the United States of America **22**(7): 439-447.
12. Obradovic, Z., K. Peng, et al. (2003). "Predicting intrinsic disorder from amino acid sequence." Proteins **53 Suppl 6**: 566-572.
13. Obradovic, Z., K. Peng, et al. (2005). "Exploiting heterogeneous sequence properties improves prediction of protein disorder." Proteins **61 Suppl 7**: 176-182.
14. Oldfield, C. J., J. Meng, et al. (2008). "Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners." BMC Genomics **9 Suppl 1**: S1.
15. Puntervoll, P., R. Linding, et al. (2003). "ELM server: A new resource for investigating short functional sites in

modular eukaryotic proteins." Nucleic acids research **31**(13): 3625-3630.

16. Romero, P., Z. Obradovic, et al. (1998). "Thousands of proteins likely to have long disordered regions." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 437-448.
17. Romero, P., Z. Obradovic, et al. (2001). "Sequence complexity of disordered protein." Proteins **42**(1): 38-48.
18. Tompa, P. (2005). "The interplay between structure and function in intrinsically unstructured proteins." FEBS letters **579**(15): 3346-3354.
19. Uversky, V. N. (2003). "A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders." Journal of biomolecular structure & dynamics **21**(2): 211-234.
20. Uversky, V. N., C. J. Oldfield, et al. (2008). "Intrinsically disordered proteins in human diseases: introducing the D2 concept." Annual review of biophysics **37**: 215-246.
21. Vacic, V., V. N. Uversky, et al. (2007). "Composition Profiler: a tool for discovery and visualization of amino acid composition differences." BMC Bioinformatics **8**: 211.
22. Vucetic, S., C. J. Brown, et al. (2003). "Flavors of protein disorder." Proteins **52**(4): 573-584.
23. Vucetic, S., H. Xie, et al. (2007). "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions." Journal of proteome research **6**(5): 1899-1916.
24. Wright, P. E. and H. J. Dyson (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." Journal of molecular biology **293**(2): 321-331.
25. Wu, H. (1995). "Studies on denaturation of proteins. XIII. A theory of denaturation. 1931." Advances in protein chemistry **46**: 6-26; discussion 21-25.
26. Xie, H., S. Vucetic, et al. (2007). "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins." Journal of proteome research **6**(5): 1917-1932.
27. Xie, H., S. Vucetic, et al. (2007). "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions." Journal of proteome research **6**(5): 1882-1898.
28. Xue, B., R. L. Dunbrack, et al. (2010). "PONDR-FIT: a meta-predictor of intrinsically disordered amino acids." Biochimica et biophysica acta **1804**(4): 996-1010.