

Prediction of Binding Sites and Secondary Structure formation on Nucleic Acid Binding Intrinsically Unstructured Proteins

Russell C. Goodman

Context of Research:

The research is being undertaken as part of a continuing independent research project. Course number: CHM 495: Research in Chemistry.

Thesis:

Intrinsically unstructured proteins (IUPs) are a novel class of proteins that exhibit conformational flexibility under physiological conditions. That is, as opposed to the classic conception of protein structure in which a protein adopts a single well-defined structure, IUPs exist as random coil-like and pre-molten globule-like structures in solution (Uversky 2002). It was determined, however, in the late 1990s that IUPs can form higher-order structure and exhibit functional roles in cells. Structure is formed through the binding of ligands, such as proteins, ions, small organic molecules, and nucleic acids (Tompa 2005). We intend to develop an algorithm to accurately predict regions within unstructured proteins that have a high propensity to bind to a ligand of known structure and ascertain secondary structure that may form on these regions.

Significance of Research:

Computational methods of analyzing IUPs to date have focused on using sequence composition methods to predict regions of Proteins that may exhibit intrinsically unstructured characteristics (Bracken et al. 2004; Ward et al. 2004; Linding et al. 2004). These algorithms are fundamentally based on the conception that IUPs differ significantly in their amino acid composition from globular proteins; therefore, to predict these regions involves only understanding the amino acids that tend to promote disorder (Tompa 2005).

With increasing evidence that the majority of IUPs exhibit an induced folding mechanism, that is, a mechanism of structural formation induced upon binding a ligand, there is confidence that,

similar to unstructured region predictors, algorithms that predict binding sites on IUPs based on amino acid composition can feasibly be developed (Wright and Dyson 2009).

Elucidating binding sites and associated structure formation on these binding sites in IUPs is significant as this is the starting point for investigations into higher-order structure, therefore, function of IUPs. In fact, as IUPs have been estimated to represent up to 30% of the eukaryotic proteome, it is becoming increasingly important to understand the similarities and differences of the structure-function paradigm as it applies to globular proteins and to IUPs (Gsponer and Babu). Deciphering these relationships will allow for greater insight into cellular processes.

Methodology:

The DisProt Database, a database hosted between the Center for Computational Biology and Bioinformatics at Indiana University School of Medicine and Center for Information Science and Technology at Temple University, was used to locate the nucleic acid binding IUPs used in our database. The binding sites of these IUPs were characterized using structures visualized by the programs Ligand Explorer and Protein Workshop available through the Protein Data Bank (PDB).

The binding sites of 7 nucleic acid binding IUPs have been characterized. Using the data from the characterization of the IUP binding sites, the average frequency of occurrence, f_r , of all 20 amino acids was determined. The equation used to determine the average frequencies is in the form of:

$$f_r = \frac{\sum n_r}{N}$$

where f_r is the average frequency of residue r , n_r is the number of amino acids with residue r in each protein of the total number of proteins characterized, N . These frequencies are used as parameters for our sequence composition algorithm used to predict binding sites in nucleic acid binding IUPs. The inequality, $f_r \cong 0.5$, characterizes the amino acid as favorable for participating in the binding site, for this indicates that the amino acid occurred at least once in every two binding sites of nucleic acid binding IUPs. These parameters are being updated continually as structures become available on the Protein Data Bank.

An algorithm for predicting binding sites on nucleic acid binding IUPs that is founded on sequence composition is being developed. However, this algorithm will differ from our previous algorithm in that it uses our sequence composition parameters as a fundamental means of finding regions with a high propensity to form binding sites yet will predict only binding sites exhibiting secondary structural patterns and characteristics.

Works Cited:

Bracken, C., Iakoucheva, L.M., Romero, P.R. and Dunker, A.K. 2004. Combining prediction, computation and experiment for the characterization of protein disorder. *Current Opinion in Structural Biology* 14: 570-576.

Gsponer J., Babu M.M. 2009. The rules of disorder or why disorder rules. *Progress in Biophysics and Molecular Biology*. 99: 94-103.

Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. 2003. Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31: 3701-3708.

Tompa, P. 2005. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters* 579: 3346-3354.

Uversky, V.N. 2002. Natively unfolded proteins: A point where biology waits for physics. *Protein science* 11: 739-756.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337: 635-645.

Status:

We have developed a sequence composition algorithm that has shown to accurately predict, on average, 80.9% of the binding sites on nucleic acid binding IUPs with, on average, 47.1% of the binding sites predicted indicative of the native structure. These results indicated that the majority of the binding sites on IUPs can be predicted through sequence composition. However, algorithms that incorporate additional constraints regarding the feasibility of the structure formation need to be developed to discriminate between regions of IUPs having a composition indicative of a binding site and actual binding sites.

Budget:

Round Trip Flight – Syracuse to San Francisco	\$400
Hotel Accommodations (three nights)	\$375
Transport from/to airport	\$30

Meals	
Breakfast	\$15
Lunch	\$15
Dinner	\$25
Registration fee	<u>\$340</u>
Total	\$1175

Prior Funding:

No previous applications have been submitted for funding.