Prediction of Binding Sites on Nucleic Acid Binding Intrinsically Disordered Proteins

Russell C. Goodman and Theresa L. Beaty
Department of Chemistry and Physics, Le Moyne College, Syracuse, NY USA

LE MOYNE
SPIRIT. INQUIRY. LEADERSHIP. JESUIT.

Abstract

Many of the methods of analyzing intrinsically Disordered proteins (IDPs) to date has been limited to sequence analysis that attempts to predict intrinsically disordered regions of proteins. However, our research is based on developing algorithms for predicting ligand binding sites and the associated secondary structure of these binding sites in IDPs. Our algorithms are founded on parameters determined through a statistical method, which used the Protein Data Bank, to calculate the frequency of all 20 amino acids occurring at the binding sites of 7 nucleic acid binding IDPs. Our primitive sequence composition algorithm for predicting binding sites, SeqCom, predicts, on average, 87.3% of the binding sites with 49.6% of the binding sites predicted representative of the native binding sites. To improve binding site prediction, we developed IUPattern. IUPattern works on the same principles as SeqCom, but it uses additional binding site constraints to better decipher between native and non-native binding sites. IUPattern predicts, on average, 70.6% of the binding sites with 58.1% of the binding sites predicted representative of the native binding sites.

Introduction

Elucidating binding sites and associated structure formation on these binding sites in IDPs is significant as this is the starting point for investigations into higher-order structure, therefore, function of IDPs. In fact, as IDPs 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 IDPs (Gsponer and Babu). Deciphering these relationships will allow for greater insight into cellular processes.

The model IDP used in our studies of the binding sites in IDPs is the nucleic acid binding IDP. This is a simple model as it provides a constant binding partner, thus ridding the need for *a priori* knowledge of the ligand structure. We developed two algorithms based on the statistical parameters developed for this model: SeqCom and IUPattern. SeqCom is a primative sequence composition algorithm that searches for regions in the IDP primary structure that have a high probability of forming binding sites. These predictions allow for the possibility of secondary structure formation at the binding site. IUPattern is a heuristic algorithm that, similar to SeqCom, searches for regions of IDPs that have a high propensity of forming binding sites but also constrains these predictions by only predicting binding regions that indicate the formation of a structured binding site (i.e., only allows for the prediction of binding sites that exhibit ordered structures, such as alpha helix formation or straight chain binding).

To benchmark the performance of our predictions, we used two scoring methods: predictive ability and accuracy. These methods are defined as:

$$\text{Predictive Ability} = \frac{\text{Number of Accurately Predicted Amino Acids in Binding Sites}}{\text{Total Amino Acids in Binding Sites in the Known Structure}}$$

$$\text{Accuracy} = \frac{\text{Number of Accurately Predicted Amino Acids in Binding Sites}}{\text{Total Number of Predicted Amino Acids}}$$

The predictive ability represents how well the algorithm predicted the known binding sites, while the accuracy represents how well the algorithm discriminates between binding and nonbinding regions.

We are in the process of finalizing a newer version of IUPattern involving the incorporation of a coupling constraint matrix, called IUPatternC. IUPatternC works on similar principles as IUPattern; however, the amino acids in the local region of predicted binding sites by IUPatternC are analyzed for their propensity to form secondary structure or straight chain binding. The intention is to further restrict the number of false positives being predicted by our nucleic acid binding IDP structure prediction applications.

Results: SeqCom & IUPattern

The predictive ability and accuracy of IUPattern and SeqCom are shown in table 1. Figures 1 through 3 show predicted binding sites and the corresponding known binding site of selected IDPs. Figures 1 and 2 contain reasonably good predictions, while figure 3 represents poor binding site predictions on p65.

Table 1. Summary of the predictive ability and accuracy of IUPattern and SeqCom. The results indicate that SeqCom finds more of the native binding sites, while IUPattern can better resolve the known binding sites from the regions of high binding site similarity.

SeqCom		
	Predictive Ability	87.3 +/- 8.4
	Accuracy	49.6 +/- 16.5
IUPattern		
	Predictive Ability	70.6 +/- 17.5
	Accuracy	58.1 +/- 15.2

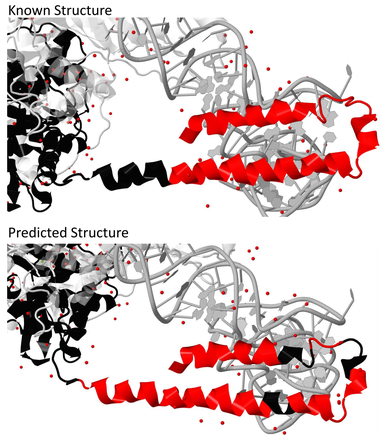

Figure 1. Known and predicted binding sites of phenylalanine-tRNA synthase using IUPattern. Predicted and known binding sites are displayed in red. This represents one of the more accurate predictions from our database.

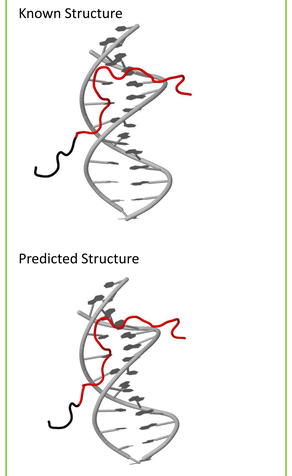

Figure 2. Known and predicted binding sites (red) of a portion of a high mobility group protein using IUPattern.

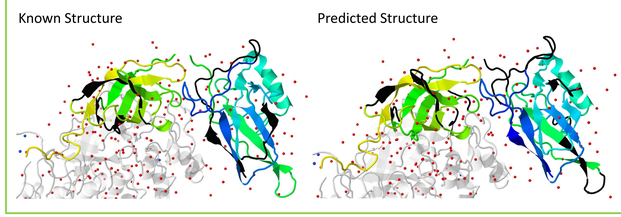

Figure 3. Known and predicted binding sites of p65. Black regions indicate binding sites. Predictions made with IUPattern.

IUPatternC

IUPatternC, or IUPattern Coupling, is an extension of our heuristic algorithm, IUPattern. IUPatternC incorporates a coupling constraint matrix (CCM) that is intended to limit the number of false positive predictions by accounting for local interactions of amino acids in IDPs. That is, since IUPattern requires that potential binding sites show patterns of ordered structure formation (e.g., α-helices and β-pleated sheets), IUPatternC requires that the local interactions of the amino acids promote the ordered structure formation for each predicted binding site.

IUPatternC incorporates three coupling matrices: straight chain, α-helix, and β-pleated sheet. The straight chain binding coupling matrix contains all pair-wise averages of the binding site parameters for a given amino acid sequence. Similarly, the α-helix and β sheet coupling matrices contain all pair-wise averages using the Chou and Fasman parameters developed for secondary structure prediction.

To incorporate coupling into the binding site predictions, some n x n neighborhood of the coupling matrices is analyzed for each amino acid predicted as a potential binding site. Our standard predictions involve a 3 x 3 neighborhood indicating that only the immediately adjacent amino acids of each position in the preliminary predicted binding sites are being coupled.

Disrupted α-helix:

K L I K I I R L I R L T K L L R L I K
α α α α α α α α α α α – – – – – α α α α

Disrupted β-pleated sheet:

K I K L R I R L K P K P R P R I K L K
β β β β β β β β – – – – – – – – β β β

Figure 4. Visual representation of coupling in IUPattern. The tyrosine residue disrupts the helix formation, while the proline residue disrupts that beta sheet formation.

Discussion

Our more advanced heuristic algorithm, IUPattern, had improved accuracy compared to our primitive sequence composition algorithm, SeqCom. The improved accuracy and lower predictive ability was not unexpected as IUPattern was designed to better resolve binding and nonbinding regions, which often involves compromising the ability to find native binding contacts. With additional refinements and packaging, IUPattern can begin to be applied by the biological community for IDP research projects as a means of gaining insight into regions of increased probability of forming binding sites in nucleic acid binding IDPs. More information and the source code of our binding site prediction applications can be found at http://openwetware.org/wiki/User:Russell_C._Goodman.

With the development of IUPatternC, we hope to improve binding site prediction and begin making crude secondary structure predictions. IUPatternC ought to have even further improved accuracy and will display the confidence in each prediction. We hope IUPatternC will be of even greater use to the biological community as a computational resource for experimental investigations of IDPs.

References

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