Prediction of Binding Sites on Intrinsically Disordered Proteins

Russell C. Goodman and Theresa L. Beaty
Department of Chemistry and Physics, Le Moyne College, 1419 Salt Springs Road, Syracuse, NY 13214

Introduction

Intrinsically disordered proteins (IDPs) are a novel class of proteins that, until over a decade ago, had not been recognized as a functional class of proteins. As opposed to globular and lipid-soluble proteins, IDPs lack a well defined structure. In solution, an IDP adopts an ensemble of conformations; however, when bound to a ligand, an IDP adopts a particular structure with a particular function. Disordered proteins my bind a multitude of ligands, exhibiting what is referred to as binding promiscuity (Uversky 2005). Therefore, rather than adhering strictly to the classical structure-function paradigm associated with globular proteins, IDPs can adopt multiple structures, each possessing a different function.

The functions adopted by IDPs are quite diverse. Peter Tompa, from the Hungarian Academy of Sciences, had outlined the various classes of unstructured proteins. Some IDPs exist as mere entropic chains that act somewhat as springs and links between other proteins. However, most are classified as recognition IDPs, which consist of transient binding and permanent binding IDPs. The former class consists of display sites, which act in post-translational modification, and chaperones, which act in assisting the folding of RNA and proteins. The permanent binding IDPs consist of effectors that act in affecting the activity of partner molecules, assemblers which assist in the formation of protein complexes, and scavengers which store and neutralize small ligands (2005). The binding partners for disordered proteins are as diverse as their functions. The ligands consist of ions, small organic molecules, other proteins, and nucleic acids, such as RNA and DNA. This, however, may not be a comprehensive list.

Computational methods of analyzing IDPs to date have focused on using sequence composition methods to predict regions of proteins that may exhibit intrinsically disordered characteristics (Bracken et al. 2004; Ward et al. 2004; Linding et al. 2004). These algorithms are fundamentally based on the conception that IDPs 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 unstructured proteins 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 IDPs 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 IDPs is significant as this is the starting point for investigations into higher-order structure, therefore, function of IDPs. In fact, as disordered proteins have been estimated to represent up to 30% of the eukaryotic proteome, it is becoming increasingly important to understand the 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 capacity to predict the structure of IDPs bound to known ligands has implications in other fields aside molecular biology. The pharmaceutical industry primarily targets structured proteins with small organic molecules to induce a particular desired effect. Little known work has been done in targeting IDPs with small organic molecules. It will be important in the future to

investigate this area as research has suggested disease and disorder are inextricably attached. This idea was deemed the D^2 concept by Dunker et al. (2008). Dunker has shown that many disease-related proteins contain disordered regions.

IDPs may have implications in synthetic biology in the future as well. Synthetic biology is a field involved in reprogramming cells to provide some useful product or effect. The ability of IDPs to be involved in multiple, promiscuous interactions allows for designing signaling pathways with a single protein hub. One such application may be in using a transcription factor with a disordered protein binding domain to respond to multiple protein signals. This may be useful if the activation of a gene was desired under multiple different conditions.

Here, an improved nucleic acid binding site prediction algorithm, called IUPatternC, is presented. IUPatternC introduces a small portion of the complexity of protein binding site prediction by considering the local amino acid composition in predictions. This provided improved resolution in binding site prediction with only a small decrease in the algorithm's ability to locate native binding sites. A parallel comparison between IUPatternC and the previous two algorithms, IUPattern and SeqCom, is also presented.

Methods

Database of Nucleic Acid Binding IDPs

The nucleic acid-binding IDPs used for both obtaining the binding site parameters and testing the algorithms were obtained from the DisProt database, which is run between the Center for Computational Biology and Bioinformatics at Indiana University and the Center for Information Science and Technology at Temple University. Corresponding structures of IDPs bound to RNA or DNA were obtained from the Protein Data Bank (PDB). The ten IDPs that were recovered

from the DisProt database are shown in Table 1. All disordered proteins in table 1 were used for testing the algorithm; however, IDPs with an asterisk were not used in calculating the binding site parameters.

Table 1. Proteins used for obtaining the binding site parameters and testing the algorithms.

Protein Name	Disprot ID Codes	PDB ID Codes
Antitermination Protein N	DP00005	1QFQ
HMG-I(Y)	DP00040	2EZD
Topoisomerase I	DP00075	1A36
Topoisomerase II	DP00076	2RGR
Transcription factor p65	DP00129	1IKN
transcriptional activator traR	DP00198	1L3L
Phenylalanine tRNA sythetase	DP00053	2IY5
Transcription factor 1*	N/A	1CQT
Vitamin D3 receptor*	N/A	1kb2
Seryl tRNA synhtetase*	DP00514	1SER

Binding Site Parameter Calculation

The average frequencies of the amino acids occurring at the binding site on nucleic acid-binding IDPs, namely the binding site parameters, were calculated as:

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

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. The inequality, $f_r \geq T_l$ where T_l = 0.7, characterized the amino acid as a high probability residue (HPR). The use of a threshold of 0.7 was a rational decision using known theory of molecular interactions and the structural and electrical properties of both the amino acids and nucleic acids.

Binding Site Prediction Algorithm: SeqCom

The algorithm for SeqCom propagates as follows:

1. Search the amino acid sequence for HPRs.
2. Assign integer values corresponding to position in the amino acid sequence to the HPRs.

3. Determine the difference between the HPRs. If the difference is greater than 3 then the region is considered as nonbinding. Else, the region is binding.

In the first step, predetermined amino acids of high statistical frequency at the binding sites of nucleic acid-binding IDPs are identified in the sequence of a polypeptide chain inputted into the algorithm. The positions of the HPRs in the sequence are stored in an array. If the spacing between the two positions of the HPRs is greater than 3, then the region is considered as a nonbinding region. By allowing for spacing of three amino acids between the HPRs, the formation of α-helices and β-strands at the binding site is allowed.

Binding Site Prediction Algorithm: IUPattern

IUPattern is an enhanced sequence composition algorithm that, similar to SeqCom, uses the binding site parameters to locate HPRs. The algorithm propagates as follows:

1. Search amino acid sequence in overlapping sections of four amino acids for:
   $\frac{\sum_{n}^{n+3} f_r}{4} > T_l$
   $\frac{f_r + f_{r+3}}{2} > T_l$
   $\frac{f_r + f_{r+2}}{2} > T_l$
   $\frac{f_{r+1} + f_{r+3}}{2} > T_l$.
2. Within each four amino acid block, the inequality evaluated in step 1 with the highest value while still being greater than T_l is used for determining which amino acids are marked as HPRs within the sequence.
3. The sequence is searched for patterns of HPRs that have patterns indicative of binding sites. These patterns are:
   Straight chain binding:
      Residue 1 through residue 4 are all HPRs.
   Alpha Helix:
      Residue 1 and reside 4 are HPRs.
   Beta-pleated sheet
      Residue 1 and 3 or 2 and 4 are HPRs.
4. Regions with listed patterns of HPRs in step 3 are predicted as binding sites.

As opposed to SeqCom which determines HPRs on a residue-by-residue basis, IUPattern determines the favorability of different combinations of amino acids in overlapping blocks of four amino acids. The combinations of these residues with the highest average frequency are marked as HPRs. Finally, the algorithm performs a search for patterns representative of possible binding site formations that include the determined HPRs. These sites are predicted as binding sites.

Binding Site Prediction Algorithm: IUPatternC

The algorithm for IUPatternC is the same as IUPattern; however, there is one additional step required before the final binding site predictions are made:

5. For regions of the amino acid sequence containing secondary structure, if the region contains an amino acid that breaks secondary structure formation – as defined by the Chou and Fasman parameters – then the sequence will considered nonbinding at i, i-1, and i+1, where i is the i-th amino acid (Orengo et al. 2003)

Although at a basic level, IUPatternC begins to incorporate the complexity of binding site prediction by considering local amino acid composition in the binding site predictions. The predictions made by IUPattern treat each amino acid as a single entity; however, in reality macromolecular folding and binding site formation in IDPs requires the coupling of local and long-range residues.

Benchmarking Binding Site Predictions

Benchmarking the binding site predictions made by SeqCom, IUPattern, and IUPatternC involved analyzing their predictive ability, a measure of the number of accurately predicted binding sites, and their accuracy, a measure of the correctly predicted residues involved in binding to the total number of residues predicted to be involved in the nucleic acid binding site.

$$Predictive\ Ability = \frac{Number\ of\ correctly\ predicted\ residues\ invoved\ in\ binding}{Number\ of\ residues\ involved\ in\ binding\ in\ the\ native\ structure}$$

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ residues\ invoved\ in\ binding}{Total\ number\ of\ residues\ predicted\ to\ be\ invovled\ in\ binding}$$

Both scoring methods are necessary to have a complete understanding of the algorithms' ability to predict binding sites. High accuracy does not imply high predictive ability, and high predictive ability does not imply high accuracy.

Results

Binding Site Prediction: SeqCom

The predictive ability (PA) and accuracy for SeqCom are provided in table 2. SeqCom has the highest predictive ability of all algorithms in this paper; however, it has to lowest accuracy. The DNA binding protein HMG-I(Y) has the highest predictive ability and accuracy compared to the other predictions produced by SeqCom. The standard deviation is relatively small for the predictive ability; however, the standard deviation for the accuracy is nearly twice as large, yet it is similar to the standard deviations for the predications by IUPattern and IUPatternC.

Table 2. Predictive ability (PA) and accuracy for all IDPs in our database. The average is provided.

	Score	
Protein (PDB code)	PA (%)	Accuracy (%)
1A36	97.3	46.1
1CQT	78.8	30.2
1IKN	78	28.9
1L3L	84	42.9
1QFQ	94.7	64.3
2EZD	100	76.2
2IY5	88.1	68.4
1KB2	78.6	44
1SER	86.4	45.2
2RGR*	81.8	3.15
Average	87.3 +/- 8.4	49.6 +/- 16.5

* Not included in the average due to erroneous result.

Binding Site Prediction: IUPattern

The predictive ability (PA) and accuracy for IUPattern are provided in table 3. IUPattern has neither the highest nor lowest predictive ability or accuracy. The predictive ability is 16.7% lower than for SeqCom; however, the accuracy is 8.5% larger than SeqCom. Similarly, the DNA binding protein HMG-I(Y) has the highest predictive ability and accuracy of all other proteins in our database. The standard deviations of 17.5% and 15.2% for the predictive ability and accuracy, respectively, are similar to the standard deviations we have observed for our other algorithms ran using the same dataset.

Table 3. Predictive ability (PA) and accuracy for all IDPs in our database. The average is provided.

	Score	
Protein (PDB code)	PA (%)	Accuracy (%)
1A36	86.3	45.0
1CQT	66.7	62.9
1IKN	40.7	33.8
1L3L	80.0	60.6
1QFQ	63.2	75.0
2EZD	100.0	76.2
2IY5	78.0	73.0
1KB2	57.1	44.4
1SER	63.6	51.9
2RGR*	72.7	4.21
Average	70.6 +/- 17.5	58.1 +/- 15.2

* Not included in the average due to erroneous result.

Binding Site Prediction: IUPatternC

The predictive ability and accuracy for IUPatternC are provided in table 4. While IUPatternC has the lowest predictive ability, it does have the highest accuracy of 61.2%. The decrease in the predictive ability for IUPatternC compared to IUPattern was only 2.4%, which is significantly smaller than the decrease observed for IUPattern from SeqCom. Similarly, the accuracy did increase; however, the increase was only 3.1% compared to the 8.5% observed from SeqCom to IUPattern. The highest predictive ability was observed for the prediction of HMG-I(Y), yet the

highest accuracy was observed for the antitermination protein N. The spread in the predictions, indicated by the standard deviation, is similar to other values for SeqCom and IUPattern.

Table 4. Predictive ability (PA) and accuracy of all IDPs in our database. The average is provided.

	Score		
Protein (PDB code)	PA (%)		Accuracy (%)
1A36	86.3		46.7
1CQT	63.6		63.6
1IKN	37.3		40.7
1L3L	80.0		69.0
1QFQ	52.6		83.3
2EZD	100.0		76.2
2IY5	72.9		72.9
1KB2	57.1		44.4
1SER	63.6		53.8
2RGR*	72.7		4.62
Average	68.2 +/- 18.9		61.2 +/- 15.3

* Not included in the average due to erroneous result.

Discussion

Comparison of Binding Site Prediction Results

Our results show that the development from SeqCom to IUPatternC shows decreasing predictive ability with increasing accuracy. This is not surprising given that our intention was to improve the accuracy of the binding site predictions since SeqCom was capable of predicting the majority of the binding sites, but it could not resolve binding from nonbinding regions, thus it had a low accuracy.

The impact on the predictive ability was far more substantial than anticipated for IUPattern. IUPattern was developed on the basis that many regions of an amino acid sequence may show the appropriate composition to form a binding site based on statistical parameters; however, it may be physically impossible to form a binding site based on a particular pattern of HPRs. Therefore, IUPattern required that the patterns of HPRs indicate the formation of α-helices, β-

stands, or extended strand formation. Clearly, with the decrease in the predictive ability, not all binding regions require these patterns. It may be interesting to investigate the allowable pattern formations or the disallowed patterns as there may be more well-defined disallowed patterns of binding site formation. Although there was a decrease in the predictive ability, there was a relatively significant increase in the accuracy. This suggests that particular HPR patterns are important, yet as discussed, we have not identified all allowed and disallowed patterns.

As with the development of IUPattern from SeqCom, IUPatternC was developed with the intent to decrease false positives, that is increase the accuracy of the predictions. This was achieved, yet it was accomplished with a small decrease in the predictive ability. IUPatternC requires the same pattern formations as IUPattern; however, the additional constraint requires that all amino acids in the binding region are favorable to exist in that particular type of structure (e.g., all amino acids in a region with an α-helix pattern should not be known to break α-helices). With the increase in accuracy of 3.1% from IUPattern to IUPatternC, it was evident that the local amino acid composition in binding site predictions was important. While we certainly would not expect the predictive ability to increase when a constraint was applied to the algorithm, we didn't necessarily expect a decrease in the predictive ability. As we did, in fact, observe that the predictive ability decreased, this may suggest two things: Either the Chou and Fasman parameters are in error, or more likely this suggests that although a region may have been predicted to form an α-helix or β-strand based on HPR patterns, this does not necessarily indicate that either an α-helix or β-strand will form in that region in the native structure.

In general, there is likely a flaw in our pattern predictions. Although we believe there to be inherent patterns in binding site formation, we are not identifying all allowable pattern formations, nor are we correctly predicting the type of structure that will form based on the patterns of HPRs currently employed. It will likely be necessary redefine the allowable patterns to prevent significant decreases in the predictive ability and to make use of the Chou and Fasman parameters in IUPatternC.

Future Work

As indicated, we do hope to redefine the allowable pattern formation for IUPattern as to incorporate those patterns that would not traditionally be considered typical. Additionally, we hope to incorporate a means of indicating the confidence in our predictions. A technique based on predictions using two sets of binding site parameters had been developed and incorporated into our original version of IUPatternC; however, as we could not get the original version of IUPatternC debugged completely, the confidence indicating method was not implemented. Finally, we hope to begin predicting secondary structure. It seemed promising to use the patterns elucidated in IUPattern and IUPatternC; however, this may not be useful given the recent results of our predictions.

For the application of these algorithms to others interested in IDP structure prediction, synthetic biology, predicting novel interactions in biological systems, and drug development it is likely not necessary to predict secondary structure. The most important development will be to improve the predictive ability and accuracy of the predictions through improved binding parameters and understanding of binding site structure. Additionally, confidence in the predictions will be

useful since 100% predictive abilities and accuracies will be difficult to achieve, particularly as we investigate the prediction of binding sites in IDPs that bind more than simply nucleic acids.

References

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.

Orengo C, Jones DT, Thornton JM, editors. 2003. Bioinformatics: Genes, Proteins, and Computers. Oxford: BIOS Scientific.

Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., Vacic, V., Obradovic, Z., and Uversky, V. N. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*. 9: 3323-3349.

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. of Mol. Bio.* 338: 1015-1926.

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 sequecnces for globularity and disorder. *Nucleic Acids Res.* 31: 3701-3708.

Shaffer, P. L., Gewirth, D. T. 2002. Structural basis of VDR-DNA interactions on direct repeat response elements. *EMBO*. 21: 2242-2252.

Tompa, P. 2005. The Interplay Between Structure and Function in Intrinsically Unstructured Proteins. *FEBS Letters*. 579 : 3346-3354

Uversky, V. N. 2002. What Does it Mean to be Natively Unfolded. *European Journal of Biochemistry*. 269 : 2-12

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.