Association of Binding Specificity and Disordered Regions in Zinc Finger Motifs

Takeshi Nagashima **Akihiko Konagaya**
t_shima@jaist.ac.jp kona@jaist.ac.jp
School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

Keywords: binding specificity, disordered regions, zinc finger

1 Introduction

Today, numerous data of genomic DNA and protein sequences are available with progress of the world's on-going genome projects. It has been acknowledged that the vast increase of available genomic sequence data makes the clarification of the functions for genes and proteins. Considering such circumstances, we focused on regulation of gene expression, especially relation between DNA binding proteins and nucleic acid sequences. Our preliminary study revealed that the binding factors can be divided into following four classes based on correlation between binding factors and binding sites :

class 1 single binding factor to single binding site (1-to-1),
class 2 multiple binding factors to single binding site (M-to-1),
class 3 single binding factor to multiple binding sites (1-to-M),
class 4 multiple binding factors to multiple binding sites (M-to-M).

The binding specificity may reflect its flexibility in the protein structure. Therefore, we predicted contents of disordered regions in all sequences belong to each class. According to our observation, much more disordered regions can be found in 1-to-M binding factors than 1-to-1 binding factors.

2 Materials and Methods

To confirm relation between binding specificity and disordered region content, we employed a simple neural network [2] as a discriminator and sequence attribute as a training data.

2.1 The data set

To discriminate ordered and disordered regions, we need sequences whose ordered and disordered regions are known to. For this requirement, we used sequences from *proteinlist* [4]. And to investigate correlation between binding specificity and disordered regions content, we used binding factors that belong to the *Zinc-coordinating DNA-binding domains Superclass* of TRANSFAC database [5]. Because zinc finger is assumed to be important for specific DNA binding [1], we used them.

2.2 Sequence attribute

An attribute is calculated from an amino acid sequence with specific window size to characterize that sequence. In this work, we employed 30 attributes. These attributes reflect 20 amino acid compositions and 10 properties of amino acid proportion. The last 10 attributes are assumed to be useful for discriminate the disordered and the ordered regions [3] and accordingly we selected them.

3 Results

We applied a simple neural network to discriminate the disordered and the ordered regions. The result shown in table 1 and 2. In table 1, the ratio of disordered regions in zinc finger motifs is denoted. In table 2, the ratio of sequences that have only ordered zinc finger motifs is denoted. Here, when every amino acid in a zinc finger is included in the ordered region, we call it ordered zinc finger. Note, any sequences that belong to M-to-1 class are not available and therefore there are only three classes represented in each table.

Table 1: Ratio of disordered regions in zinc finger motifs.

Class	DR
1 to 1	5.2 %
1 to m	18.0 %
m to m	19.3 %

Table 2: Ratio of sequences in which all zinc finger motifs are ordered regions.

Class	OZF
1 to 1	26.7 %
1 to m	23.8 %
m to m	8.0 %

Our results shows that 1-to-1 class contain the disordered regions one-third of the M-to-M class and contain the ordered zinc finger sequences four times greater than that of M-to-M. That is, when a transcription factor that contain zinc finger binds to DNA more specific, it contains fewer disordered regions in zinc finger and contents of sequences that have only ordered zinc fingers is higher than the other classes. So, we conclude that this result supports our hypothesis, that is "the multiple binding factors may result from flexible protein structures."

Acknowledgements

We would like to thank the members of Genetic Knowledge Systems Laboratory of Japan Advanced Institute of Science and Technology for fruitful discussions. This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from the Ministry of Education, Science, Sports, and Culture of Japan.

References

[1] Choo, Y. and Klug, A., Physical basis of a protein-DNA recognition code, *Current Opinion in Structural Biology*, 7(1):117–125, 1997.

[2] Romero, P., Obradović, Z., Kissinger, C., Villafranca, J. E., and Dunker, A.K., Identifying disordered regions in proteins from amino acid sequence, *Proc. IEEE International Conference on Neural Networks*, 1:90–95, 1997.

[3] Xie, Q., Arnold, G. E., Romero, P., Obravic, Z., Garner, E., and Dunker, A.K., The sequence attribute method for determining relationships between sequence and protein disorder, *Genome Informatics*, 9:193–200, 1998.

[4] http://disorder.chem.wsu.edu/proteinlist.html

[5] http://transfac.gbf.de/TRANSFAC/