Prediction of DNA Target Sites by Regulatory Proteins Based on Structure-Derived Potential

Hidetoshi Kono **Akinori Sarai**
hkono@rtc.riken.go.jp sarai@rtc.riken.go.jp
The Institute of Physical and Chemical Research (RIKEN)
3-1-1, Koyadai, Tsukuba, Ibaraki 305-0074, Japan

1 Introduction

Gene expression in higher organisms is controlled by a variety of regulatory proteins. Structural analysis of protein-DNA complexes has revealed that the same amino acids often interact with different bases and vice versa. Thus, no general rules yet exist to explain how proteins discriminate among DNA sequences or that predict their target sites. On the other hand, the amount of structural information on protein-DNA recognition has been increasing rapidly. The structures of more than 200 protein-DNA complexes are now registered in the Protein Data Bank (PDB) [1], and they serve as a rich source of information about the interactions between amino acids and base pairs at the atomic level. Statistical analysis of contact potentials between pairs of amino acids, empirically derived from protein tertiary structures, has been used to predict amino acid sequences that fold into particular structure [2]. We applied this strategy to the structures of protein-DNA complexes in order to predict DNA binding sites recognized by regulatory proteins.

2 Methods

The structures of 52 protein-DNA complexes were selected from the PDB making certain that redundant and low resolution data (resolution ≤ 3.2Å) were excluded. A set of pairwise potentials between DNA bases and Cα atoms of all amino acids was then empirically determined by statistical analysis of known protein structures using a modification of the method of Sippl [3]. For a pair of base a and amino acid b at grid point s, the potential is given by the following expressions:

$$\Delta E^{ab}(s) = -RT \ln \left[\frac{f^{ab}(s)}{f(s)} \right], \qquad f^{ab}(s) = \frac{1}{1+m_{ab}w} f(s) + \frac{m_{ab}w}{1+m_{ab}w} g^{ab}(s) \tag{1}$$

where m_{ab} is the number of pairs ab observed, w is the weight given to each observation, $f(s)$ is the relative frequency of occurrence of any amino acids at grid point s against any bases, $g^{ab}(s)$ is the equivalent relative frequency of occurrence of amino acid a against base b. R and T are gas constant and absolute temperature, respectively.

DNA sequences were fit to a template of protein-DNA geometry, and the energy potentials were calculated. The sum of the potentials for a sequence with a given length is defined as the energy for the sequences. The length of DNA sequence depends on the interface size of the binding proteins. The energy for the sequences in the crystal structures were characterized by their Z-scores against random sequences.

3 Results and Discussion

Our ability to predict DNA binding sequences from patterns of atomic interaction between bases and amino acids was examined by jack-knife test for 15 protein-DNA crystal structures exhibiting distinct characteristics.

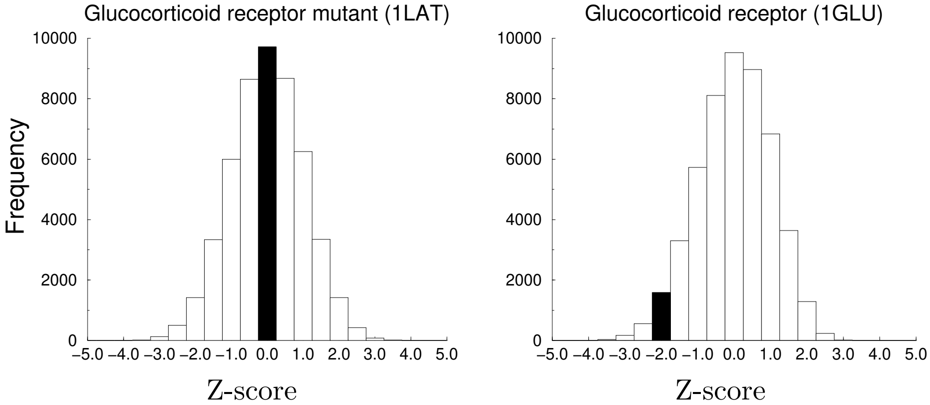

Figure 1: Histograms of Z-scores for random DNA sequences 50,000 base pairs in length. In each histogram, the positions of the co-crystallized DNA sequence with protein are shown as filled bars. In each case the database includes 52 complex-structures, except where the test complex was itself in the database, in which case it has been excluded.

For the 15 protein-DNA complexes, the co-crystallized DNA sequence with protein were detected with average Z-score of -2.8. This means that we can find the target DNA sequence within the top 3 among 1,000 random sequences. An interesting example for the structure-specificity relationship is the cognate and non-cognate complex structures of nuclear receptor. Gewirth and Sigler solved the crystal structure of an estrogen receptor (ER)-like DNA-binding domain (a glucocorticoid receptor (GR) DNA-binding domain altered by mutation) bound to the wrong type of half-site (a glucocorticoid response element; GRE) and revealed an interface resembling the specific interfaces of the GR or ER bound to their correct response elements [4]. The subtle difference in binding specificity can be tested by the present analysis. When this non-cognate complex (1LAT; Fig. 1 (left)) was used as a template, the GRE site was not detected ($Z = -0.1$). On the other hand, when a specific complex structure between GR and GRE (1GLU) [5] was used, a more favorable Z-score ($Z = -1.9$; Fig. 1 (right)) was obtained. Thus, this result indicates that our method can detect a subtle difference in binding specificity for target sites with some sequence variations. In addition, we considered effect of cooperativity on binding specificity, effect of DNA deformation and role of water molecules and the analyses have provided some insights into these effects.

We found that energy potentials extracted from distributions of Cα atoms around DNA bases of the known protein-DNA complex structures are sufficiently sensitive to detect the DNA binding sites of regulatory proteins. Moreover, this method can also be applied to proteins of unknown structure but with homology to known proteins, from which structures can be modeled and binding sites predicted. Increases in the amount of structural data should enable further refinement of the potential, and make this method a powerful tool for predicting multiple target sites and genes for regulatory proteins.

References

[1] Bernstein, F.C., *et al.*, The Protein Data Bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.*, 112:535–542, 1977.

[2] Bowie, J.U., Lüthy, R., and Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, 253:164–170, 1991.

[3] Sippl, M., Calculation of Conformational Ensembles for Potentials of Mean Force An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins, *J. Mol. Biol.*, 213:859–883, 1990.

[4] Gewirth, D.T. and Sigler, P.B., The basis for half-site specificity explored through a non-cognate steroid receptor-DNA complex, *Nat. Struct. Biol.*, 2:386–394, 1995.

[5] Luisi, B. *et al.*, Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA, *Nature*, 352:497–505,1991.