

Prediction of DNA Binding in Proteins from Composition, Sequence and Structure

Shandar Ahmad¹ Michael M. Gromiha²
shandar@rtc.riken.go.jp michael-gromiha@aist.go.jp

Akinori Sarai¹
sarai@rtc.riken.go.jp

¹ RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, 305-0074, Ibaraki, Japan

² Computational Biology Research Center (CBRC), AIST, 2-41-6 Aomi, Koto-Ku, Tokyo, Japan

Keywords: DNA binding, neural network, prediction

1 Introduction

In this work, we analysed the characteristic features of DNA binding proteins and their binding sites, in order to determine the factors that distinguish a binding protein from non-binding ones and a binding residue from other residues. The analysis has been made with different sets of data and properties. The datasets include (i) a non-redundant protein DNA complexes comprising of information about the structure and location of binding sites of DNA binding proteins and (ii) a non-redundant database of binding sequences for which structure information or the location of DNA binding sites is not known. The properties such as amino acid composition, sequence information, secondary structure, solvent accessibility and number of contacting residues have been employed in this work. We found that residue composition of DNA-binding proteins has two levels of specificity. One is at the sequence level, which can be used to classify sequences as binding or otherwise. The other is at the binding site level, which, when coupled with residue neighbourhood information and local structural information, can be helpful in locating binding sites in a totally new sequence even if no homology with previously known binding sequences existed. Solvent accessibility is found to correlate most strongly among the properties studied. On the whole there is no preference of residues to occur in any particular secondary structure, but there are interesting exceptions to this generalisation.

2 Method and Results

We calculated relative occurrence of residues in the binding region in each ASA range and results of such analysis are presented in Figure 1. Relationship between secondary structure and DNA-binding has been investigated for six main DSSP-defined secondary structures. Occurrence of the residue in binding regions, relative to non-binding regions was also investigated (Results not shown). Some residues have shown interesting preferences in this analysis, although the relationship is weaker than that of ASA.

Three different designs of neural network have been used for sequence based binding prediction. Residue and neighbor information are supplied to the input layer, which is then propagated through the network using linear activation function

$$X_{(i+1)k} = \Sigma(W_{ijk}X_{ij}) \quad (1)$$

where X_{ij} is the activation of the j^{th} unit of i^{th} layer and W_{ijk} is the connection weight between j^{th} units of i^{th} layer to k^{th} unit of the next layer ($i+1$).

The final output received at the output layer (single unit) is transformed to a value between 0 and 1 through a sigmoidal function

$$P = 1/[1 + \exp(-X_o)] \quad (2)$$

where P is the predicted probability and X_o is the activation of the unit in the output layer.

Experimental (desired) values of binding probability D for each residue is set as 0 or 1 depending on its being in the non-binding or binding state respectively. Error function is the sum of all the absolute errors in these probabilities for all the residues.

$$E_o = \sum |P - D| \quad (3)$$

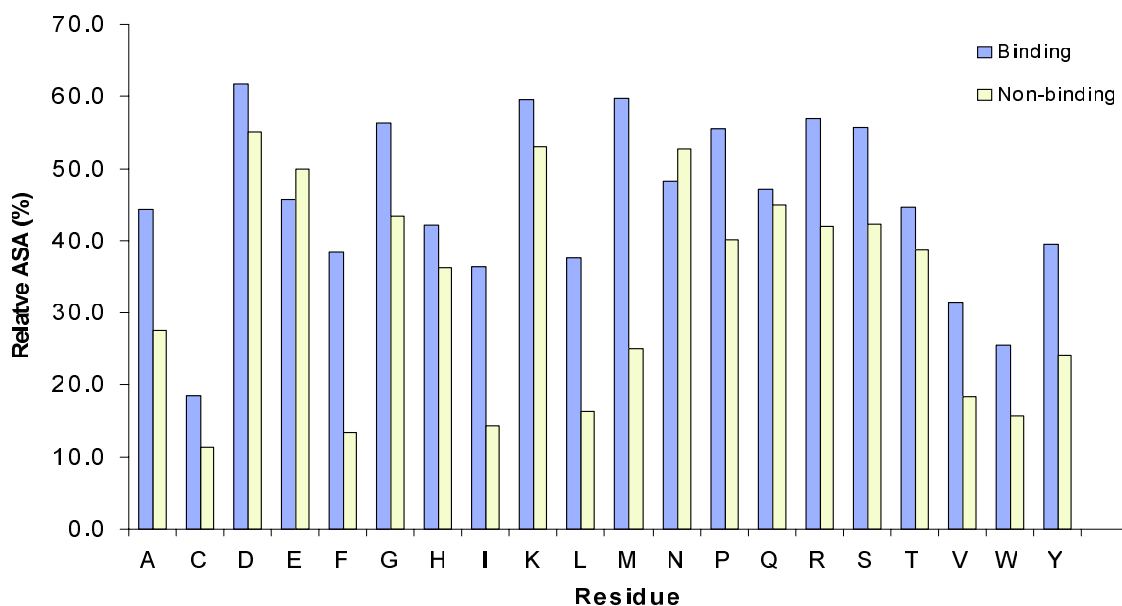


Figure 1: Residue ASAs in binding and non-binding regions.

First of these networks (*nnet-1*), simply inputs the residue and its two neighbors information, whereas in the second (*nnet-2*) and third (*nnet-3*) networks, we include residue ASA information as the input. The difference between them is that *nnet-2* feeds ASA information as a single bit and *nnet-3* provides it in a 21 bit vector, allowing to distinguish the residue again. Cross-validation of the datasets has been achieved by a procedure described elsewhere [1].

Sensitivity of prediction could be manipulated by introducing a bias so that the training error is defined as

$$E_b = E_o / \exp(k * P) \quad (4)$$

where E_o and E_b are unbiased and biased values of error respectively. k is a constant, which may be adjusted to move the most sensitive regions of our predictions from 0 to 1. Typical results of predictions obtained in this way are summarized in Table 1.

Table 1: Prediction results from different network designs for the cross validation data.

Network Inputs	Sensitivity(%)	Specificity(%)	Accuracy(%)
Nnet-1	40.6	76.2	73.6
nnet-2	32.2	86.2	79.9
nnet-3	40.3	81.8	79.1

References

- [1] Ahmad, S., Gromiha, M.M., and Sarai, A., Real value prediction of solvent accessibility from amino-acid sequence, *Proteins*, in press.