Gene 428 (2009) 25–30

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Sequence-dependence and prediction of nucleotide solvent accessibility in double stranded DNA

Shandar Ahmad *

National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki-shi, Osaka 567-0085, Japan

ARTICLE INFO

Article history:
Received 8 July 2008
Received in revised form 6 September 2008
Accepted 30 September 2008
Available online 9 October 2008

Received by L. Marino-Ramirez

Keywords:
DNA conformation
Solvent accessibility
Sequence dependent conformation
Prediction

ABSTRACT

Solvent accessibility of amino acid residues in proteins has been widely studied and many methods for its prediction from sequence and evolutionary information are available. Some of the advantages of studying amino acid solvent accessibility also apply to DNA. However, currently there are no methods to estimate the solvent accessibility of nucleotides, as most works on DNA structures have focused on elastic deformations and other structural attributes. In this work, an attempt has been made to analyze the distribution of different nucleotides in various accessibility ranges. Effect of neighboring nucleotides on the predictability of exposure has been evaluated by developing a linear perceptron model that takes sequence information as the input. Five different types of solvent accessibility (overall nucleotide, side chain, main chain, polar and non-polar) have been predicted. From the analysis, it is observed that Thymine stands out in terms of its higher exposed surface area, particularly its side chain and non-polar atoms. It is also concluded that the solvent accessibility of a nucleotide strongly depends on its sequence neighbors and can be predicted with fair success using this information.

1. Introduction

The concept of solvent accessibility was first introduced to understand and estimate the hydrophobic and solvation effects in proteins and their complexes (Lee and Richards, 1971; Sprang et al. 1979). Soon, solvent accessibility or accessible surface area (ASA) became an important descriptor of the protein structure itself (Lesk and Chothia 1980). Solvent accessibility of residues has been widely used in the identification of binding sites, scoring interfaces, function prediction and genome annotation (e.g. Gohlke et al. 2000; Kelley et al. 2000; Ahmad et al. 2004; Raih et al. 2005; Ofran et al. 2007). Based on its importance, methods to predict burial classes and real values at single residue as well as atomic level using sequence and evolutionary profiles have been developed (Rost and Sander 1994; Ahmad and Gromiha 2002; Ahmad et al. 2003; Garg et al. 2005; Araúzo-Bravo et al. 2006; Xu et al. 2006; Singh et al., 2006; Wang et al. 2007; Chen et al. 2008;) and it has been shown that even predicted values of ASA can be helpful in improving some of the functional properties of amino acid residues (Ofran et al. 2007). ASA has also been used as an important descriptor to predict deleterious mutations in proteins (Saunders and Baker 2002), especially because other well-known local structural descriptors such as secondary structure are less sensitive to single residue changes. Unusually exposed surface in proteins has been implicated in the prediction of protein-protein interacting residues (Raih et al. 2005).

DNA-protein interactions strongly control transcriptional and translational processes in the course of expressing genetic information (e.g. Riley et al. 2008). The interaction of DNA and proteins strongly depends on the structure of both the interacting molecules (Sarai and Kono 2005). Also, the behavior of DNA molecule depends on its conformation and in turn on the sequence, which plays a role in favoring or not favoring that conformation (Ahmad et al. 2006). Structural properties of DNA are typically characterized in the form of elastic deformations and knowledge-based potentials are derived from the observed frequencies of these deformations in different nucleotide positions (base steps) (e.g. Lankas et al. 2000). To the best of author's knowledge there has been no study which explicitly analyzes the nucleotide solvent accessibility in terms of sequence environments, although accessibility of nucleotides to proteins and solvents has been studied in different contexts and refers to a different type of accessibility (Robinson and Sligar 1998). There has been interest in nucleotide accessibility as indicated by experimental techniques to determine them (Zhang et al. 2003), but none of the published works represent a sequence-based fine-tuned analysis of nucleotide ASA. However, DNA conformation, including its ability to form double helix or unwinding strongly depends on hydration and solvent availability (Lee et al. 1981; Lipanov et al. 1989) and hence role of exposed solvent accessible area in DNA structure and function seems to have been under-estimated. Thus, solvent accessibility appears to be a useful descriptor of sequence-dependent parameters

Abbreviations: Ade, Adenine; ASA, Accessible surface area; Cyt, Cytosine; Cor, Coefficient of correlation; Gua, Guanine; MAE, Mean absolute error; Thy, Thymine.

* Tel.: +81 72 641 9848.
E-mail address: shandar@nibio.go.jp.

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.09.031

of DNA conformations, largely ignored thus far. In this context it may be mentioned that a single nucleotide change from GAATTC to TAATTC has been reported to change the number of water molecules released by EcoRI binding by 70 (Robinson and Sligar 1998).

In the current study, a statistical analysis of nucleotides with different ranges of ASA has been presented. A high quality data set of 3D structures in protein-bound or unbound form (Berman et al. 2000) has been compiled. Redundancy is removed and a representative set of sequences is selected. Solvent accessibility of the main chain, side chain, polar and non-polar atoms has been analyzed. Prediction of all five types of solvent accessibility has been made by a (i) linear perceptron and (ii) multilayer neural network. A linear expression to predict nucleotide ASA from the identity of target nucleic acid and its neighbors performs well and can estimate the nucleotide accessibility with a mean absolute error of 52 Å^2 using three neighbor information.

It is hoped that the prediction of nucleotide ASA from sequence information will be helpful in providing additional understanding of DNA–water–protein interactions, supply SNP-sensitive sequence descriptors for DNA and help in predicting functional implications of single to multiple synonymous as well non-synonymous DNA mutations.

2. Methods

2.1. Data sets

DNA structural coordinates are available from Nucleic Acids Database as well as Protein Data Bank (PDB) (Berman et al. 2000). PDB repository was searched for three-dimensional structures solved by X-ray diffraction, which resulted in 1948 entries, many of which are complexed with protein chains (complexes with RNA were excluded). Complexes containing Single stranded DNA structures were then removed by keyword search. Structures solved at poorer than 2.5 Å^2 were also removed at this stage. As a result, 681 DNA chains were obtained. However many of these sequences are similar and redundancy needs to be removed.

2.2. Redundancy removal

To obtain a representative, non-redundant data-set of DNA-sequences, standalone version of *Blastclust* program (part of *BLAST* package) is used (Altschul et al. 1990). Default word size of 32 in the program had to be reduced to 15 due to relatively smaller sequences in the database. Clustering is performed at 25% sequence identity and longest chain from each cluster is selected as the representative. Finally selected list of 85 DNA chains, obtained in this way are shown in Table 1, along with the chain length for each one of them.

2.3. Calculation of solvent accessibility

Solvent accessibility (ASA) of all DNA structures is computed using freely available program *NACCESS* (Hubbard and Thornton 1993). Original environments as in the complex or unbound structure available in PDB is used for these calculations, although it is not clear if complexation with protein affects the prediction results as sufficient data in bound and unbound categories are not available.

2.4. Nucleotide encoding

All target nucleotides (for which ASA prediction is carried out) and their sequence neighbors are represented by four-bit sparse vectors, in which one bit, representing nucleotide identity is set to 1 and all other bits are set to zero. Formally, complete environment of a nucleotide $\boldsymbol{N_i}$ (input to a prediction model) is represented as:

$$N_i := \{X_{i-n}, X_{i-n+1}, \ldots, X_i, X_{i+1}, \ldots, X_{i+n}\} \quad (1)$$

Table 1
PDB codes of DNA coordinates, their chain identifier (fifth letter), and their chain size selected for prediction and analysis

PDB ID	Size	PDB ID	Size	PDB ID	Size	PDB ID	Size	PDB ID	Size
1eqzI	146	1akhC	21	1mnmE	26	1i3jB	21	3crxF	34
1xo0C	35	1du0C	21	2ex5X	26	1k61E	21	1jeyD	31
1a74C	21	1tsrE	21	6paxB	26	1p78D	21	2owoB	26
1mowB	23	2a07A	21	1au7C	25	1t2tB	21	1ewqC	23
1kx5I	147	1hwtA	20	1bl0B	24	1tc3A	21	1ewqD	22
2vicD	26	1je8C	20	1d3uC	24	1yrnC	21		
1g9yC	24	1l3lE	20	1m5xC	24	1zrfX	21		
1murB	21	1lliD	20	1u0cC	48	2aorC	21		
2hddC	21	1zg1C	20	2fldC	24	2ntcC	21		
1ihfE	20	1zg5C	20	2i3pC	24	1b72D	20		
1ihfC	35	2or1A	20	2i3qC	24	1jnmC	20		
1kf1A	22	1k4tD	22	2pi4T	22	1ouzE	20		
2ht0D	22	1a6yD	20	2is6C	23	1perA	20		
1k78C	27	2vjuC	35	1bdtE	22	1pufD	20		
1h89D	26	1pvpD	34	1p47C	22	1rpeA	20		
1r8eB	23	2iszE	33	2b9sE	22	1ubdA	20		
1h6fC	24	2pi0E	32	3c25C	22	2hanC	20		
1r7mC	24	1ytbC	29	1am9F	21	2hapA	20		
1gxpC	23	1rioT	27	1f4kD	21	3crxE	35		

Where n is the number of nucleotide neighbors considered for a given prediction model, i is the position of that nucleotide within the DNA sequence and $\boldsymbol{X_i}$ is a four-dimensional binary vector such that $\boldsymbol{X_i}$: = $\{\delta_{ij}\}$; j = 1,4.

Where $\delta_{ij} = 1$ if $V(i) = V(j)$,
0 otherwise

$V(i)$ and $V(j)$ are the identities of nucleotide such as Adenine (Ade), Cytosine (Cyt), Guanine (Gua) or Thymine (Thy).

For unknown nucleotides or absent positions all bits in $\boldsymbol{X_i}$ are set to zero.

2.5. Prediction output vectors

In all prediction models reported in the current work, output is a five-dimensional vector with values in each dimension ranging from 0 to 1. These values correspond to five types of ASA for the nucleotide under consideration (all atom ASA, side chain ASA, main chain ASA, non-polar ASA and polar ASA). The original ASA values computed by *NACCESS* are transformed to finite range *[0,1]* by dividing all values by 400. After the training is completed, the predicted vector is multiplied by the same factor of 400 to obtain the ASA values on the original scale.

2.6. Performance measurement

All ASA values are in absolute area units (Å^2) and mean absolute error in this ASA is used to estimate the prediction performance of the predictors. Mean absolute error of a predictor is defined as:

$$\mathrm{MAE}(j) = 1/N \sum ||O_{ij} - P_{ij}|| \quad (2)$$

Where O_{ij} and P_{ij} are the observed and predicted values of (one of the five types of) ASA (identified by j) in Å^2 and the summation is carried out over all nucleotides (i).

In addition to MAE, coefficient of correlation is also used to measure the prediction performance and defined in a standard way as used in our previous works (Ahmad et al. 2003).

2.7. Linear perceptron

A linear perceptron consists of a simple linear function connecting the sparse-encoded nucleotide environment to the real-encoded