BIOINFORMATICS

Vol. 20 no. 4 2004, pages 477–486
DOI: 10.1093/bioinformatics/btg432

Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information

Shandar Ahmad[1,2,], M. Michael Gromiha[3] and Akinori Sarai[1]*

[1]Department of Biochemical Science and Engineering, Kyushu Institute of Technology, Fukuoka, Iizuka 820 8502, Japan, [2]Department of Biosciences, Jamia Mila Islamia, New Delhi 110025, India and [3]Computational Biology Research Center (CBRC), AIST, 2-41-6, Aomi, Koto-ku, Tokyo 135 0064, Japan

Received on February 26, 2003; revised on May 26, 2003; accepted on July 26, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: Though vitally important to cell function, the mechanism of protein–DNA binding has not yet been completely understood. We therefore analysed the relationship between DNA binding and protein sequence composition, solvent accessibility and secondary structure. Using non-redundant databases of transcription factors and protein–DNA complexes, neural network models were developed to utilize the information present in this relationship to predict DNA-binding proteins and their binding residues.

Results: Sequence composition was found to provide sufficient information to predict the probability of its binding to DNA with nearly 69% sensitivity at 64% accuracy for the considered proteins; sequence neighbourhood and solvent accessibility information were sufficient to make binding site predictions with 40% sensitivity at 79% accuracy. Detailed analysis of binding residues shows that some three- and five-residue segments frequently bind to DNA and that solvent accessibility plays a major role in binding. Although, binding behaviour was not associated with any particular secondary structure, there were interesting exceptions at the residue level. Over-representation of some residues in the binding sites was largely lost at the total sequence level, but a different kind of compositional preference was observed in DNA-binding proteins.

Availability: Online predictions of DNA-binding proteins and binding sites are available at http://www.netasa.org/dbs-pred/

Contact: shandar@bse.kyutech.ac.jp

INTRODUCTION

The rapid progress in genome analysis has made available the complete genome sequences of many organisms. Subsequent annotation of the genes, enabling their function to be inferred from sequence homology, is an important next step in the post-sequence analysis of genomes. In that regard, X-ray crystallographic and NMR spectroscopic analyses of DNA-binding proteins, which play key roles in the regulation of gene expression, have provided valuable information about the general features of protein–DNA interactions. In recent years, for example, Luscombe and Thornton (2002) analysed amino acid conservation and the effects of mutations on the binding specificity within protein–DNA complexes. Pabo and Nekludova (2000) developed geometrical models for characterizing side chain–base interactions and in related studies, Mandel-Gutfrend and Margalit (1998) and Mandel-Gutfrend *et al*. (1998) demonstrated the importance of hydrogen bonding and hydrophobic and CH $\cdots$ O interactions to protein–DNA interactions. Nadassy *et al*. (1999) analysed the importance of the interface surface area between the protein and the DNA for protein–DNA recognition. In addition, our systematic analysis of the contacts between amino acids and base pairs in a set of protein–DNA complexes has enabled us to construct models with which to predict the DNA target sites of regulatory proteins (Kono and Sarai, 1999). Since then, we have refined our analysis of base–amino acid interactions and have applied it successfully to cognate, non-cognate, symmetric and asymmetric binding of DNA-binding proteins (Selvaraj *et al*., 2002). Still, the mechanism underlying protein–DNA recognition is not yet completely understood.

In the present work, we analysed the characteristic features of DNA-binding proteins and their binding sites in order to determine the factors that distinguish binding proteins from non-binding ones and binding residues from all others. The data sets used for this analysis included (i) non-redundant protein–DNA complexes containing information about the structure and location of binding sites within DNA-binding proteins and (ii) a non-redundant database of DNA-binding proteins for which structural information and the location of DNA-binding sites were not known. Amino acid compositions, sequence information, secondary structure,

*To whom correspondence should be addressed at present address: Department of Biochemical Science and Engineering, Kyushu Institute of Technology, Fukuoka, Iizuka 820 8502, Japan.

solvent accessibility and the number of contacting residues were analyzed.

We found that the residue composition of DNA-binding proteins has two levels of specificity. One of them is at the sequence level, which can be used to classify sequences as binding or non-binding. The other is at the binding site level, which, when coupled with residue neighbourhood information and local structural information (particularly solvent accessibility), can be helpful for locating binding sites in a totally new sequence, even if there is no homology with known binding proteins. Among the properties studied, solvent accessibility was found to correlate most strongly with binding. On the whole there was no tendency for binding residues to occur in any particular secondary structure, though there were a number of interesting exceptions to this generalization. Predictive models developed to make use of these findings achieved a fair degree of accuracy.

MATERIALS AND METHODS

Data sets

Three different data sets were used for the studies described below:

(i) PDNA-62: This is a database of representative protein–DNA complexes from the Protein Data Bank (PDB) (Table 1) that we used previously in related studies (Selvaraj *et al*., 2002). Identity among the sequences is <25%, and the resolution of the structures is 2.5 Å or better.

(ii) NRTF-915: This is a non-redundant, representative database from SWISS-PROT used for composition-based prediction of DNA-binding proteins, which was limited to transcription factors for this study. A search in SWISS-PROT (version 39, available online at the time of this study) using 'transcription factor' as a key word returned 1003 SWISS-PROT and 2514 TrEMBL entries Boeckmann *et al*., 2003. Redundancy among sequences was first removed by using *CD-HIT* program from http://bioinformatics.burnham-inst.org/cd-hi (Li *et al*., 2001) with a threshold of 40% sequence identity. This resulted in 1486 sequences. We aligned these sequences against one another using pairwise sequence alignment program *bl2seq* (Altschul *et al*., 1997). A data set was thus created, by retaining only the representative ones such that no two sequences in the resulting data set have more than 25% sequence identity. We call this database NRTF-915. The criterion of sequence identity is similar to that used by Holm and Sander (1998) to cluster sequences to reduce redundancy.

(iii) CNTR-3332: A control database of sequences not including transcription factors was generated by searching for SWISS-PROT sequences excluded by the key word 'transcription factor' this database included sequences from human only, and fragments were omitted. Redundancy was removed as in (ii) and the resulting database contains 3332 sequences, called CNTR-3332.

Table 1. PDB codes of protein–DNA complexes selected for prediction of binding sites

1a02	1bl0	1dp7	1hdd	1mdy_a	1per	1tc3
1a74	1c0w_b	1ecr	1hlo	1mey_c	1pnr	1tf3
1aay	1cdw	1gat	1hry	1mhd_a	1pue_e	1tro_a
1azq	1cf7_a	1fjl_a	1hwt	1mnm	1pvi_b	1tsr_b
1b3t	1cjg	1gcc	1if1	1mse	1pyi_a	1ubd
1ber_a	1cma	1gdt	1ign_a	1oct	1rep_c	1xbr_a
1bf5	1d02_a	1hcq	1ihf	1par_b	1srs	1yrn_a
1bhm_a	1d66_a	1hcr	1lmb_4	1pdn	1svc	1ysa
1yui	2bop	2drp_a	2hdc	2gli	3cro_1	—

Definition of a binding residue and of binding density

An amino acid residue within a protein sequence was designated as a binding residue if its side chain or backbone atoms fell within a cutoff distance of 3.5 Å from any atom within a binding DNA. In this work, the term binding site has often been used to refer to the residues that are found to be binding as defined above. This is somewhat different from the more general meaning of binding site, which may refer to a region in protein, spanning several residues not necessarily sequence neighbours. All residues in the PDNA-62 database were labelled as binding or non-binding according to this criterion. Binding density was defined as the number of binding residues within a segment of fixed size and was calculated only for those segments in which the central residue was binding.

DNA-binding segments and sequence motifs

We collected statistics on three- and five-residue segments from DNA-binding proteins in the PDNA-62 database to explore the possibility of frequently occurring segments or sequence motifs. A segment was defined as binding if any of the atoms from its central residue fell within the cutoff distance from any atom of the DNA.

Local and overall amino acid composition

We collected statistics on amino acid residues around DNA atoms within a sphere of 3.5 Å. An attempt was then made to determine whether there was a preference for any particular amino acid composition. Frequency of occurrence for each residue type is calculated and corresponds to the relative number of residues of that type out of all the residues that were found in the DNA-binding region of proteins (as defined above). Thus, if there are R_i number

of residues of type i, which bind to DNA, then relative occurrence of i for the binding regions is given as $100 * (R_i / \Sigma R_i)$. The relative occurrence of residues in non-binding regions is also defined in the same way. In contrasts to the above, the overall composition of DNA-binding sequences refers to the residue composition in the whole amino acid sequence of the corresponding DNA-binding (or non-binding) protein.

Calculation of solvent accessibility or accessible surface area

Solvent accessibility or accessible surface area (ASA) values of 62 protein–DNA complexes were calculated using *DSSP* program (Kabsch and Sander, 1983). Absolute values of ASA, thus obtained are normalized to relative values as described in our earlier work (Ahmad and Gromiha, 2002).

Neural network design and training

For composition-based prediction of DNA-binding sequences, a fully connected, layered neural network was constructed using 20 input units encoding the relative abundance of the corresponding residue in a sequence. After varying the number of units, and hidden layers, it was found that a network with three units in the hidden layer and a single output unit performed slightly better than other choices. This neural network, used for the prediction of DNA-binding sequences is named *SeqPredNet*. The data sets used here were the NRTF-915 and CNTR-3332 sequences described above. The validation procedure was the same as in the other networks (see next paragraphs).

For sequence-based prediction of binding residues, three different architectures were implemented to introduce sequence and related information used in making predictions. In the first of these networks (we call it *SitePredNet-1*), the residue about which a prediction was to be made and its two nearest neighbours were encoded as 21-bit vectors (Ahmad *et al.*, 2003). The hidden layer consisted of two units, and the output layer had just one unit (Fig. 1). Information about the residue and its neighbours was supplied to the input layer, which was then propagated through the network using a linear activation function,

$$X_{(i+1)_j} = \sum (W_{ijk} X_{ijk}) \qquad (1)$$

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

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

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

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 was set to 1 or 0, respectively, depending on whether it was binding or not binding. The error function was the sum of all the absolute errors in these probabilities for all the residues:

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

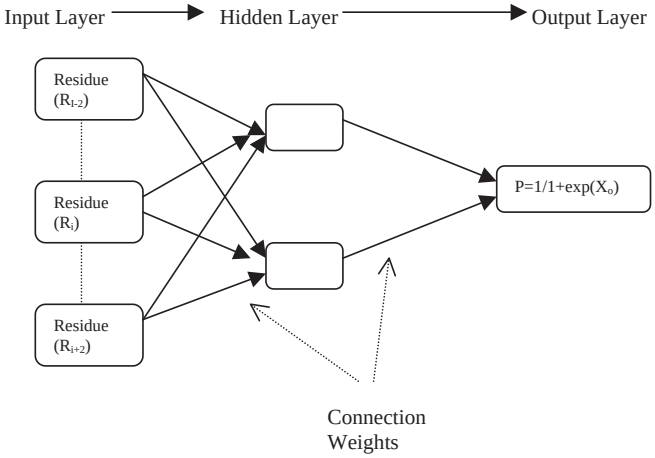

Fig. 1. Typical architecture of a neural network used for binding-site prediction. Each residue and its neighbour are represented by a 21-bit vector in the input layer. The activation of the j-th units in $(i + 1)$-th layer is given by: $X_{(i+1)j} = \Sigma W_{ijk} X_{ik}$. Output layer activation X_o is transformed by a sigmoidal function $P = 1/1 + \exp(X_o)$, which is taken as the predicted probability of the residue to bind DNA.

In the second network (SitePredNet-2), we included one additional unit that carried information about the ASA of the central residue in the input layer. In all other respects, *SitePredNet-1* was the same as *SitePredNet-2*. In the third network (*SitePredNet-3*), ASA information was sent through 21 units, all of which were set to zero except one that identified the residue type. The value of this unit was the same as the relative solvent accessibility or ASA calculated as described in the next section. This allowed for residue-wise variation in the ASA-dependence of the binding.

Training and validation in these three networks were accomplished by first expanding the PDNA-62 database into a residue-wise database of residues and their neighbours (two neighbours were selected), along with their desired binding state (1 for binding and 0 for non-binding). This database was then divided into three approximately equal parts. One part was used for training the network (training data). During the training the network weights were saved every time there was an improvement in the second (test) data prediction. The final set of weights thus represented the stage in the training history when the value of the error function output for the test data was minimal. The third part of the data set (validation data set) was kept aside from the training procedure and used for final validation of the predictions. The values reported in this work are actually those obtained from the

validation data. All six combinations of the three parts of the data set were used for training, testing and validation, and the reported values are the means of the validation results. Similar procedures for cross-validation have been used previously in related studies (Cuff and Barton, 2000; Ahmad *et al*., 2003).

Manipulating the desired sensitivity of prediction

The number of binding sites in the data set was quite small (about 12%); consequently, the networks trained to minimize the error function defined by Equation (3) might have a low sensitivity value. Selecting a lower probability threshold for predicting binding would likely change this sensitivity, but the added flexibility provided by this change would not be more than 5–10%. This is because the experimental values underlying the change in probability have only integer values (0 or 1), and the predicted probabilities also tend to have extreme values, causing the probability distribution to have sharp peaks near those values. This problem was overcome by choosing the following biased error function during network training:

$$E_b = E_o / \exp(k * P) \tag{4}$$

where E_o and E_b are unbiased and biased error values, respectively, and k is a constant that could be adjusted to give the most sensitive regions of the probability distribution values ranging from 0 to 1. However, this increase in the sensitivity of the binding prediction is at the cost of specificity. In order to compare the results of predictions from one network to others, we calculated the average of sensitivity and specificity [called 'net prediction' (NP)], which is a fair measure of the total amount of prediction quality obtained after training.

Accuracy scores

Correlation between the predicted (P) and desired (D) values was defined as the ratio of the covariance in the predicted and desired (experimental) states to the product of the standard deviations (SDs) in x and y:

$$R = S_{xy} / S_{xx} * S_{yy} \tag{5}$$

where

$$S_{xy} = \sum (P - P_o)(D - D_o),$$

$$S_{xx} = \sqrt{\left[\sum (P - P_o)^2\right]}$$

and

$$S_{yy} = \sqrt{\left[\sum (D - D_o)^2\right]}$$

Subscript 'o' represents the mean value of the corresponding variable. This correlation is the same as conventional Pearson's coefficient of correlation, which gives identical results as Matthews' correlation for a two-state prediction, as discussed previously (Ahmad and Gromiha, 2002).

Sensitivity and specificity of the predictions were defined as:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \tag{6}$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \tag{7}$$

(T-True, F-False, P-Positive, N-Negative).

Net prediction was introduced here as the average of the sensitivity and specificity.

$$\text{NP} = (\text{Sensitivity} + \text{Specificity})/2 \tag{8}$$

Accuracy scores were simply the ratios of the number of correct predictions to the total number of predictions made.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}).$$

In this work, term accuracy is employed to refer to this last parameter, whereas NP may serve as a better measure of predictability in some cases, especially when data sizes in the two states are not balanced.

P-value t-tests of significance were conducted for the assessment of statistical relevance of quantities such as the difference between solvent accessibility of binding and non-binding residues. P-value represents the probability that this difference could have occurred by a statistical chance and hence a lower P-value suggests a greater significance of that result. These P-values were obtained from the means, SD and number of data items using the online public domain t-test calculator available at http://home.clara.net/sisa/t-test.htm

RESULTS AND DISCUSSION

Compositional specificity of DNA-binding proteins and binding sites

The residue-wise composition of the DNA-binding sequences (NRTF-915) and binding sites (PDNA-62) were calculated, to assess any specificity present therein. Interestingly, the over-representation of some residues in the binding regions of 62 complexes was not reflected in the overall composition of transcription factor sequences. For instance, charged residues (e.g. Arg and Lys) were significantly over-represented within the binding sites of these 62 complexes (Fig. 2), but the same was not true for the overall composition of transcription factors (Fig. 3). This is mainly because the specificity of binding-site residues is lost in the overall composition due to the large number of non-binding residues present in the sequences. In addition, the bias in the overall frequency of occurrence for the DNA-binding sequences seems to differ from that of the binding sites within these sequences, suggesting there may be two levels of information present in DNA-binding proteins. Of particular significance is the residue serine, which is highly over-represented in DNA-binding sequences compared with the control data. Thus, serine seems to play an important role in the overall recognition of DNA targets. Higher occurrence of glycine in

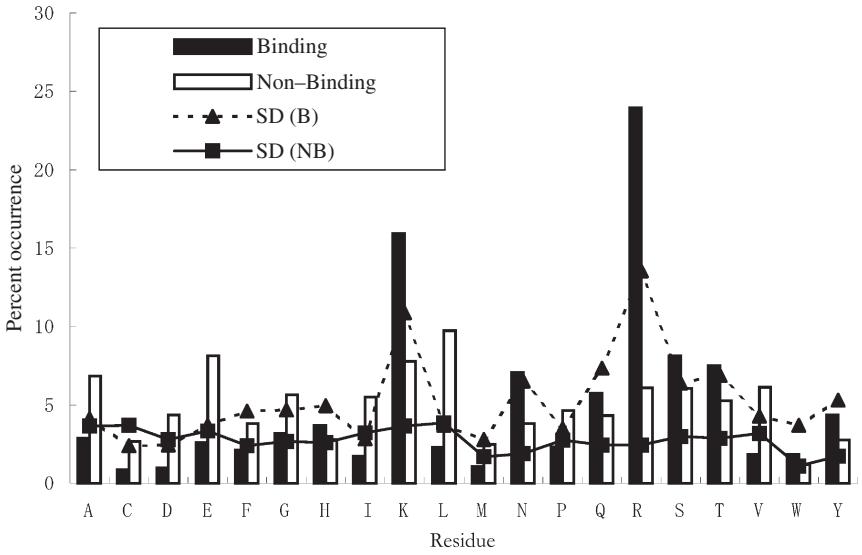

Fig. 2. Residue-wise occurrence of all residues in the binding and non-binding regions of 62 complexes. Percent occurrence of each residue type in the binding and non-binding regions were calculated and the graph shows averages and standard deviations over all 62 proteins. SD stands for standard deviation; B for binding and NB for non-binding regions.

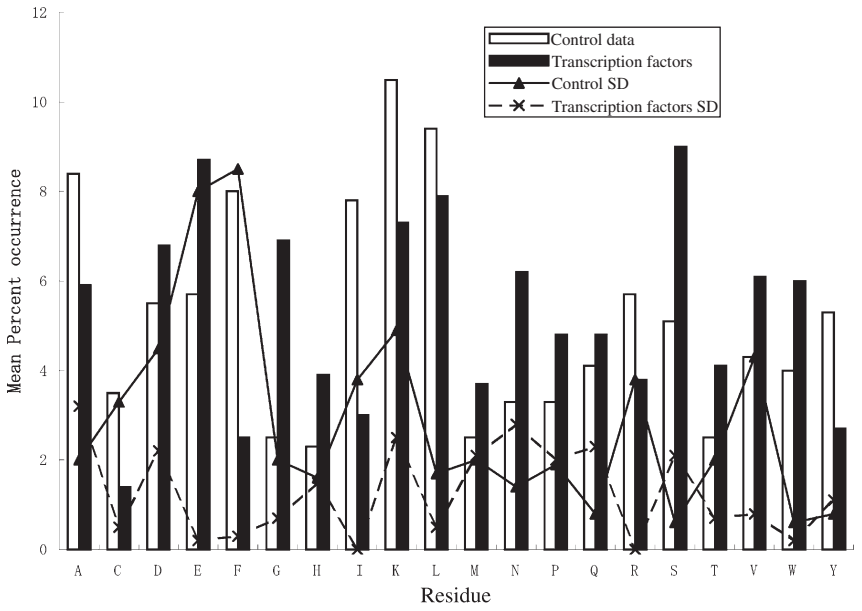

Fig. 3. Comparison of residue-wise composition of transcription factors (NRTF-915 data) and other proteins (CNTR-3332 data). SD stands for standard deviation in the corresponding data.

these proteins also suggests that this residue may contribute to conformational flexibility needed during the process of binding.

RESIDUE POPULATIONS AROUND DNA

We considered the residue populations around DNA while searching for specific environments that support binding. Interestingly, within the cutoff distance (3.5 Å), a very large number of nucleic acid bases (more than 50%) were found to be surrounded by just one residue, though there were several bases that had two, three and more amino acid residues within the cutoff distance. This suggests that, in some cases two different amino acid residues may contribute to the two hydrogen bonds needed for conferring specificity (Seeman *et al.*, 1976). It is quite possible that bases co-ordinated with

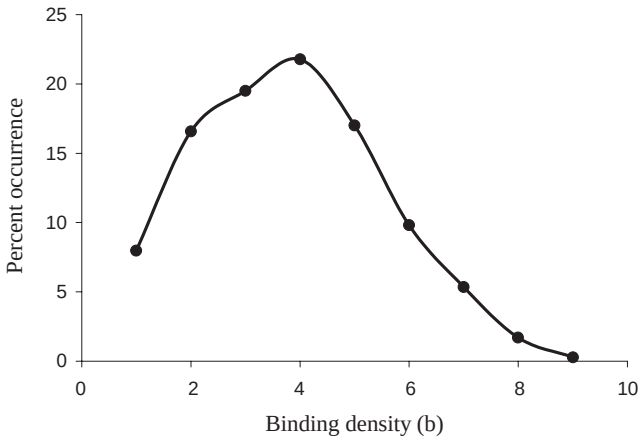

Fig. 4. Relative occurrence of 11-residue binding patterns with different binding density values in 62 protein–DNA complexes. Binding density is defined as the number of binding residues within 11-residue segments. Only segments with a binding central residue are plotted.

a higher number of residues may actually determine the overall binding, while those with just one residue within the cutoff distance may be of secondary importance.

Residue neighbours and binding site density

The numbers of DNA-binding sites within 11-residue segments approximates a normal distribution with a maximum at $b = 4$ (Fig. 4). This indicates that DNA-binding is unlikely to involve only a single residue–base interaction; more likely, several residues participate in the process. On the basis of that information, we collected a group of sequence fragments comprised of DNA-binding triplets and quintets. We found a total of 893 DNA-binding triplets out of a possible maximum of 1481 (the number of DNA-binding sites in the entire data set) from among 8000 possible triplets. Likewise, 1070 DNA-binding quintets were formed, and some notable examples of each are shown in Table 2. Some of these triplets and quintets are thought to be sequence motifs, given their relatively high frequency of occurrence and their high propensity for binding.

In addition, when these three- and five-residue patterns were arranged with respect to their central residue, we noticed that Arg residues were frequently surrounded by residues such as Lys or Arg, suggesting that certain neighbours enhance Arg's ability to bind DNA. Similarly, Lys residues within binding regions seem to favour Gly as their immediate neighbour on either side (an observation, which becomes more interesting in view of the finding presented above that Gly residues are more abundant in DNA-binding proteins). Interestingly, some of the five-residue patterns (e.g. AGIAT, ITRGS, MSQRE, SLKAA, MLTPD and IATIT) occur more frequently than would be expected and always bind to DNA. These patterns are therefore thought to support binding, and their occurrence in DNA-binding proteins may help in locating binding sites.

Table 2. Selected triplets and quintets and the number of times they happen to bind as compared with their total occurrence frequency

Sequence triplet/ quintet	Times binding	Binding propensity (%)
RER	6/13	46.1
QFN	4/6	66.7
KHQ	3/4	75.0
GKQ	3/3	100
GKS	3/3	100
KKI	4/4	100
VKC	3/3	100
PKT	5/7	71.4
PKG	3/5	60.0
SNS	4/4	100
PQF	4/6	66.7
QRE	4/4	100
RRC	3/3	100
TRG	6/7	85.7
RRI	4/5	80.0
LRW	4/6	66.7
SRA	4/7	57.1
RRK	11/36	30.6
KRK	11/42	26.2
SSK	4/4	100
ISN	3/4	75.0
MRERR	4/6	66.7
PQFNL	4/6	66.7
GAGIA	3/4	75.0
AGIAT	4/4	100
SLKAA	4/4	100
LPKVE	4/7	57.1
GSNSL	4/4	100
QFNLR	4/6	66.7
MSQRE	4/4	100
MPQFN	4/6	66.7
FSRSD	3/3	100
ITRGS	4/4	100
SQREL	4/4	100
DRRKA	4/6	66.7
NLRWP	4/6	66.7
RRRLS	4/6	66.7
TMRER	4/6	66.7
RGSNS	4/4	100
AATMR	4/6	66.7
MLTPD	4/4	100
IATIT	4/4	100

The relative number of times a pattern shows binding may be thought of as the binding propensity of the pattern.

ASA and the probability of a residue binding

We next attempted to establish and quantify the relationship between solvent accessibility and DNA-binding in two ways. We first calculated the number of binding residues within different solvent accessibility ranges. Figure 5 shows that the probability of binding systematically increased as the ASA increased. Double-sided probabilities (P-values) have been plotted to assess the significance of the difference, which have

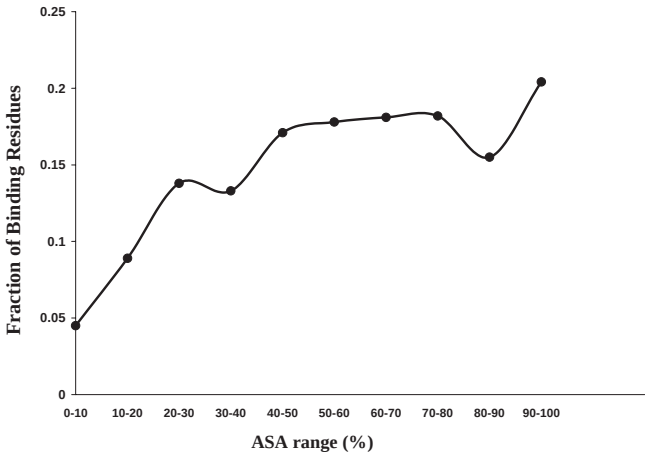

Fig. 5. Relative occurrence of residues in binding regions in different ranges of solvent accessibility.

been found to be low (implying higher significance) in most residues. These average ASA-values for the binding and non-binding regions revealed that the binding of hydrophobic and aromatic residues (e.g. Phe, Gly, Ile, Leu and Tyr) seemed to have a higher specificity for solvent accessibility than other residues (Fig. 6; P-value < 0.001). This may reflect the fact that a large number of these residues are in buried regions, so they are involved in binding only when they come to the surface. Consequently, there is a better separation between binding and non-binding residues in terms of their ASA.

Asn residues showed a negative correlation between binding and ASA. This negative relationship appears to arise because Asn forms a large number of turn structures, very few of which bind DNA: as turn conformations are highly accessible to the solvent, their low-binding propensity leads to a negative correlation between binding and ASA. This conclusion is however tentative due to a relatively higher P-value of difference in this residue. It also stands true for Glu for which a negative relationship between ASA and binding is accompanied by higher P-value. On the other hand, the negative relationship between ASA and Glu binding could be due to conformational changes necessary to expose certain atoms of Glu at the cost of others.

Binding probability and secondary structure

We evaluated six secondary structural states for every residue type and calculated the relative occurrence of each within binding regions. A t-test of significance as carried out for ASA does not seem to be viable due to a small number of binding sites, becoming even smaller when broken into each secondary structure for each protein. This leaves very little statistical significance for most of the data obtained in this category. We therefore looked at the *binding propensity* of amino acid residue on the whole. The binding propensity of a particular secondary structure was thus defined as the relative number of times a residue in that structure happens to be binding (Table 3). The last column of Table 3 gives the overall binding propensity of the corresponding residues in any of the six secondary structures. A value higher than this average for any residue implies a preference for binding within the secondary structure in question. In contrast to ASA, there was little preference for any particular secondary structure among the binding residues (see last row). Nonetheless, there were some notable features to these data. First, the highest value (50%) was obtained for Trp when located within a turn conformation. This may indicate that Trp has a strong preference for binding when it is in the turn conformation, but this conclusion cannot be asserted very strongly, as there are very few total number of Trp residues, occurring in this secondary structure. It was notable that Tyr and Arg also prefer binding in turn conformations. With the exception of Lys, it appears that residues with long side chains (Trp, Tyr and Arg) are more likely than average to be involved in DNA-binding when in a turn conformation. This may be because the greater length of their side chains provides additional flexibility in the turn conformation; thus binding may occur even if the backbone at these locations is a little far.

Hydrophilic residues such as Asn and Ser, show a preference for binding within a helical conformation, which may mean that they are able to make better contact with the DNA-helix when in that conformation. This greater tendency to bind when in a helical conformation is shared to some extent by Cys, His and Pro. In addition, Ala, Phe, Gly, Met and Val (mostly hydrophobic) appear to prefer a 3–10-helix over other structures. Gln also seems to have a preference for the 3–10-helix, but the number of 3–10- helices formed by Gln is less than 10 (including non-binding), making a conclusive statement about this residue difficult. Despite the apparent preferences shown by individual residues to bind within particular secondary structures, the overall abundance of binding residues within all secondary structures was very similar (except beta-strand, which has a very low binding propensity) (Table 3, bottom row).

Prediction of DNA-binding in proteins

To assess and utilize the compositional specificity of transcription factors, we trained a neural network (*SeqPredNet*) to make predictions about the sequences binding to DNA. This prediction simply calculates the relative abundance of each residue in the sequence and feeds the resultant 20-unit vector to the network. After trying different numbers of hidden layers and units therein, a network with three hidden units (described above) was found to give the best performance (Table 4). The fact that the best predictions were made when there were three units in the hidden layer indicates that the relationship between composition and binding is non-linear and that some interplay between compositions of two or more residues may be involved in giving a protein the capacity

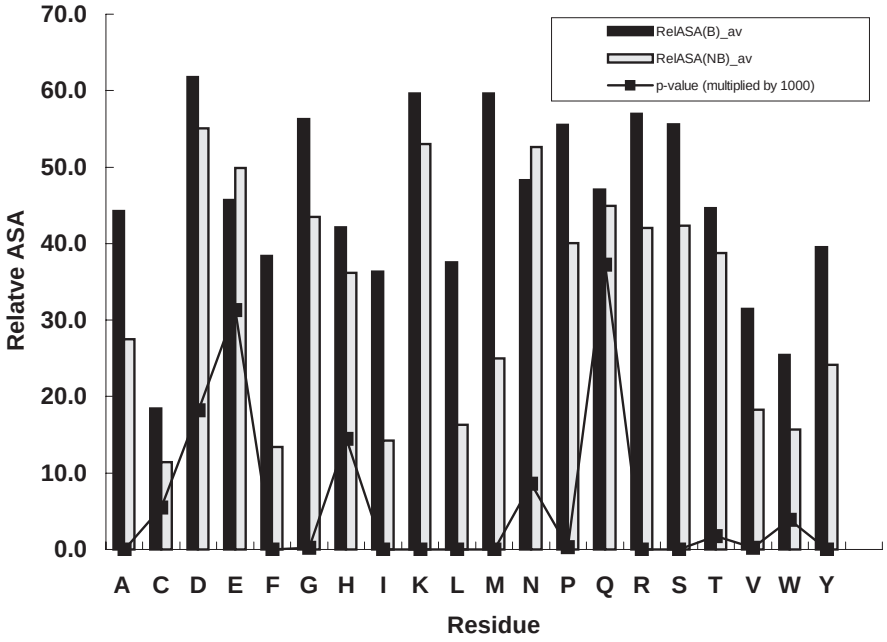

Fig. 6. Average ASA of residues in binding and non-binding regions and the *P*-values of the significance in their difference (lower *P*-value indicated greater statistical significance). B stands for binding and NB for non-binding residues. Corresponding *P*-values represent *t*-test probability that these differences could have occurred by chance. Most residues have a (statistically significant) higher ASA when in the binding state.

to bind DNA. A fairly high value of accuracy suggests that DNA-binding proteins have some characteristic residue composition that is, however, not identical to the composition of the binding sites.

Prediction of a residue's binding probability from its sequence neighbourhood

The results obtained when binding sites were predicted using the three neural networks described in Methods section are summarized in Table 5. There was a lot of fluctuation in the sensitivity and specificity of the predictions, which made it difficult to compare accuracy scores. In addition, introducing a biased error function during training altered the sensitivity of the predictions: the bias constant *k* of the error bias was varied from 1 to 2, so that the sensitivity of resultant predictions for the training and test data sets was at least 40%. Actual prediction scores were calculated by choosing a probability cutoff above which a residue is designated as binding. Results reported in Table 5 were obtained at a probability cutoff of 0.5. An additional 10–20% sensitivity could be manipulated by changing the probability threshold of binding.

For *SitePredNet-1*, predictions were made using just the residue information; for *SitePredNet-2* and *SitePredNet-3*, additional ASA information was used, which gave a small (2%) improvement in prediction quality when measured by 'NP', defined above. No significant improvement was achieved by adding information about secondary structure or about more distant neighbours. This may be due to the fact that local residue neighbourhood and structure carry only partial information about DNA binding. The remaining information must be contained in long-range contacts between distant residues. This inference is consistent with an earlier observation that stability and conformation in proteins are in part dependent on long-range contacts (Gromiha and Selvaraj, 2001; Gromiha *et al*., 1999). Moreover, due to the limited amount of available binding data, networks of arbitrarily large size cannot be used without diminishing the reliability of predictions.

Online prediction of DNA binding

Online predictions for new sequences based on the above methods are available at www.netasa.org/dbs-pred/. The detailed explanation of this prediction server (DBS-Pred) may be seen at www.netasa.org/dbs-pred/description.html.

CONCLUSIONS

The composition of binding regions differs significantly from non-binding regions; however, the compositional specificity of binding sites does not extend to the total composition of the protein sequence, where it takes a different form. Although binding residues show no overall preference for a secondary structure, some suggestive biases were observed (without much statistical significance, due to the

Table 3. Residue-wise binding propensity of six secondary structures

Residue	Helix	Beta	Ex-beta	G (3-10)	T (Turn)	S (bend)	Overall
A	3.4	0	3.5	14.3	5.4	7.0	5.0
C	7.0	0	5.7	0	0	5.6	4.5
D	4.8	0	0	0	5.2	1.7	3.3
E	4.0	0	1.9	5.3	1.1	7.5	3.8
F	4.0	0	6.6	18.2	5.6	0	7.4
G	6.3	0	7.0	14.3	6.0	10.2	8.1
H	20.3	0	13.0	5.9	8.8	23.3	15.2
I	3.3	0	3.6	0	0	5.6	4.1
K	21.8	0	26.5	9.1	19.8	21.9	23.0
L	2.7	7.1	0.6	0	5.6	5.2	2.7
M	1.5	0	2.2	40.0	4.3	4.5	5.6
N	29.3	40.0	24.4	28.6	8.7	15.0	20.1
P	8.2	0	2.4	13.0	3.5	6.3	6.3
Q	17.0	0	18.2	37.5	3.9	17.1	15.0
R	36.2	0	27.9	22.2	39.4	38.6	35.4
S	22.0	0	5.7	10.5	1.3	12.2	15.2
T	16.4	20.0	10.1	10.0	17.2	18.8	15.9
V	1.4	0	2.4	12.5	6.9	10.9	3.4
W	17.8	0	7.3	0	50.0	0	18.1
Y	18.6	0	11.5	7.7	24.0	26.7	16.6
Total	12.6	3.6	8.6	10.5	9.5	13.1	11.8

This table shows the percent-binding event of the indicated secondary structures for every residue, with the overall shown in the last column. If the propensity of a residue in a given secondary structure is more than the corresponding value in the last column, it would generally suggest a preference of the residue to bind in that conformation. Due to scarcity of data in each of these states, these values are merely suggestive and not conclusive. The last row shows that there is no real statistically preferred secondary structure in which all the residues will prefer to bind. Some propensities very different from averages have been highlighted.

Table 4. Mean validation data prediction results (*SeqPredNet*) for the composition-based prediction of binding sequences at 50% probability cutoff

	Accuracy (%)	Sensitivity (%)	Specificity (%)	NP (%) (Sens. + Spec.)/2
Self-consistency	71.1	71.3	71.0	71.2
Cross-validation	64.5	68.6	63.4	66.1

scarcity of data at this stage)—i.e. some residues seem to prefer to bind when in a particular secondary structure. DNA-binding sites can be predicted with moderate success based on sequence information alone, and introduction of solvent accessibility information improves those predictions. Using the sequence information as the input, DNA-binding sites are predicted and online system for such predictions is provided. A residue-level approach to binding complements the motif-based approach, in the sense that residue-wise information is likely to have a stronger capacity to locate previously undetermined binding motifs and those binding regions, which may not be really conserved enough to form a detectable motif. This method of prediction will be especially useful when a new binding protein is found with no significant homology with any known binding protein.

Table 5. Binding site cross-validation data prediction results from different network designs

Network inputs	Sensitivity (%)	Specificity (%)	Accuracy (%)	NP (%) (Sens. + Spec.)/2
Residue + 2 nbrs (*SitePredNet-1*)	40.6	76.2	73.6	58.4
Residue + 2 nbrs + 1bit ASA (*SitePredNet-2*)	32.2	86.2	79.9	59.3
Residue + 2 nbrs + 21bit ASA (*SitePredNet-3*)	40.3	81.8	79.1	61.1

REFERENCES

Ahmad,S. and Gromiha,M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–24.

Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) Real value prediction of solvent accessibility from amino-acid sequence. *Proteins*, **50**, 629–635.

Altschul,S.F., Thomas,L.M., Alejandro,A.S., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Gromiha,M.M., Oobatake,M., Kono,H., Uedaira,H. and Sarai,A. (1999) Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**, 549–555.

Gromiha,M.M. and Selvaraj,S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.*, **310**, 27–32.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.

Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, **17**, 282–283.

Luscombe,N.M. and Thornton,J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.

Mandel-Gutfrend,Y. and Margalit,H. (1998) Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.

Mandel-Gutfrend,Y., Margalit,H., Jernigan,R.L. and Zhurkin,V.B. (1998) A role for CH · · · O interactions in protein–DNA recognition. *J. Mol. Biol.*, **277**, 1129–1140.

Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.

Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.

Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.

Selvaraj,S., Kono,H. and Sarai,A. (2002) Specificity of protein–DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.*, **322**, 907–915.