5922–5932 *Nucleic Acids Research, 2008, Vol. 36, No. 18* *Published online 18 September 2008*
doi:10.1093/nar/gkn573

Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins

Shandar Ahmad[1,2], Ozlem Keskin[3], Akinori Sarai[4,*] and Ruth Nussinov[5,6]

[1]National Institute of Biomedical Innovation, 7-6-8, Saito-asagi, Ibaraki, Osaka 567-0085, [2]Graduate School of Frontier Biosciences, Osaka University, Japan, [3]Koc University, Center for Computational Biology and Bioinformatics, College of Engineering, Rumeli Feneri Yolu, Sariyer, 34450, Turkey, [4]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, 820-8502, Japan, [5]Center for Cancer Research Nanobiology Program, SAIC, NCI-Frederick, MD, USA and [6]Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Israel

Received May 16, 2008; Revised August 4, 2008; Accepted August 25, 2008

ABSTRACT

Amino acid residues, which play important roles in protein function, are often conserved. Here, we analyze thermodynamic and structural data of protein–DNA interactions to explore a relationship between free energy, sequence conservation and structural cooperativity. We observe that the most stabilizing residues or putative hotspots are those which occur as clusters of conserved residues. The higher packing density of the clusters and available experimental thermodynamic data of mutations suggest cooperativity between conserved residues in the clusters. Conserved singlets contribute to the stability of protein–DNA complexes to a lesser extent. We also analyze structural features of conserved residues and their clusters and examine their role in identifying DNA-binding sites. We show that about half of the observed conserved residue clusters are in the interface with the DNA, which could be identified from their amino acid composition; whereas the remaining clusters are at the protein–protein or protein–ligand interface, or embedded in the structural scaffolds. In protein–protein interfaces, conserved residues are highly correlated with experimental residue hotspots, contributing dominantly and often cooperatively to the stability of protein–protein complexes. Overall, the conservation patterns of the stabilizing residues in DNA-binding proteins also highlight the significance of clustering as compared to single residue conservation.

INTRODUCTION

In protein–protein interfaces, conserved residues have been widely studied and shown to correlate with hotspot residues (1–5). A hotspot is a residue whose mutation to alanine leads to a drop of over 2 kcal/mol in the binding free energy (6). Conserved residues are tightly packed and form clusters dubbed 'hot regions'. Due to the tight packing, it was proposed that within hot regions residues contribute to the stability of the complex cooperatively; in contrast, between hot regions, the contributions of hotspot residues are independent, i.e. additive (1,7). The relationship between conservation, structural hotspots and protein function has also been established in protein interactions with other types of molecules (8,9).

Over the last few years, protein–DNA interactions have received considerable attention. Recent studies include development of physical models from high-throughput data (10), prediction of p53 affinity for DNA elements (11), mutual information on the protein and DNA (12), prediction of DNA-binding proteins (DBPs) and their binding sites or specificity from sequence and structure (13–17) and genome-wide predictions of transcription factor binding affinities (18). Position-specific scoring matrices (PSSM) have identified DNA-contacting residues with reasonable success (14,19). However, (i) they ignore cooperativity between structurally close residues; and (ii) focus on the prediction of contacting residues irrespective of their contribution to the stability.

Earlier studies (20–23) analyzed additivity of protein–DNA interactions from the nucleic acid perspective. Conservation analysis has also focused on DNA-bases and base pairs rather than amino acids (e.g. 24). In some cases, studies were extended to take into account small set of proteins such as a single family of DBPs (e.g. 25,26). Most studies concluded that the interactions

*To whom correspondence should be addressed. Tel: +81 948 29 7811; Fax: +81 948 29 7841; Email: sarai@bio.kyutech.ac.jp

are not additive. However, to the best of our knowledge, a comprehensive analysis of conserved residues in DBPs, and their relationship to stability and clustering has not been carried out. Such an analysis would assist in understanding residue couplings in protein–DNA interaction and in predictions from the sequence and structure of proteins, not only the functional residues but also the contributions of residue interactions.

In this work, we analyze the stability, structural properties and clustering patterns of residues in DBPs. Since alanine-scanning data are unavailable for protein–DNA interfaces, we compile a list of mutations in protein–DNA complexes for which experimental free-energy changes are available in the thermodynamic data of protein–DNA interactions (ProNIT) (27). Due to the analogy between the most destabilizing mutations and hotspots in protein–protein interfaces, we dub the residues whose mutations led to the highest change in stability 'putative hotspot residues'. We observe a dependence of the free energy on both the conserved residues and the number of conserved neighbors of residues with the highest change in stability, indicating a larger contribution of conserved residues to the stability of a complex.

MATERIALS AND METHODS

Data sets

Stability data (smddg). Single mutation free-energy data were extracted from our thermodynamic data of nucleic acid interactions, ProNIT (27). This is a regularly updated database of experimentally known observations of free-energy values of binding between DNA and wild type and mutant proteins. Only single amino-acid mutations, with full structural and thermodynamic information have been considered in the current work. The final data consist of 511 entries. These data are called the *smddg* data and are provided in Supplementary Table S1. The calculated conservation scores and the number of conserved neighbors for each of the 511 entries calculated at different conservation score (Co) cutoff values are also included in the same file. The free-energy change upon mutation has been calculated as

$$\Delta\Delta G = \Delta G(\text{mutant}) - \Delta G(\text{wild}) \quad \mathbf{1}$$

A higher value of $\Delta\Delta G$ for a given mutation indicates larger destabilization by the mutation. Residue identity and any possible mutations in the corresponding DNA sequence have not been considered due to small data size.

Protein–DNA complex data (PDNA140). A complete list of protein–DNA complexes, with resolution better than 2.5 Å, solved by X-ray diffraction was downloaded from the Protein Data Bank (PDB) (28) (the August 2007 release). The list consists of 1178 protein chains. These chains are made up of 208 clusters (at 25% sequence ID), obtained by BLASTCLUST (29). The chain with the highest number of DNA contacts was selected from each one. Complexes with less than 10 bases or with breaks and structure anomalies were removed. Sequences with <10 homologs in the NCBI NR data (required for obtaining conservation) were also removed from the list. These conditions of quality and redundancy were satisfied by 140 chains and were used here. The list of selected proteins is provided in Supplementary Table S2.

Structural classification of proteins

The classification of DBPs is as reported by Luscombe *et al.* (30). This classification divides the DBPs into eight groups, namely, Group I (helix-turn-helix(HTH)), Group II (zinc-coordinating), Group III (Zipper type), Group IV (other α-helical), Group V (β-sheet), Group VI (β-hairpin/ribbon), Group VII (enzymes) and Group VIII (histone/histone-like) proteins. A ninth group for unclassified proteins is also used. All 140 protein chains (PDNA140) were manually assigned to these groups based on their structural characteristics or previously available assignments.

Multiple alignments

Multiple alignments with similar sequences were obtained by carrying out a Basic local alignment search tool (BLAST) search of the NCBI NR database, choosing a maximum of 50 aligned sequences for the multiple alignments (29). Multiple alignments were obtained using *clustalw* (31) with default parameters and Gonnet substitution matrix (32).

Conservation scores

The SCORECONS web server was used to convert multiple alignment outputs into conservation scores (33). According to this method, conservation scores for individual residues may be obtained using normalized substitution matrices and multiple alignments of a set of sequences.

Clustering procedure

A C program was written to obtain the number of conserved neighbors and identify unique clusters of conserved residues. First, the geometric center of all residues was calculated. Conservation scores and other computed properties (e.g. Accessible Surface Area (ASA), DNA contacts) were also read from the corresponding files. Using simple distance calculations, the number of conserved neighbors for each residue was calculated. To obtain the unique clusters of conserved residues, first each conserved residue was selected as a cluster seed. Then clusters were allowed to evolve by systematically calculating the distance between new residues with all the current members of the (evolving) cluster. If the geometric center of the new residue was in contact (6 Å) with any other member of that cluster, this residue was assigned to the cluster. This leads to a number of overlapping clusters, totaling the number of conserved residues (one cluster for each seed residue). The unique set of clusters obtained from different seeds were compared and if two clusters had any common members, the smaller cluster is discarded or if the two clusters are identical, only one is retained. This lead to a nonoverlapping unique set of clusters. Five iterations ensured cluster convergence.

Computation of packing density

The definition of packing density is as described earlier (1): packing density is calculated as the number of residues whose C_α position falls within a sphere of 6 Å radius from the C_α position of a target residue. In the packing density definition for DNA nucleotides, the C_α position is replaced by a backbone phosphate. Definitions based on the residue's geometrical center were also tested, giving similar statistics and hence the corresponding results are not included in the article.

Computation of solvent accessibility

ASA was calculated using the Dictionary of secondary structures of proteins (DSSP) program (34) and normalized as in our previous works on solvent accessibility prediction (35).

Computation of DNA contacts

A residue was defined to be in contact with the DNA if any of its atoms fell within a cutoff distance from any DNA atom. The cutoff distance is 3.5 Å, as in our previous works on the prediction of DNA-binding sites (36). A recent study shows that this distance for any-to-any atom contact gives the best prediction performance in machine learning methods (37). A cluster was said to be in contact with DNA if any of its member residues has a DNA contact.

Statistical significance

In many cases, the difference between the means of the data sets does not provide sufficient information about its statistical significance due to the large standard deviations. In most of these cases a two-tailed Student's *t*-test for determining statistical significance of the difference between means can ascertain the statistical validity. These tests are carried out in the open source programming environment of *octave* using their *t_test_2* function (http://www.gnu.org/software/octave).

Linear predictor optimization

A unique cluster of conserved residues is defined as an interface cluster if any atom of any residue in that cluster falls within a 3.5 Å distance from any atom of the DNA. All other clusters are labeled as noninterface or nonbinding. A linear predictor tries to predict if a cluster is in the interface (binding) or nonbinding. The relative amino acid composition of a cluster is reduced to four dimensions by grouping residues into four categories, namely, hydrophobic (Ala, Cys, Phe, Ile, Leu, Met, Pro, Val and Trp), hydrophilic (Gly, His, Asn, Gln, Ser, Thr and Tyr), negatively charged (Asp and Glu) and positively charged (Arg and Lys). A linear relationship is then defined as follows:

$$\text{State of the } i\text{-th Cluster} = \begin{cases} \text{binding} & \text{if } \chi_i > \lambda_o \\ \text{nonbinding} & \text{otherwise.} \end{cases}$$

where,

$$\chi_i = \lambda_{pho}\rho_i(\text{pho}) + \lambda_{phil}\rho_i(\text{phil}) + \lambda_{neg}\rho_i(\text{neg}) + \lambda_{pos}\rho_i(\text{pos}) \quad \mathbf{2}$$

ρ is the relative frequency of occurrence of the corresponding residue type within a cluster, phil stands for hydrophilic, pho for hydrophobic, neg for negatively charged or acidic residues and pos for positively charged or basic residues (the sum of the four ρ-values for each cluster is 1). λ's are the coefficients, obtained by optimizing the prediction. Optimization is performed with cross-validation to estimate predictability and by a simple self-consistency model fitting to obtain the most suitable parameters. In the cross-validation scheme, the entire set of clusters was divided into 10 parts and in each training step nine parts are combined to train the model and the 10th left-out part is used to assess the prediction performance. After 10 cycles of training and testing, the average prediction scores were computed. Measuring prediction performance on the left-out test data ensures that the scores are not exaggerated for the data under consideration, but would be useful for new clusters, not used in training. For the self-consistency set, trained parameters were determined as: $\lambda_{pho} = 0.096$, $\lambda_{phil} = 0.951$, $\lambda_{neg} = -0.073$ $\lambda_{pos} = 10.741$, $\lambda_o = 0.55$. Due to the small number of independent parameters, no over-training was observed.

Definitions

Conserved region: A set of all residues in a protein which satisfy the condition that their conservation score $C \geq C_0$. C_0 is the threshold at which residues are labeled as conserved. A default value for C_0 is fixed at 0.8.

Number of conserved neighbors: The number of conserved residues whose geometric center lies within a cutoff distance (6 Å) of the geometric center of the target residue. The number includes the target residue and therefore has a minimum value of one for residues with no conserved neighbors.

Clustered-conserved regions/residues: A subset of conserved residues with at least one conserved neighbor.

Conserved residue singlet: Conserved residues with no neighboring conserved residue.

Unique clusters of conserved residues: A cluster of contiguous conserved residues such that no pair in the cluster in a protein has any common residue members with it and all residues in the cluster are conserved. Assignment of residues to such cluster follows clustering criteria.

Putative hotspot residues: Assignment is based on experimental data of free-energy changes. Residues in the *smddg* database are ranked by the free-energy change on mutations in these positions (most destabilizing mutations ranked higher). The *N*-highest ranking mutant positions are termed *putative hotspots*.

RESULTS AND DISCUSSION

Conserved residues in protein–DNA complexes

We first analyze the distribution of conserved residues in 140 protein–DNA complexes, selected as described in Materials and methods section.

Overall occurrence. Conserved residues may occur as singlets (one conserved residue) or associated with other conserved residues. In Figure 1a, the number of residues occurring as singlets or with conserved neighbors is plotted. These values are calculated for each protein separately and error bars show their variations amongst all proteins. For most residues, the number of conserved neighbors ranges between one and seven; very few residues have more than seven conserved neighbors. Most conserved residues (>10 per protein) have three conserved neighbors, and only a small number (about five residues per protein) occur as singlets. Overall, there are 659 singlet and 5552 clustered-conserved residues. Thus, clustered-conserved residues are about 8.4 times as abundant as conserved residue singlets. Figure 1b and c compare the observed clustering with a random distribution, obtained by reassigning the conservation scores in the sequence and re-clustering them in the same way. Both figures show that DBPs have substantially fewer singlets and larger unique clusters in these proteins are far more frequent than in a randomly generated distribution. This indicates that conserved residues are not scattered in the protein structures but form (tight) clusters (see subsection on packing density). Supplementary Figure S1 gives additional statistics, showing that the occurrence of conserved residues in clusters versus singlets is not peculiar to some proteins; rather it is a general property of almost all proteins.

Distribution on the surface. We analyzed the distribution of conserved residues, and conserved residue clusters on the surface in different solvent accessibility (ASA) ranges. Detailed results are provided in Supplementary Figure S2. Conserved residues and conserved-clustered residues occur more frequently in the lower ranges of the ASA, with a ~10% difference in the relative abundance of these two types of residues compared with that of the entire database. The difference between these frequencies falls sharply and in higher ASA ranges the distributions are very similar. The higher percentage of both residue categories in buried regions could reflect structural constraints. In summary, conserved residues are found to be buried in DBPs, and more interestingly, they are surrounded by other conserved residues forming tight residue interaction networks.

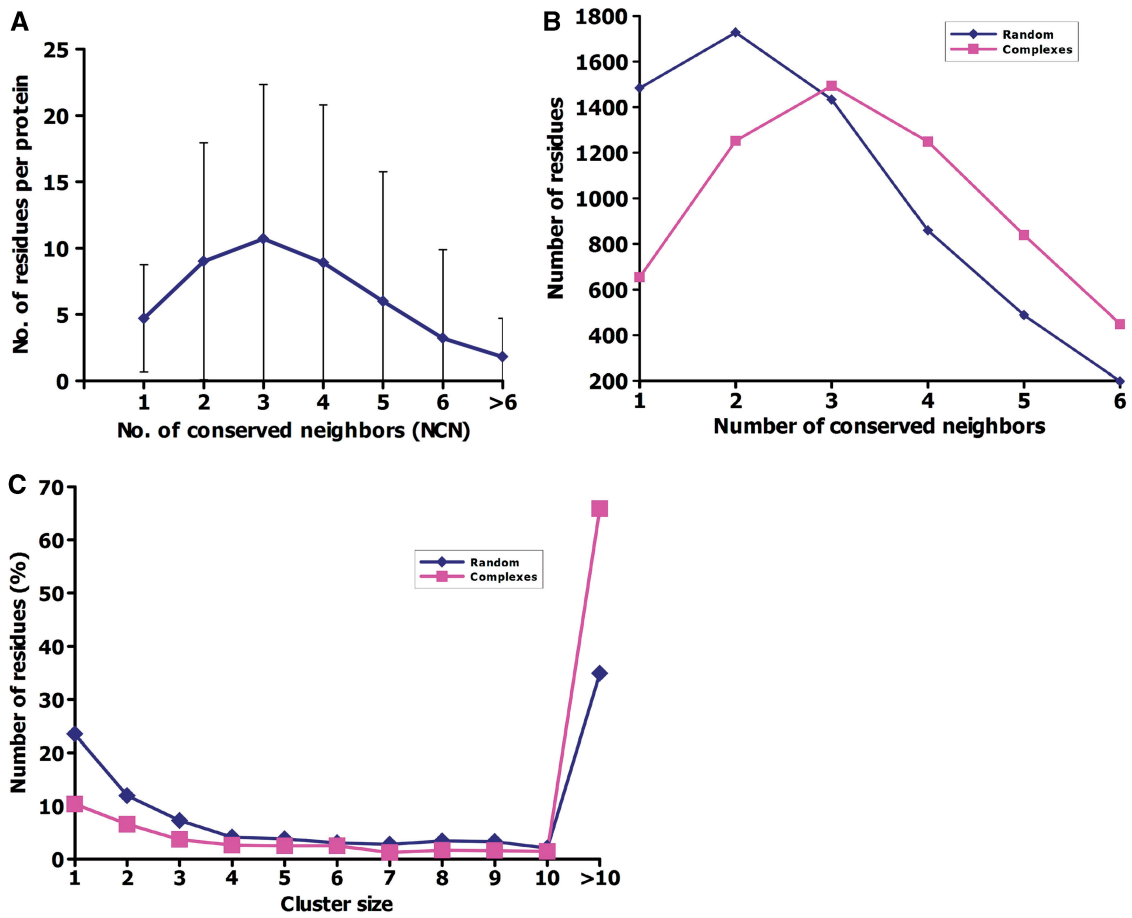

Figure 1. (**A**) The number of conserved residues as a function of the number of its conserved neighbors. As an example, the first point on the graph indicates that there are about five residues per protein in the data set, which have no other conserved residue in their structural neighborhood. The values are computed for each protein and the standard deviations are plotted in vertical error bars. The *x*-axis shows whether a conserved residue occurs as a singlet (first point) or a cluster with conserved neighbors. NCN is the number of conserved neighbors. (**B**) NCNs of a residue expected by chance. Random distribution was constructed by reassigning conservation scores randomly along the sequence. (**C**) Histogram of cluster size in unique clusters of conserved residues in observed protein–DNA complexes, compared to randomly distributed conservation scores.

Distribution in secondary structures. No statistically significant difference between helical, strand and coil distributions was observed.

Packing density. The packing density histogram of the conserved and clustered residues (with the number of conserved neighbors >1) in the protein–DNA interface is shown in Figure 2. A similar histogram is also obtained for noninterface residues (data not shown). Statistical analysis indicates that the packing density average for interface residues is 5.4 compared to 6.1 in conserved clusters. Conducting a *t*-test on packing density data for all residues, the *P*-value was found to be 1.6×10^{-7}, confirming the statistical significance of the difference. A higher packing density in the conserved clusters compared to the rest of the protein supports cooperativity between clustered residues.

Conservation score and $\Delta\Delta G$

Above, we analyzed the overall occurrence of conserved residues in DBPs. We now focus on the functional role of these residues at the DNA interface. We analyze the thermodynamic data of free-energy changes of single residue mutations in DBPs. We examine whether the most destabilizing mutations occur at positions with higher conservation scores and greater number of conserved neighbors; and vice versa i.e. whether the residues with high conservation scores and with more conserved neighbors contribute more to the stability of complexes. Table 1 gives the statistics of conservation scores and the number of conserved neighbors for the putative hotspot residues, the mean and the standard deviation of the free-energy change upon mutating the given position. These hotspots are the residues with the highest loss of stability upon mutations, ranked by the single mutation free-energy change data (*smddg* database, see Materials and methods section). We choose the 10–50 top ranking residues (2–10% of the *smddg* data). We observe a clear difference between the conservation and the number of conserved neighbors as compared to putative nonhotspots (e.g. for 10 top-ranked residues, average conservation for putative hotspot is 0.78; for putative nonhotspots 0.68; the number of conserved neighbors for the putative hotspots is 3.2; for nonhotspots 2.2). The difference between the scores diminishes as we use a more permissive definition for the putative hotspots. The statistical significance of these differences has been assessed using a *t*-test. For the 10 top positions, there are very few data and a statistical test may not be able to capture the significance and for larger *N*, the difference between the control and target becomes less obvious. Therefore the best *P*-values are obtained for 20 and 30 top-ranked mutations. Among the top-20 ranked mutations, the mutations studied are in wide range of proteins such as Endonuclease (PDB code 1azo), Gene activator protein (1run), Operon repressor (1lbg), EBNA nuclear protein (1b3t), ARC repressor operator (1par), Purine repressor (1bdh) and lambda CRO operator (4cro), indicating that the data consist of a fairly representative set of proteins.

We examine the reverse argument, namely is there a difference between the stability changes caused by *all* mutations (not just the most destabilizing mutations) in positions occupied by (i) nonconserved residues; (ii) all conserved residues; and (iii) conserved residues

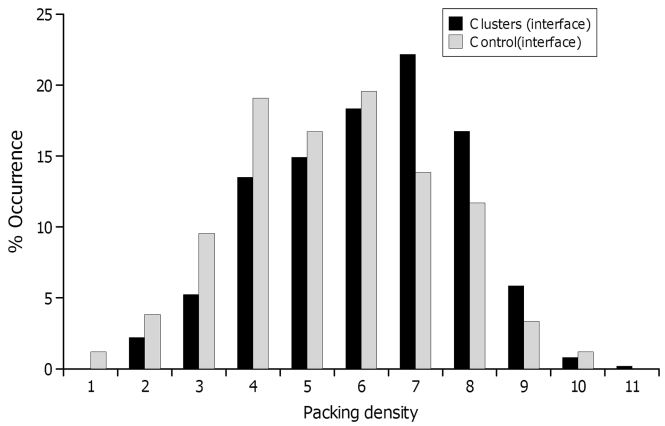

Figure 2. Interface residue packing density histogram. Conserved-clustered residues are more tightly packed than rest of the residues in the protein.

Table 1. The relationship between conservation scores (*C*) and the number of conserved neighbors (NCN) of the most destabilizing mutant positions in the $\Delta\Delta G$ data

N	Data	$<\Delta\Delta G>$	$\sigma(\Delta\Delta G)$	$<C>$	σ (*C*)	<NCN>	σ (NCN)	*P*-value (*C*)	*P*-value (NCN)
10	Target	4.837	1.000	0.779	0.154	3.200	2.044	0.154	0.138
	Control	−0.057	1.369	0.679	0.220	2.236	2.034		
20	Target	3.841	1.259	0.768	0.154	3.250	2.221	0.072	0.027
	Control	−0.132	1.273	0.678	0.221	2.222	2.032		
30	Target	3.311	1.274	0.769	0.146	2.867	2.113	0.022	0.086
	Control	−0.186	1.230	0.675	0.221	2.210	2.027		
40	Target	2.974	1.248	0.730	0.164	2.450	1.987	0.137	0.510
	Control	−0.234	1.197	0.676	0.223	2.229	2.040		
50	Target	2.735	1.213	0.741	0.160	2.540	2.002	0.044	0.304
	Control	−0.280	1.168	0.675	0.224	2.228	2.046		

Top *N* mutations in *smddg* data with highest values of $\Delta\Delta G$ (most destabilizing mutant positions) are used to form the target data and the rest is the control. *N* top-ranked mutations were selected for each pair of rows and *N* varied from 10 to 50 (about 2–10% in the *smddg* data set). $<X>$ stands for the mean value and $\sigma(X)$ for the standard deviation of quantity *X*. All $\Delta\Delta G$ values are in *kcal/mol* and conservation scores range from 0 to 1.

Table 2. Difference between average $\Delta\Delta G$ for mutations at nonconserved, conserved singlets and conserved-clustered positions and its statistical significance

Average $\Delta\Delta G$ (kcal/mol)	Nonconserved (NC)	−0.131 (345)
	All conserved (AC)	0.264 (167)
	Conserved singlets (CS)	−1.035 (18)
	Conserved clustered (CC)	0.421 (149)
P-value of difference between means	*P* (NC,AC)	0.00459
	P (NC,CS)	0.01009
	P (NC,CC)	0.00012
	P (CS,CC)	9.0E−05

Conservation score cutoff is 0.8, all conserved residues with at least one conserved neighbor are treated as conserved-clustered (CC), whereas conserved residues with no conserved neighbor are treated as conserved singlets (CS). Values in the brackets are the actual number of observations in the given category.

surrounded by other conserved residues (clustered-conserved regions). Supplementary Table S3 provides the details of the statistically significant differences in $\Delta\Delta G$ for mutations in conserved regions and clustered-conserved regions at different cutoffs. We observe that at conservation scores above 0.8, most *P*-values are <0.01 and the main results at this conservation score cutoff are presented in Table 2. A better estimate of the role for conserved residues and conserved-clustered regions could be made if we were to group mutations by residue identities and exclude the difference in the DNA sequence (nucleotide) to which they bind. However, available free-energy data are currently insufficient for such a study. Based on the overall statistics the average $\Delta\Delta G$ for mutations in conserved positions is 0.26 kcal/mol compared with −0.13 kcal/mol for nonconserved positions (conservation score cutoff 0.8). Within the overall set of conserved residue positions, those with at least one more conserved neighbor (clustered residues) have a much higher $\Delta\Delta G$ (0.42 kcal/mol) compared with the background set of all conserved residues. This suggests that cooperativity plays a greater role in the most stabilizing residues compared to the conservation, which exists and thus is likely to play a role even at relatively less stabilizing positions.

Cooperativity

The statistics of the distribution of conserved residues and their contributions to the stability of protein–DNA complexes as obtained from experimental binding data of mutations and the packing density suggest that the interaction between the protein and DNA is cooperative. The thermodynamic data indicate that conserved-clustered residues contribute more to stability as compared to conserved singlets; and that the packing density of clustered residues in the interface is significantly higher than other residues, similar to protein–protein interaction hotspots (1). A more direct indication of cooperativity would involve an analysis of a diverse set of experiments showing that the sum of the binding free-energy changes caused by multiple single mutations differs from the total free-energy change caused by *simultaneous* multiple mutations; that is, the cooperativity K between residues 1 and 2 is nonzero if $K(1,2) = \Delta\Delta G(1,2) - (\Delta\Delta G(1) + \Delta\Delta G(2)) \neq 0$.

For such an analysis, the experiments on the simultaneous mutations $\Delta\Delta G(1, 2)$ and individual mutation $\Delta\Delta G(1)$ and $\Delta\Delta G(2)$ must be performed under the same conditions (e.g. temperature, pH, buffer, ionic concentrations) and the DNA to which binding is studied should be identical. A search of the ProNIT database resulted in only 16 such pairs of mutations [data in Supplementary Table S3(b)]). Unfortunately, out of these 16, 12 did not occur in the same cluster. For the remaining four we observe that the cooperativity scores in two of them (1.03 and 0.90 kcal/mol) are almost five times that of an average in control data (−0.22 kcal/mol), whereas only one of them has a slightly smaller *K*-score.

We also look at thermodynamic data of more than one (individual) mutation in the same protein, which cause $\Delta\Delta G$ changes >2 kcal/mol, similar to hotspots in protein–protein interactions. Data relating to five proteins fell in this category (Table 3). The summary in the last row of this table shows that eight out of such 12 mutations are coclustered. Two of those not coclustered are borderline cases in the 0.8 conservation score cutoff. Overall, unfortunately current direct experimental data are limited; thus although it points toward cooperativity, the statistics is too small.

The observation of a role for conserved residues and clustered-conserved residues in stability, leads us to the analysis of conserved residues, their occurrence in the DNA interface and clustering patterns in protein–DNA complexes.

Conserved residues, conserved-clustered residues and DNA-binding residues

To determine if the conserved residues or conserved-clustered residues are enriched in the interface, we calculate the fraction of interface residues in each category (binding ratio; or the number of DNA-binding residues in a given category with respect to the total number of residues in that category). The statistics for each of the 20 amino acid types have been analyzed for nonconserved residues, conserved residue singlets and clustered-conserved residues (Supplementary Figure S3). The *P*-value results for all category pair combinations are shown in Table 4. The results suggest that (i) the ratio of interface residues for singlets is small; it is significantly higher for clustered-conserved regions (Met is an exception). This highlights the significance of clusters of conserved residues in DNA recognition; (ii) almost all Arg and ~80% of Lys among the conserved-clustered residues are in contact with DNA, considerably more than in nonconserved regions or singlets (~40% and ~20%); (iii) some residues, especially those with acidic side chains (Asp and Glu) only occur significantly in conserved clusters probably due to electrostatic repulsion with the DNA; (iv) for some hydrophobic residues there is almost no difference in binding ratio in the three regions. Ala, Cys, Leu, Met, Pro and Val are prominent amongst them. Phe and Trp are exceptions, as nearly 20% of these residues occur in the interface in conserved-clustered regions, compared to near absence as singlets or as nonconserved. Tyr is also enriched in clustered regions.

Table 3. Multiple experimental hot spots in the same protein

PDB Code	Mutations [$\Delta\Delta G$ kcal/mol]	CCR positions
1lmb	Q44S/Q44Y (av = 3.7); Q33S (4.5); A49V (4.6)	Q44; Q33;
1aay	R18A (2.7); R24A (3.5)	R18; R24
1b3t	Y518A (2.6); R522A (4.4); R469A (3.4)	Y518; R522
1run	D138A/D138V/D138L/D138T (av = 4.1); T127L (2.8)	D138 (T127 at C = 0.6 cutoff)
1mse	K128M (2.4); V103L (2.2)	K128 (V103 at C = 0.8 cutoff)
Summary		**8 of 12 at cutoff $C = 0.8$;** **9 of 12 at cutoff $C = 0.7$;** **10 of 12 at cutoff $C = 0.6$**

Some mutations have different mutant residue for the same position; the $\Delta\Delta G$ data have been averaged in such cases. CCR stands for clustered-conserved positions i.e. conserved residues occurring as part of a cluster.

Table 4. Statistical significance tests between the fractional numbers of DNA-binding residues of each type in the three defined regions

Residue	No. NCR (% binding)	No. CRS (% binding)	No. CCR (% binding)	*P*-value (NCR/CRS)	*P*-value (NCR/CCR)	*P*-value (CRS/CCR)
Ala	1771 (4.8)	36 (3.4)	385 (4.4)	0.641	0.810	0.780
Cys	285 (2.0)	6 (0.0)	114 (0.0)	0.477	0.070	–
Asp	1130 (5.4)	32 (0.0)	253 (15.8)	0.326	0.023	0.026
Glu	1777 (4.1)	33 (0.0)	318 (13.1)	0.060	0.001	0.018
Phe	810 (6.4)	34 (0.0)	283 (15.7)	0.095	0.072	0.120
Gly	1122 (15.7)	82 (5.8)	458 (25.8)	0.046	0.044	0.001
His	548 (20.0)	16 (8.3)	160 (35.5)	0.323	0.048	0.098
Ile	1293 (8.1)	14 (0.0)	267 (11.0)	0.232	0.469	0.293
Lys	1683 (27.1)	39 (17.6)	372 (74.9)	0.299	3.1E−06	0.031
Leu	2120 (2.2)	86 (0.0)	614 (3.5)	0.006	0.200	0.014
Met	475 (8.5)	4 (0.0)	64 (4.2)	0.669	0.508	0.706
Asn	908 (25.5)	10 (0.0)	187 (51.7)	0.271	0.017	0.165
Pro	940 (7.0)	65 (4.3)	282 (9.9)	0.397	0.402	0.246
Gln	1032 (16.8)	12 (0.0)	163 (28.3)	0.242	0.073	0.132
Arg	1484 (45.5)	81 (12.5)	490 (100)	0.002	5.8E−06	4.6E−05
Ser	1334 (19.2)	12 (0.0)	203 (48.1)	0.036	0.002	0.096
Thr	1098 (17.9)	11 (0.0)	268 (30.0)	0.128	0.070	0.238
Val	1413 (3.8)	31 (4.5)	317 (11.4)	0.841	0.016	0.330
Trp	298 (15.9)	18 (0.0)	97 (24.7)	0.092	0.262	0.043
Tyr	655 (16.3)	37 (0.0)	257 (32.8)	0.024	0.044	0.022

Nonconserved residues (NCR), conserved-residue singlets (CRS) and clustered conserved residues (CCR). *P*-values for (X/Y) are obtained using two-tailed students *t*-test on protein-wise distribution of *X* and *Y* in the data and indicate the probability that the two types of regions are similar. Some larger *P*-values, showing low statistical confidence are due to a small number of binding residues of that type in one or both regions compared. *N* refers to the total number of residues in a given category. Overall, there are 659 singlets and 5552 (about 8.4×) clustered-conserved residues and for the overall data all three pairs have statistically significant difference.

The occurrence of Phe, Tyr and Trp in protein–protein interaction hotspots has been described earlier (1) and these results similarly show that these residues also prefer to be conserved in clusters in the DNA-binding interface. Their aromatic nature should play critical roles in DNA binding.

Solvent-accessible residues and binding

We analyze the occurrence of conserved-clustered residues in the interface (all DNA-contacting residues) with respect to their solvent accessibilities. Data are too sparse for residue-wise comparison. Figure 3 shows the distribution of residues (singlets, nonconserved and conserved-clustered) in different accessibility ranges. As noted above (Data sets section) the total number of residues falls sharply with the ASA (complete statistics in Supplementary Figure S2). However, interface residues become more abundant within the clustered and conserved categories at higher accessibility. Figure 3 shows the averages of raw values computed for each protein. Additional details, including *P*-values of *t*-tests are given in Supplementary Table S4. The results show that the binding ratio (number of DNA-binding to overall residues within a category) of conserved-clustered regions is clearly higher than that of the nonconserved ones or singlets and the difference is even higher in more exposed regions. This observation leads us to conclude that (i) a highly exposed conserved-clustered residue is likely to be in the interface (nearly 90% of such residues with ASA > 20% are in the interface); and (ii) for singlets and nonconserved residues only a small number (about 35% and 15% of the most exposed, respectively) are in the interface, whereas most others are not in contact with the DNA despite being on the surface.

Properties and distribution of clustered-conserved regions

We analyze the intra-cluster organization since Figure 1 and Supplementary Figure S1 do not provide the details of

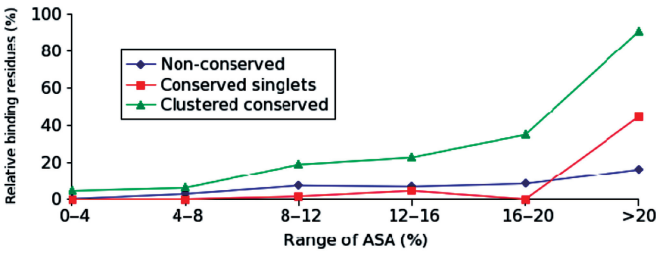

Figure 3. The relative frequency of DNA-binding residues in three identified regions: nonconserved residues; conserved residue singlets with conservation score at least 0.8 and no conserved neighbors; and clustered-conserved regions with conservation score at least 0.8 and at least one conserved neighbor, in different ranges of ASA.

the interaction and network of the conserved residues. We aim to understand how the conserved clusters are found in DBPs. We thus proceed to identify unique sets of clustered-conserved regions (see Materials and methods section), and carry out a systematic analysis of features of these sets at different cutoffs of clustering distance and conservation scores. While the number of clusters and their sizes vary, the overall trend remains unchanged within the tested parameter range (data not shown). Since the statistically most significant values were obtained with 0.8 conservation score cutoff and 6.0 Å clustering distance, these values were retained for further analysis.

Supplementary Table S5 gives a complete list of unique clusters in each structure class and includes the information about DNA contacts. Here, we discuss salient features. Overall two groups stand out: enzymes have the highest number of clusters (~5.1 conserved clusters per protein with 267 clusters in 52 proteins) and the Histone-like group, with the fewest (~1.6 per protein i.e. 13 clusters in eight proteins). In other groups, the ratio is more or less consistent (~3 clusters per protein). Specifically, the HTH group has 91 clusters in 31 proteins, the zinc-coordinating group has 40 in 12, the Zippers group has 12 in six, the β-sheet group has five in two, and the β-hairpins group has 28 in seven. There are only 11 proteins in which no clustered-conserved regions are observed probably due to the high conservation score cutoff that we used (the multiple alignment contained relatively distant sequences with low conservation scores). There are 129 proteins with at least one clustered region. Among these, only six proteins (PDB codes: 1oe5A, 1oupA, 1r7mB, 1rxwA, 1zetA and 1jeyA; entry numbers 84, 86, 90, 91, 92 and 140 in Supplementary Table S5) do not have any clusters in contact with the DNA (but they do have conserved clusters elsewhere in the structure). Interestingly, all but one of these proteins were classified as DNA-binding enzymes. Enzymes with the highest number of clusters per protein are characterized by a large number of small unique clustered regions (sometimes in addition to a large interface cluster), many of which are not on the interface. Figure 4a shows a typical example of a DNA-binding enzyme (PDB code 1qaiB, reverse transcriptase). This protein has five unique clusters located in different regions of the protein, only one of which is in the interface. The interface cluster of this protein like many other cases is also the largest. Another enzyme, the TAQ MUTS protein (PDB code 1ewqA) has several unique clusters, mostly small but two large clusters are centered on Leu260 (81 residues) and Gln19 (16 residues), as shown in Figure 4b. One is in the DNA interface and the other in contact with its dimerization partner forming a protein–protein interface. Probably, other smaller clusters stabilize the enzyme scaffold.

As noted above, DBPs belonging to histone-like structures (Group V) proteins (see Supplementary Data) have the fewest and largest clusters. If we take the largest cluster from each protein of this group, the smallest has 30 residues, which is much larger than the typical cluster sizes in the entire data set. The largest of all histone clusters is composed of 113 residues (in the nucleosome core particle, PDB code: 1eqz, chain G; total protein chain length: 136), which is rather exceptional amongst the clusters being analyzed. This example is shown in Figure 4c. As can be seen, almost the entire chain of a histone molecule is conserved, apparently a requirement to bind DNA and other structural proteins both of which make numerous contacts with other histone protein chains in the complex.

Although, not as striking as the above two DBP classes some interesting features are also observed in other groups. For example, in the widely studied HTH group of proteins, we observed 91 clusters in 31 proteins; 52 of these are on the interface. The remaining 39 clusters are typically much smaller in size and have no contacts with DNA. In some cases (e.g. Paired box protein, PDB code 1k78I), there is just one cluster occurring on the recognition helix (in the interface). The occurrence of small clusters outside the interface is observed in many proteins. One such example, is the phosphate region transcription regulatory protein (PDB code: 1gxpB) shown in Figure 4d. As observed in this example and in many other HTH proteins there is a relatively larger cluster in the recognition domain and one or more smaller clusters occur in the stabilizing helix and sometimes in the linker regions. Two clusters on the same helical segment are a rare occurrence, suggesting that the recognition is more localized in these proteins.

In the zinc-coordinating group of proteins (Group II, data entry 32–43), the most interesting observation is the occurrence of small unique clustered regions consisting of two or more Cys residues, sometimes accompanied by one or two Arg residues, in almost all members of this family. In a few cases, this zinc-binding cluster is larger. These clusters are typically not in the interface with DNA, and there is often another (larger) cluster in the interface. There are additional small clusters, which have essentially the same properties as that in HTH proteins and are not strikingly peculiar to the zinc-coordinating group of proteins. A typical example of a Tandem zinc finger (Zif 268; PDB code 1p47A) is shown in Figure 4e. This protein has five conserved clusters, three of which (shown in blue) consist of a pair of Cys residues (sometimes accompanied by a few more residues) coordinated to a zinc ion (shown in red). A large cluster (shown in green) runs through the DNA interface of the complex and links to one of the zinc ions via a His residue. This arrangement of two Cys and

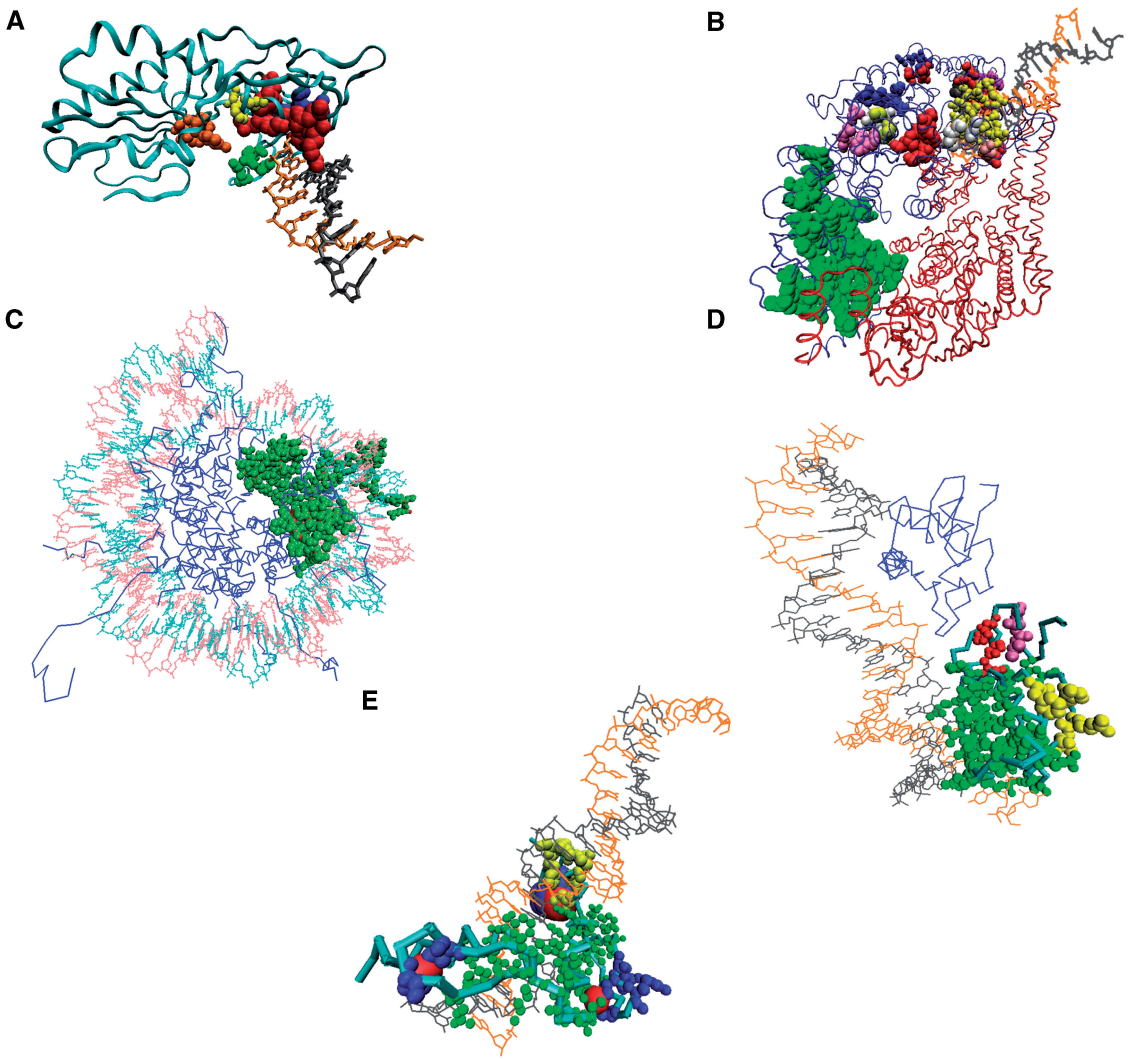

Figure 4. Clustering patterns of conserved residues (**A**) a typical enzyme (PDB code 1qai, chain B, reverse transcriptase). Several small clusters of conserved residues are observed in most enzymes. (**B**) Another DNA-binding enzyme TAQ MUTS protein (PDB code 1ewqA). One large cluster of conserved residues is observed in the oligomerization domain forming a protein–protein interface. Several other small clusters occur in recognition domain and scaffold. (**C**) Nucleosome core particle (PDB code 1eqzG) protein is a typical example of histone-like proteins with highly conserved residues throughout their structure. Usually a single large cluster is observed as most residues are conserved. (**D**) Phosphate region transcription regulatory protein (PDB code: 1gxp chain B) is a typical HTH protein with a few small clusters, usually one in the recognition helix, one in linker region and the other in the stabilizing helix. (**E**) A typical zinc-coordinating protein Tandem zinc finger (Zif 268; PDB code 1p47A). Small clusters of two Cys residues (shown in blue)—sometimes accompanied by other residue—form small clusters of conserved residues away from DNA interface and coordinate zinc ions (shown in red). A large cluster is observed in the interface in contact with DNA major groove. Sometimes, this cluster extends to include conserved His residues from the C2H2 motif.

their corresponding His residues (part of a C2H2 motif, typical of zinc coordination) lying on the other DNA-interface cluster, is quite common in this group of proteins, although in two cases (PDB codes 2a66A and 1ga5E), the C2 and H2 regions lie on the same clusters, as the residues on the opposing sides of zinc ion come in contact through their side chains merging the two clusters.

In the unique clustered regions of Zipper Group III and other alpha helical (Group IV) DBPs, one (typically the largest) or more clusters are in the interface, whereas a few small clusters occur in other regions, presumably to provide structural stability.

There are only two β-sheet (Group IV) DBPs (PDB codes 1qn7B and 1rm1C, data entry numbers 57 and 58) each with two interface clusters. Although, these proteins interact with the DNA minor groove, no special features could be extracted due to a small number of members in that family. Similarly, the β-hairpin proteins group members also do not show any special features.

Prediction of interface clusters

About 50% of the unique conserved-clustered regions are in the interface with the DNA, whereas most others are either in the protein–protein interface or have other structural roles, such as zinc-coordination or fold stabilization. It is useful to identify DNA-interface clusters from their simple properties as it could assist in predicting the most

Table 5. The number of interface and noninterface clusters falling in the specified hydrophobic, hydrophilic, negatively charged and positively charged residues composition ranges

Number of clusters (%)								
Composition (%)	Hydrophobic		Hydrophilic		Negatively charged		Positively charged	
	Bind	NB	Bind	NB	Bind	NB	Bind	NB
0–10	19.7	12.3	17.7	52.5	67.9	64.9	28.9	71.7
10–20	3.6	0.7	14.1	5.1	20.9	8.0	28.9	6.1
20–30	6.4	4.3	23.7	8.3	4.0	5.1	13.3	4.7
30–40	15.7	7.6	17.3	8.7	4.4	5.4	11.2	6.5
40–50	26.5	32.6	16.5	22.5	2.8	14.1	13.7	10.1
>50	28.1	42.4	10.8	2.9	0.0	2.5	4.0	0.7
P-value	1.17E−13		2.5E−08		5.96E−04		6.01E−13	

Clusters in the DNA-interface (bind) and with no DNA contact (NB) significantly differ in their compositions. For example 71.7% of noninterface clusters have <10% positively charged residues, whereas just 28.9% DNA-interface clusters have such low composition of positively charged residues. This difference in composition leads to statistically significant difference between DNA-interface and noninterface residues (see *P*-values), which could be used for prediction. Linear predictor using just four parameters of a cluster can identify DNA-interface clusters with high confidence. Residue classification: hydrophobic (Ala, Cys, Phe, Ile, Leu, Met, Pro, Val, Trp, Tyr), hydrophilic (Gly, His, Asn, Gln, Ser, Thr), negatively charged (Asp, Glu), positively charged (Lys, Arg). Mean prediction scores on 10-fold validation, Sensitivity (true positive/actual positive): 87.3%, Specificity (true negative/actual negative): 67.3%.

significant DNA-binding sites. Since DNA contacting and other clusters have different functions, we compared their hydrophobic, hydrophilic and charge type composition, calculated the statistical significance of the differences, trained a linear predictor on the data set and tested their predictability using 10-fold cross-validation (see Materials and methods section). Table 5 summarizes the results. As expected, DNA-interface clusters have a higher content of positive charges. Hydrophobic and negatively charged residues in interface clusters are significantly lower, occurring more frequently in protein–protein or in intrachain contacts. Based on the four composition parameters, we predicted 86% of the interface clusters from their composition with 68% specificity. As expected, a cluster with positively charged residues and with a higher value of χ-score [see Equation (2)] is more likely to be in the interface. Only 22% of the clusters are misclassified by a simple count of their four types of residues.

Comparison with protein–protein interfaces

We compared the hotspot cluster organization in protein–protein versus protein–DNA interfaces. Previously, we found that computational hotspots are not homogeneously distributed in protein–protein interfaces, but form tightly packed cooperative clusters. Here, we similarly observe that conserved residues are tightly packed in clusters of varying sizes. Interestingly, despite the difference in the overall character of the interfaces, when the types of amino acids in the clusters of both interface types are compared, we notice that some of the same type of residues namely, Trp, Phe and Tyr are abundant in both sets. However, the high occurrence of positively charged residues Arg and Lys in conserved clusters of DBPs distinguishes them from protein–protein interface clusters.

CONCLUSIONS

Here, we investigate a potential relationship among free energy, sequence conservation and structural cooperativity of conserved residues in protein–DNA recognition. We analyzed a dataset of 3D structures of protein–DNA complexes using experimental thermodynamic data of mutations, and identified putative residue hotspots in their interfaces. According to our definition, in analogy to protein–protein interactions, putative hotspots contribute over 2 kcal/mol to the binding free energy. Our results show that the most stabilizing residues tend to occur in distinct clusters. About half of the clusters of conserved residues are in contact with DNA; the others are at the interfaces with proteins or between elements of the protein structures. Based on the properties of these clusters, we developed a classifier and were able to predict with high confidence clusters interacting with DNA. Our comprehensive analysis of the hotspots, conservation and their structural environments suggest that similar to protein cores and protein–protein interfaces, cooperativity plays an important role in protein–DNA interactions, while residue conservation can take place also at relatively less stabilizing positions. In particular, the most destabilizing mutations (top ~5%) appear to be more conserved than nonhotspot residues and these residues occur more often in clusters of conserved residues.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. This research was also funded by Grants-in-Aid for Scientific Research 16014219 and 16041235 to A.S. from

the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES:

1. Keskin,O., Ma,B. and Nussinov,R. (2005) Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.
2. Guharoy,M. and Chakrabarti,P. (2005) Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.
3. Burgoyne,N.J. and Jackson,R.M. (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, **22**, 1335–1342.
4. Kortemme,T. and Baker,D.A. (2002) Simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
5. Ma,B., Elkayam,T., Wolfson,H. and Nussinov,R. (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.
6. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
7. Li,X., Keskin,O., Ma,B., Nussinov,R. and Liang,J. (2004) Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J. Mol. Biol.*, **344**, 781–795.
8. George,R.A., Spriggs,R.V., Bartlett,G.J., Gutteridge,A., Macartur,M.W., Porter,C.T., Al-Lazikani,B., Thornton,J.M. and Swindelss,M.B. (2005) Effective function annotation through catalytic residue conservation. *Proc. Natl Acad. Sci. USA*, **102**, 12299–12304.
9. Nordlund,A. and Oliveberg,M. (2006) Folding of Cu/Zn superoxide dismutase suggests structural hotspots for gain of neurotoxic function in ALS: parallels to precursors in amyloid disease. *Proc. Natl Acad. Sci. USA*, **103**, 10218–10223.
10. Kinney,J.B., Tkacik,G. and Callan,C.G. Jr. (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc. Natl Acad. Sci. USA*, **104**, 501–506.
11. Veprintsev,D.B. and Fersht,A.R. (2008) Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res.*, **36**, 1589–1598.
12. Mahony,S., Auron,P.E. and Benos,P.V. (2007) Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, 297–304.
13. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
14. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
15. Ahmad,S., Kono,H., Arauzo-Bravo,M.J. and Sarai,A. (2006) ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res.*, **34**, W124–W127.
16. Ahmad,S. and Sarai,A. (2004) Moment-based prediction of DNA-binding proteins. *J. Mol. Biol.*, **341**, 65–71.
17. Zhang,Y., Xi,Z., Hegde,R.S., Shakked,Z. and Crothers,D.M. (2004) Predicting indirect readout effects in protein-DNA interactions. *Proc. Natl Acad. Sci. USA*, **101**, 8337–8341.
18. Hallikas,O., Palin,K., Sinjushina,N., Rautiainen,R., Partanen,J., Ukkonen,E. and Taipale,J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.
19. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence, *Bioinformatics*, **23**, 347–353.
20. 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.
21. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
22. Faiger,H., Ivanchenko,M. and Haran,T.E. (2007) Nearest-neighbor non-additivity versus long-range non-additivity in TATA-box structure and its implications for TBP-binding mechanism. *Nucleic Acids Res.*, **35**, 4409–4419.
23. O'Flanagan,R.A., Paillard,G., Lavery,R. and Sengupta,A.M. (2005) Non-additivity in protein-DNA binding. *Bioinformatics*, **21**, 2254–2263.
24. Mirny,L.A. and Gelfand,M.S. (2002) Structural analysis of conserved base-pairs in protein-DNA complexes, *Nucleic Acids Res.*, **30**, 1704–1711.
25. Liu,J. and Stormo,G.D. (2005) Quantitative analysis of EGR proteins binding to DNA: assessing additivity in both the binding site and the protein. *BMC Bioinformatics*, **6**, 176.
26. Sathyapriya,R. and Vishveshwara,S. (2004) Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.*, **32**, 4109–4118.
27. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
28. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
30. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, 1.
31. Larkin,M.,A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) ClustalW2 and ClustalX version 2. *Bioinformatics*, **23**, 2947–2948.
32. Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1145.
33. Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
34. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bond and geometrical features. *Biopolymers*, **22**, 2577–2637.
35. Wang,J.Y., Lee,H.M. and Ahmad,S. (2007) SVM-Cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. *Proteins*, **68**, 82–91.
36. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
37. Andrabi,M., Ahmad,S., Mizuguchi,K., and Sarai,A. (2008) Benchmarking and analysis of DNA-binding site prediction using machine learning. In Proceedings of International Joint Conference on Neural Networks (IJCNN), World Conference on Computational Intelligence (WCCI), June 1–8, Hong Kong, IEEE. pp. 1746–1750