Conformational changes in DNA-binding proteins: Relationships with precomplex features and contributions to specificity and stability

Munazah Andrabi,1,2 Kenji Mizuguchi,1 and Shandar Ahmad1*

1 Bioinformatics project, National Institute of Biomedical Innovation, 7-6-8 Saito-Asagi, Ibaraki City, Osaka 567-0085, Japan
2 Genome Resource & Analysis Unit (GRAS), Center for Developmental Biology (CDB), RIKEN, Kobe 650-0047, Japan

ABSTRACT

Both Proteins and DNA undergo conformational changes in order to form functional complexes and also to facilitate interactions with other molecules. These changes have direct implications for the stability and specificity of the complex, as well as the cooperativity of interactions between multiple entities. In this work, we have extensively analyzed conformational changes in DNA-binding proteins by superimposing DNA-bound and unbound pairs of protein structures in a curated database of 90 proteins. We manually examined each of these pairs, unified the authors’ annotations, and summarized our observations by classifying conformational changes into six structural categories. We explored a relationship between conformational changes and functional classes, binding motifs, target specificity, biophysical features of unbound proteins, and stability of the complex. In addition, we have also investigated the degree to which the intrinsic flexibility can explain conformational changes in a subset of 52 proteins with high quality coordinate data. Our results indicate that conformational changes in DNA-binding proteins contribute significantly to both the stability of the complex and the specificity of targets recognized by them. We also conclude that most conformational changes occur in proteins interacting with specific DNA targets, even though unbound protein structures may have sufficient information to interact with DNA in a nonspecific manner.

INTRODUCTION

Almost all biological pathways depend on the expression and regulation of proteins and their interaction with DNA is a major biochemical event controlling these processes (e.g., Refs. 1 and 2). Protein-DNA interactions are often accompanied by a change in the conformation of one or both the partners to carry out the required function. Although conformational changes can also occur in response to environmental perturbations such as site-specific chemical modifications (e.g., phosphorylation and methylation) or varying pH and temperature levels, the events of binding interactions resulting in the formation of complexes between protein–protein, protein–ligand, and protein–DNA are of particular interest. Protein conformational changes in systems such as protein–protein and protein–ligand complexes have been the subject of extensive investigation owing to their implications in docking and drug design.3–5 Some of these studies are aimed at understanding signaling and energy propagation through conformational changes,6,7 whereas others have investigated molecular recognition mechanisms from this perspective.8 Several models describing recognition mechanisms underlying conformational changes have been postulated,9,10 of which “conformational selection” has been described as the main driver of conformational changes.3–5 Using the conformational selection model,
intrinsic flexibility, and dynamics of the unbound protein, some studies have also attempted to model/predict the conformational change expected on complex formation.\(^5\)

While these studies have greatly enhanced our understanding of the conformational changes accompanying protein–protein/protein–ligand interactions, it is not immediately clear whether the observations made from these systems can be extrapolated to the protein–DNA system. Protein–DNA interactions are unique in terms of their target specificity, the dominant electrostatic nature of the interface and the ability of proteins to find specific sites on a much larger DNA molecule. To gain insights into the nature of protein-DNA interactions in their entirety, it therefore becomes essential to understand the conformational changes in the interacting entities viz. proteins and DNA brought about by complex formation. DNA conformational changes are well documented and widely discussed,\(^11-17\) as the only requirement for such an analysis is the availability of protein-bound DNA-structure; the unbound structure can be assumed to be similar to a standard canonical form such as B-DNA to a fair degree of approximation. However, the protein side of the story is far more complex owing to their structural diversity, necessitating each individual unbound structure to be solved explicitly. Unfortunately, only a limited number of DNA-binding proteins have been crystallized in both the unbound and DNA-bound forms, imposing a limitation in describing and understanding the basic principles underlying the recognition and subsequent interaction by them, even when the final complex is already known. Availability of fewer pairs of structures has resulted in equally small number of analyses of conformational changes in DNA-binding proteins. Among the few relevant studies on the subject is one carried out on a small data set of 24 proteins, including 8 disordered structures, as part of an overall analysis of structural features of protein-nucleic acid complexes.\(^18\) Since this study focused on general structural features of protein-DNA complexes, it did not provide a detailed understanding of the conformational changes. More recently, a protein-DNA docking benchmark was reported, in which the authors compiled a dataset of 47 free/bound structure pairs of DNA-binding proteins and mainly analyzed their conformational changes with the sole purpose of establishing basic standards for modeling protein-DNA complexes through docking.\(^19\) Hence this study focused on docking-related issues without carrying out a detailed analysis of conformational changes and their role in biological function. Apart from these, conformational changes in DNA-binding proteins have often been analyzed in greater details in the original papers reporting the three-dimensional structure of complexes or in reviews of the recognition mechanisms of a specific family of proteins such as endonucleases and polymerases.\(^20-22\)

In this work, we address three issues of conformational changes in DNA-binding proteins: (a) types of conformational changes, their distribution among proteins of different functions and their relationships with physico-chemical features of proteins such as charges and dipole moments, (b) the role of intrinsic flexibility of the unbound protein in inducing conformational changes and (c) contributions of conformational changes to the stability and specificity of protein-DNA recognition. For this purpose, we created a dataset of unbound and DNA-bound pairs of protein structures, manually examined each of them by superimposing their bound and unbound variants and studied the nature of conformational changes observed in them. We classified the conformational changes into six types and related them to various functional classes of DNA-binding proteins. Subsequently, we performed a normal mode analysis of unbound protein structures using an elastic network model and examined the agreement between the predicted and observed conformational changes. Finally, we investigated a relationship between the extent of conformational changes in DNA-binding proteins and the stability and specificity of protein-DNA complexes, for which such information was available from published experimental results. Our results provide a broad picture of conformational changes in DNA-binding proteins and hence contribute to our current understanding of protein-DNA interactions.

**METHODS**

**Data sets**

**DNA-bound proteins or protein–DNA complexes**

Protein-DNA complexes solved by X-ray crystallography and having a resolution better than 2.5 Å were collected from a local mirror of the Protein Data bank (PDB) on November 9, 2011.\(^23\) These comprised of 1142 complexes with 2121 protein and 2863 DNA chains. Data redundancy was then removed at a single protein chain level by clustering them using the **BLASTCLUST** module of the NCBI BLAST package\(^24\) at a 25% sequence identity threshold, resulting in 317 clusters. From each of these clusters, a protein chain with the highest number of contacts (number of residues within 3.5 Å from any DNA atom) was selected as a representative. Protein chains making no contact with DNA at the given distance cutoff were discarded, reducing the number or representative protein chains to 310. Proteins in complex with single-stranded DNA were removed, leaving 274 protein chains for further processing.

**Selection of unbound partner**

A data set of 71,418 protein structures (172,790 chains) solved without a nucleic acid (DNA or RNA) in their structure was compiled from the PDB in November 2011. FASTA- formatted sequence files for all the proteins
were downloaded. Protein chains with less than 40 amino acids and having unknown residues were discarded, reducing the number to 160,906 chains. Protein chains from the complex data (274 protein chains) were then aligned to this data set using the BLASTALL module of the NCBI Blast package. In addition to the sequence identity obtained from BLASTALL, two additional parameters, Query Coverage (QC) and Subject Coverage (SC), were defined to select only the most reliable aligned pairs. QC and SC are calculated as the percentage of aligned length of the protein and the total sequence length where Query is the DNA bound protein chain and Subject is the unbound protein chain. Unbound protein chains having a sequence identity, QC and SC > 90% were selected as the partners of the DNA-bound protein chains.

**Manual curation of bound-unbound pairs**

**DNA structure.** The bound/unbound pairs of DNA-binding proteins were manually examined using PyMOL and tools and annotations in the NDB to find any anomalies in the DNA structure. Based on the Structural Features of DNA defined in the NDB, DNA structures were further assessed and only the B-DNA structures classified as Double Helix were included and complexes with their DNA chains classified as Single Stranded, Z-DNA, or A-DNA or with ambiguous structure were eliminated.

**Protein structure.** Visual inspection was also carried out to ensure that both the unbound protein and its corresponding complex with DNA are in identical protein oligomeric states. Information about the oligomeric states was obtained from the PDB and examined using PyMOL. Protein structures with changing oligomeric states were either excluded or replaced by an alternative member with a similar oligomeric state from the original cluster of proteins (BLASTCLUST results), leaving behind a clean set of 90 bound-unbound pairs.

**Recovering additional hits and final selection.** Another round of data selection was carried out to examine the original BLASTCLUST-generated clusters, which were lost due to any of the above filters on the selected representative. In this round, alternative representatives from these clusters were systematically selected and examined through the same filters until a suitable member could be selected or no member of the cluster satisfied the quality conditions imposed above. This operation gave us five additional proteins and produced a final data set of 90 proteins. This data set is referred to as DBP90 in this manuscript. Several proteins in DBP90 do not have complete coordinate data in either the complex or unbound forms, some of which is due to the presence of intrinsically disordered positions. Excluding all such proteins, we are left with 52 proteins with complete structures and no missing residues. This data set is referred to as DBP52 from here on.

**Quantifying conformational change**

In this work, we use the phrase conformational change to denote differences in DNA-bound and unbound structures in identical protein oligomeric states. Evaluating whether such differences represent true physical movements induced by DNA binding or simply random variations will require additional tools and data sets and goes beyond the scope of the current study. Here we assume that the differences between unbound and bound structures, especially with a large magnitude, represent real conformational changes. The degree of conformational changes is estimated by comparing bound and unbound structures using two independent parameters as follows.

**Superimposition and RMSD**

Superimposition of the bound and unbound proteins was performed and the root mean squared deviation (RMSD) between $C_a$ atoms in the two states was calculated using the least squares fitting method as implemented in the ProFit program. For visual inspection, “align” and “cealign” tools of PyMOL were also utilized.

**Global distance test (GDT_TS) and GDT_TS5**

Although the RMSD can accurately quantify small scale changes between two closely related conformations of a protein, it has been reported to exaggerate the estimates of global changes in structures. Several alternative metrics for quantifying structural changes have been developed to address this issue. We chose the Global Distance Test measure introduced by Zemla as part of LGA. The GDT score has shown to be a reliable measure to evaluate the local and global structural similarity as it iteratively collects the largest set of pairs of residues that can be aligned under a given distance cutoff. In this work a whole-protein score GDT_TS is defined as the relative number of residues ($C_a$ atoms) from the complex, which can be aligned with the corresponding residue of the reference (unbound) protein within a 5 Å distance. Since GDT_TS is a more general term, we used the notation GDT_TS5 to indicate that we have used a 5 Å cutoff in our calculation of GDT_TS.

**Classification of conformational changes**

Several authors have characterized conformational changes while reporting individual structures of protein-DNA complexes. However, no standard classification of these changes has been established. In order to develop such a scheme, we compiled the characterization of conformational changes by the authors from the primary literature and classified the types of changes using unified terms. In cases where the complex and unbound structures have been solved independently by different authors and/or conformational changes have not been adequately characterized, we examined the structures using multiple
measures and assigned them to an existing type or defined a new type to create a comprehensive list of conformational changes.

**Secondary structure and order/disorder calculation**

Secondary structure of both the unbound and bound protein chains was calculated using the DSSP program. A residue position is called disordered if no corresponding data was found in the ATOM records of its PDB coordinate file. A protein is annotated as containing a disordered region if at least 5 residues fulfill this condition.

**Calculation of DNA-contacts**

For each protein chain in our data set, we computed the number of contacts with DNA atoms, grouped as the major and minor groove atoms assuming the B-DNA conformation as in our previous work. Similarly DNA-atoms were also grouped as the base or the backbone atoms and corresponding contacting residues in each category were identified. Finally, the number of DNA-contacts of any type was defined as the unique number of residues forming one or more contacts with the corresponding group of DNA atoms.

**Residue-wise propensity and packing density**

The propensity of a residue type (e.g., Arg) to undergo conformational changes was measured by the average displacement (after optimal superimposition) of all occurrences of this residue in the data set relative to the average displacement of all residues of any type. These values were computed both for the entire protein and the DNA-interface residues only. Packing density was defined as the number of residues within a distance of 7 Å from the Cα atom of each reference residue. The backbone phosphate was used instead of Cα for DNA residues. This definition of packing density is consistent with our previous work on the subject.

**Temperature factors**

PDB coordinate files often provide atom-wise temperature factors. Since we used Cα positions for all the other analyses, the temperature factors were also collected for Cα positions and treated as whole residue B-factor values.

**Normal mode analysis**

Normal modes have been used to study the collective motions in biological macromolecules. In the current study, normal modes were calculated using the available software PDBMAT and DIAGRTB, downloaded from (http://ecole.modelisation.free.fr/modes.html). The Cα only elastic model network was implemented with a cutoff distance of 10 Å. Using these tools the conformational features were computed as follows.

**NMA predicted collectivity**

Collectivity (K) describes the collective motion of protein atoms or the number of highly mobile atoms corresponding to each low frequency mode (j). The NMA-Predicted value of this parameter was calculated as follows.

\[
K = \frac{1}{N} \exp \left( -\sum_{i=1}^{N} a_{ij}^2 \log a_{ij}^2 \right)
\]

where \(a_{ij}\) is the displacement of atom \(i\) under mode \(j\) and \(N\) is the total number of residues.

**Observed collectivity**

This was measured by the formula as above, where the displacement \(a_{ij}\) of each Cα atom was obtained by superimposing the bound and unbound structures using the least square fitting method.

Correlation between the NMA predicted atomic displacements and those observed upon the superimposition of unbound and bound structures was calculated to estimate the agreement between the predicted and observed local conformational changes.

**Best matching mode**

The best matching mode was selected from the nontrivial lowest frequency modes (normal modes 7 to 18), in such a way that the agreement with the experimental data for that mode was better than the other normal modes. To find the best-matching normal mode, observed displacements between complex and free proteins were compared with those predicted by each mode from the free protein alone. The normal mode with the highest correlation coefficient was selected as the “best match” normal mode.

**Sequence specificity and stability of DNA structure**

All 90 protein structures were manually annotated to be binding to their targets specifically or nonspecifically using information collected from the literature. Specificity information was typically inferred from the description of recognition targets in the original publication reporting the three-dimensional structure. Structural DNA-binding proteins such as histones and chromatin binding proteins were treated as nonspecific for this analysis. Stability of protein-DNA complexes, as measured by the free energy change on complex formation, was obtained from the thermodynamic databases of protein-nucleic acid complexes (PRONIT).
**Charge and electric moment**

Total charges, dipole moments and quadrupole moments were computed as in our earlier studies, i.e., by assigning a +1 charge to \( C_\alpha \) positions of Arg and Lys and -1 to those of Asp and Glu. All other residues were treated as neutral. This method of computing electric moments has been found to be successful in characterizing DNA-binding proteins and is robust against missing atoms in the coordinate data of the protein and uncertain ionization states of His, as reported in our earlier papers.44,45

**RESULTS**

**Types of conformational changes**

We manually examined the superimposed structures of 90 protein-DNA complexes in our data set as listed in Table I and classified them into seven groups (one group of negligible conformational change and six different types of conformational changes). (A protein was allowed to be in multiple groups, if it showed different types of changes and the redundant lists of groups is shown in Table II).

The first group consisted of 8 “highly rigid” proteins, in which no conformational change (displacement of \( C_\alpha \) position or secondary structure change) was observed and the RMSD was less than 1.0 Å (GDT_TS5 was more than 95% in all these cases). The other six groups are: (1) local loop motion (LL) (69 proteins), (2) secondary structure change (SS) (68 proteins with at least 10 residues undergoing change), (3) order/disorder transition (OD) (28 proteins, with at least 5 residues undergoing transition), (4) single arm movement (SA) (10 proteins), (5) open/close conformational switching (OC) (7 proteins) and (6) inter-domain or quaternary structure change (ID) (10 proteins). A typical example of each proposed group and the number of proteins in each group are illustrated in Figure 1 and Table II respectively. Previously, Nadassy et al.18 had characterized conformational changes into four basic categories without analyzing detailed features of individual groups. With a larger data set that spans greater diversity of conformational changes, we believe the proposed classification will be more informative. A detailed analysis based on this classification also leads to novel findings as described below.

We found that except for a few (only 8) highly rigid proteins, all DNA-binding proteins undergo one or more types of conformational changes upon complex formation. The RMSD values for these groups are distinct and follow the general order (mean values in Å):

- **ID** (8.3) < **OC** (5.2) < **OD** (3.9) < **SA** (3.3) < **LL** (2.8) = **SS** (2.8).

Consistently GDT_TS5 follows the reverse order (see Methods for the definitions of GDT_TS5 and RMSD):

- **ID** (66.6) < **OC** (72.1) < **OD** (81.9) < **SA** (86.7) < **SS** (87.1) < **LL** (87.4).

In order to understand relationships between different types of conformational changes and structural and functional categories, we annotated each protein with its functional class, DNA-binding motifs and the nature of the targets (specific/nonspecific). We then analyzed the distribution of conformational changes across these categories (Supporting Information Table S1 and Figure 2). Although various types of conformational changes are present across functional categories and DNA-binding motifs, members with largest degree of conformational changes (about 2/3rd of all proteins with RMSD>3.0 Å and 100% with GDT_TS5<85%) are either polymerases or endonucleases. In particular, changes characterized by open/close and inter-domain motions are prevalent only in endonucleases. However, these conformational changes do not appear to be essential for the endonuclease function, since some endonucleases such as *E. Coli* endonuclease IV and Nei endonuclease VIII are rigid with less than 1.0 Å RMSD.

Structural proteins along with some enzyme groups show only small changes. Among them, glycosylases, ligases, and other unclassified enzymes consistently show only small local loop or secondary structure transition (with the only exception of DNA-helicase II, which undergoes a large domain rearrangement upon DNA-binding). Similarly conformational changes in structural proteins vary in a narrow range (of RMSD values between 0.7 and 1.6 Å and GDT_TS5 between 95.5 and 99.9%) with only a couple of exceptions. Transcription factors lie between these two general types (large and small conformational changes), suggesting diversity in their recognition mechanism.

While functional categories and DNA-binding motifs are only loosely connected to the nature of conformational changes, a more convincing relationship of the latter is observed with the specificity of the targets recognized by these proteins. Overall, the average RMSD for proteins recognizing specific DNA-targets is significantly higher than those recognizing nonspecific DNA sequences [Fig. 3(a)]. If we look at the types of changes observed in the two groups, we find that all nonspecific proteins undergo only small or localized changes such as the SS transition or the LL movements, whereas the specific binding proteins undergo all types of changes to form a stable complex. None of the nonspecific DNA-binding proteins show large-scale conformational changes classified above as OC, ID, or SA types of motions [Fig. 3(b)].

In summary, conformational changes are widely distributed and although differences do appear between functional and structural categories on the one hand and the degree and type of conformational changes on the other, the nature seems to have optimized different oligomeric states, conformational changes and recognition mechanisms on a case-to-case basis, and the best available interaction mechanism is used in individual
Figure 1
Types of conformational changes between DNA-bound proteins compared to their unbound forms (blue is the structure in complex with DNA, superimposed on the unbound conformation in green). Red arrows highlight sites of conformational changes.
instances of similar interactions. Notwithstanding this argument, some groups (e.g., nonspecific target binding proteins) do show a much narrower range in the degree of conformational changes than others, suggesting that these groups have optimized a universal binding mechanism.

**Residue-wise displacements and packing density**

To characterize the preference of each amino acid to undergo conformational changes in DNA-binding proteins, we computed the average displacement of all 20 amino acid residue types in the entire data set. The propensity of each of the 20 amino acids was computed for the interface and noninterface residues (see Methods) and the ratio of the two propensity values was used to estimate whether the residue prefers to undergo greater conformational changes in the interface (Supporting Information Table S2). The scatterplot of displacements observed in the interface compared to the whole protein can be seen in Figure 4(a). To understand whether a relationship exists between the residue-wise displacements and changes in packing density, the latter was also computed in the interface and noninterface positions [Figure 4(b), detailed data in Supporting Information Table S3].

Figure 4(a) shows that most residues undergo displacements in the interface and noninterface positions with comparable magnitudes (average interface displacements are less than twice the average displacement anywhere in the structure). On the other hand packing density of residues undergoes greater changes in the interface. This implies that physical displacements of residues may not always be accompanied by an increase in packing density. Residue-wise differences in Figure 4(a,b) also indicate that changes in packing density are less dependent on the residue type than the displacements are. Surprisingly, Arg, the most important residue for DNA-binding, does not show either a large change in packing density or more than the average degree of
Presumably positively charged residues do not require special conformational changes to bind DNA, as they primarily interact with the backbone. On the other hand, some hydrophobic residues such as Cys, Ala, and Gly undergo much more conformational adjustments in the interface compared to elsewhere. An analysis of propensities for conformational changes and packing density has been reported for a smaller number of proteins by Bhardwaj and Gerstein and our results are in broad agreement with theirs. However, DNA-binding proteins owing to the electrostatic nature of protein-DNA interactions provided additional insights, particularly in the comparison between interfaces and overall structures.

**Bulk electrostatic properties and conformational changes**

Since charged residues play a major role in DNA-binding, we compared the bulk electrostatic properties of proteins in different groups of conformational changes (Table III). Despite the small sample size and thus, limited statistical significance, several general trends are apparent. For example, the only protein in the open/close conformational switch (OC) group with complete coordinates (DBP52 data) has a much larger positive charge than the proteins in the other groups. Yet, the net charge on the interface residues computed from the unbound conformation of this protein is smaller than the corresponding value for the proteins outside this group. On the other hand, the average dipole moment of this protein is higher than the control. (In general, we took all the proteins other than those in the target group as the control.) In contrast, the single arm (SA) conformational changes are associated with proteins having a lower dipole moment but a more positive net charge on the interface. Thus the open/close conformational switch seems to occur through dipole interactions and possibly brings more distant positive charges towards the interface, whereas single arm motions do not involve dipole interactions or significant rearrangements of charges, as the interface is already positively charged in the unbound form.

**Conformational changes and the nature of protein-DNA contacts**

One of the special features of protein–DNA interactions, in contrast to protein–protein or protein–ligand interactions, is the presence of well-defined atomic groups in DNA, which characterize their geometrical and physicochemical properties. For example, backbone atoms are negatively charged, whereas base atoms support stacking and hydrogen-bonded interactions. Thus, we made a comparison between conformational changes and protein atomic contacts in each of the four geometrical regions of DNA viz., the nucleic acid base, the backbone, and the major and minor grooves (Table IV). It was observed that proteins in the rigid group show the smallest number of DNA-contacts (9.1 residue per protein compared to 17.3 overall average), whereas proteins with large-scale conformational changes in DBP52 (the SA and ID groups) had a larger number of DNA-contacts. Highly rigid proteins also had a lower net charge on the interface, as shown in Table III. Thus it can be hypothesized that conformational changes enhance the stability of protein-DNA complexes, i.e., rigid proteins form less stable protein–DNA complexes than those undergoing any form of conformational changes. (Further support to this hypothesis will be provided in a later section on the stability analysis of proteins with conformational changes.)

The analysis of DNA-contacts in various geometrical regions provided further insights into contact distributions across conformational changes. For example, in the rigid group of proteins (on the average) 35% DNA-binding residues have at least one contact with DNA-base atoms, compared with smaller values overall (21%) and in most other groups (e.g., 21% in SS group and 20% in LL group). The numbers of the major and minor...
| Function | Complex | Free | Name | Part of DBPS2 | Conformational change observed | Chain RMSD (Å) | Complex RMSD (Å) | GDT_TS5 (chain) (%) |
|----------|---------|------|------|--------------|-------------------------------|--------------|-----------------|-------------------|
| endonucleases | 2oaaA | 2o9B | Restriction endonuclease MvaI | No | SS,OD,OC,LL | 7.29 | 7.91 | 61.3 |
| 3bamA | 1bamA | Type II restriction endonuclease BamHI | No | SS,OD,SA,LL | 6.07 | 4.83 | 96.6 |
| 1ipgA | 1evxB | I-Ppol His-Cys homing endonuclease | Yes | SS,LL | 0.91 | 0.97 | 96.6 |
| 1v15D | 1f5jB | HNH endonuclease | Yes | SS,LL | 1.28 | 0.63 | 97.9 |
| 1khvB | 2auA | Type II restriction endonuclease HincII | No | SS,OD,OC,LL | 1.37 | 1.38 | 93.1 |
| 1qpsA | 1qv3C | EcoRI endonuclease | No | SS, OD, LL | 2.11 | 2.08 | 93.3 |
| 1eyuB | 1puB | Type II restriction endonuclease PvuII | No | SS,ID,LL | 2.7 | 2.82 | 82.5 |
| 1wteA | 1wtdB | Type II restriction endonuclease EcoO109I | No | SS,OD,OC,LL | 3.13 | 3.40 | 74 |
| 1iawB | 1ev7A | Type II restriction endonuclease NaeI | No | SS, OD, LL | 3.46 | 2.69 | 71.4 |
| 1dfmB | 1es8A | Restriction endonuclease BglII | No | SS,OD,ID,LL | 4.63 | 4.35 | 87.5 |
| 1stxA | 1rveB | Type II restriction endonuclease EcoRV | Yes | SS,LL | 1.94 | 1.88 | 88.9 |
| 2fkcB | 1ynmA | Restriction endonuclease HinP1I | Yes | SS,SA,LL | 4.41 | 4.39 | 91.1 |
| 3a46B | 3a45B | Endonuclease VIII (Nei) DNA glycosylase | Yes | SS,LL | 0.69 | 0.66 | 96.3 |
| glycosylases | 3g3B | 2o7mB | Homing endonuclease I-CreI | Yes | SA, LL | 3.85 | 2.72 | 90.1 |
| 3c25B | 3vqB | Restriction endonuclease NotI | No | SS,OD,OC,LL | 1.94 | 1.82 | 78.6 |
| 2nqjA | 2nqhA | E.Coli endonuclease IV | Yes | SS | 0.46 | 0.46 | 99.3 |
| 2vs7A | 2i3pB | Homing endonuclease I-CreI | Yes | SA, LL | 3.85 | 2.72 | 90.1 |
| 3c25B | 3vqB | Restriction endonuclease NotI | Yes | SS,OD,OC,LL | 1.94 | 1.82 | 78.6 |
| 2nqjA | 2nqhA | E.Coli endonuclease IV | Yes | SS | 0.46 | 0.46 | 99.3 |
| polymerases | 3frA | 3foaA | DNA polymerase from enterobacteria phage RB69 | Yes | SS,OD,LL | 5.21 | 5.21 | 65.2 |
| 3mfaA | 1hjB | DNA polymerase eta | No | SS,SA,LL | 3.31 | 3.17 | 76.2 |
| 3tapa | 1pxA | DNA polymerase I | Yes | SS,SA,LL | 1.39 | 1.39 | 91.8 |
| 3magA | 3skOA | DNA polymerase II | No | SS, OD, LL | 2.58 | 2.53 | 74.5 |
| 3qjaA | 5kqaA | Thermus thermophilus uracil-DNA polymerase | Yes | SS | 0.57 | 0.57 | 98.8 |
| 2pyjB | 1rxhD | Phage B-family replicative DNA polymerase | Yes | SS,OD,LL | 2.06 | 1.46 | 78.9 |
| ligases | 3q8A | 2diA | Y-family DNA polymerase IV | No | SS,OD,ID,LL | 11.52 | 11.42 | 65.4 |
| 3bebB | 3qbaA | DNA clamp polymerase III | Yes | SS | 0.62 | 0.68 | 99 |
| 3rh6A | 1bpaA | DNA polymerase beta | No | SS,OD,OC | 11.9 | 11.90 | 60.2 |
| 3sdpA | 3spaA | DNA polymerase beta | No | SS,OD,OC | 11.9 | 11.90 | 60.2 |
| 2r9B | 2riA | Neighbor homologous end joining (NHEJ) DNA-ligase | No | SS,OD,LL | 0.49 | 0.50 | 98.8 |
| transfers | 2g1pA | 2oreF | DNA-adenine-N6-methyltransferase | No | SS | 0.6 | 0.48 | 98.6 |
| 3bkZa | 3qnaA | Alpha-ketoglutarate-dependent dioxygenase aikB | Yes | None | 0.66 | 0.66 | 97.9 |
| 1m5rB | 1qkaA | T4 phage beta-glucosyltransferase | Yes | SS,LL | 0.41 | 0.31 | 99.6 |
| 1fjaA | 2hmyB | HHAI methyltransferase | Yes | SS,SA,LL | 3.92 | 3.92 | 88.4 |
| 2q9D | 2admB | Adenine-specific methyltransferase T7Q | No | SS, OD, LL | 1.17 | 1.09 | 92.8 |
| other enzymes | 2is8B | 3fiaA | DNA helicase II | No | SS,OD,ID,LL | 13.58 | 13.49 | 45.6 |
| 1dnkA | 3dniA | DNase I | No | SS,OD,LL | 0.36 | 0.35 | 99.8 |
| 2ve9C | 2ve8E | DNA translocase FTSK | Yes | Ll | 0.39 | 0.79 | 99.6 |
| 3dsdB | 1i7A | DNA double-strand break repair protein mre11 | Yes | Ll | 0.59 | 0.61 | 98.2 |
groove contacts relative to the whole protein length are also higher in the rigid group proteins (data not shown). On the other hand, the fraction of backbone contacting residues is always high (>80% in all groups), with no statistically significant differences between the groups. Thus, even though the rigid groups of proteins form a smaller number of contacts, their geometry seems to be suitable enough to fit well in the DNA structure and form contacts with the major and minor groove atoms as well as the base atoms in the helix interior.

Agreement with the intrinsic flexibility of the unbound protein

Intrinsic flexibility has been measured at the residue level by the observed temperature factors (B-factors)
Table II
Redundant Lists of Proteins Within Each Conformational Change Group (one protein may belong to more than one groups)

| Conformational change group                     | Number of proteins | PDB Ids (with selected chain names) of protein-DNA complex |
|------------------------------------------------|--------------------|----------------------------------------------------------|
| Highly rigid                                    | 8                  | 1j4A, 1mwiA, 1utdB, 2bnwaA, 2qhbB, 2rfaA, 3bzxaA, 3c0R   |
| Secondary structure change (SS)                 | 68                 | 1b3tA, 1b1DA, 1dfmB, 1dknA, 1eyuB, 1fxaA, 1h8cA, 1iawB, 1ppA, 1qsaA, 1reaA, 1txaA, 1v5DF, 1wtxA, 1xvbA, 1zs4C, 2d6A, 2c1CA, 2eFwA, 2LCB, 2r1CB, 2g1PA, 2s5B, 2g3D, 2nqA, 2oaAA, 2p7FA, 2p10D, 2p2YB, 2r9IB, 2s7A, 2w42A, 2xwvB, 2xrvB, 2yvhB, 2zkdB, 3a46B, 3BaM, 3BeP, 3c25B, 3craA, 3cwaB, 3e6CC, 3h5BA, 3jxoA, 3jy1A, 3kaMA, 3mqaA, 3msHA, 3nibB, 3o0S, 3p0vA, 3q05B, 3zq8A, 3y3HA, 3sm4B, 3spdA, 3tapA |
| Local loop motion (LL)                          | 69                 | 1b3tA, 1b1CA, 1b1GA, 1dfmB, 1dknA, 1eyuB, 1fxaA, 1h8cA, 1iawB, 1ppA, 1qsaA, 1reaA, 1txaA, 1v5DF, 1wtxA, 1xvbA, 1zs4C, 2d6A, 2c1CA, 2eFwA, 2LCB, 2r1CB, 2s5B, 2g3D, 2nqA, 2oaAA, 2p7FA, 2p10D, 2p2YB, 2r9IB, 2s7A, 2w42A, 2xwvB, 2xrvB, 2yvhB, 2zkdB, 3a46B, 3BaM, 3BeP, 3c25B, 3craA, 3cwaB, 3e6CC, 3h5BA, 3jxoA, 3jy1A, 3kaMA, 3mqaA, 3msHA, 3nibB, 3o0S, 3p0vA, 3q05B, 3zq8A, 3y3HA, 3sm4B, 3spdA, 3tapA |
| Single arm movement (SA)                        | 10                 | 1b3tA, 1b1CA, 1b1GA, 1dfmB, 1dknA, 1eyuB, 1fxaA, 1h8cA, 1iawB, 1ppA, 1qsaA, 1reaA, 1txaA, 1v5DF, 1wtxA, 1xvbA, 1zs4C, 2d6A, 2c1CA, 2eFwA, 2LCB, 2r1CB, 2s5B, 2g3D, 2nqA, 2oaAA, 2p7FA, 2p10D, 2p2YB, 2r9IB, 2s7A, 2w42A, 2xwvB, 2xrvB, 2yvhB, 2zkdB, 3a46B, 3BaM, 3BeP, 3c25B, 3craA, 3cwaB, 3e6CC, 3h5BA, 3jxoA, 3jy1A, 3kaMA, 3mqaA, 3msHA, 3nibB, 3o0S, 3p0vA, 3q05B, 3zq8A, 3y3HA, 3sm4B, 3spdA, 3tapA |
| Order/disorder transition (OD)                  | 28                 | 1b1DA, 1dfmB, 1dnkA, 1eyuB, 1fxaA, 1h8cA, 1iawB, 1ppA, 1qsaA, 1reaA, 1txaA, 1v5DF, 1wtxA, 1xvbA, 1zs4C, 2d6A, 2c1CA, 2eFwA, 2LCB, 2r1CB, 2s5B, 2g3D, 2nqA, 2oaAA, 2p7FA, 2p10D, 2p2YB, 2r9IB, 2s7A, 2w42A, 2xwvB, 2xrvB, 2yvhB, 2zkdB, 3a46B, 3BaM, 3BeP, 3c25B, 3craA, 3cwaB, 3e6CC, 3h5BA, 3jxoA, 3jy1A, 3kaMA, 3mqaA, 3msHA, 3nibB, 3o0S, 3p0vA, 3q05B, 3zq8A, 3y3HA, 3sm4B, 3spdA, 3tapA |
| Open/close conformational change (OC)            | 7                  | 1eyuB, 1wteA, 1xvhB, 2oaAA, 2c25B, 3x0A, 3h6A            |
| Inter-domain motion or quaternary structural changes (ID) | 10                 | 1bg1A, 1dfmB, 1h8cA, 2s6B, 2zpfA, 3craA, 3isoB, 3o9xA, 3q05B, 3qz8A |

Table III
Bulk Electrostatic Properties of DNA-Binding Proteins Grouped by the Type of Conformational Changes Observed in Them

| Proteins with complete atomic coordinates data (DBP52) |
|--------------------------------------------------------|
| Counts | Total charge | Interface charge | Complex moment (complex) | Unbound Dipole moment (P-value) | Complex Quadrupole moment (P-value) | Unbound quadrupole moment (P-value) |
|--------|--------------|------------------|--------------------------|-------------------------------|-----------------------------------|-----------------------------------|
| Rigid  | 7            | 1.4 (0.99)       | 3.0 (0.21)               | 0.65 (0.82)                   | 0.58 (0.71)                       | 1.80 (0.05)                       |
| LL     | 34           | 1.8 (0.51)       | 4.1 (0.22)               | 0.66 (0.15)                   | 0.64 (0.76)                       | 2.08 (0.04)                       |
| SS     | 33           | 0.8 (0.27)       | 4.0 (0.61)               | 0.58 (0.08)                   | 0.57 (0.35)                       | 0.63 (0.47)                       |
| OD     | 0            | NA               | NA                       | NA                            | NA                                | NA                                |
| SA     | 7            | −0.1 (0.49)      | 6.4 (0.002)              | 0.50 (0.31)                   | 0.47 (0.20)                       | 0.04 (0.41)                       |
| OC     | 1            | 5.0 (0.18)       | 3.0 (0.19)               | 0.72 (0.03)                   | 0.86 (0.03)                       | 1.94 (0.84)                       |
| ID     | 3            | 0.3 (0.87)       | 4.7 (0.44)               | 0.57 (0.61)                   | 0.66 (0.80)                       | 2.35 (0.44)                       |
| Average (DBP52) |         | 1.44          | 3.86                     | 0.62                         | 0.6                              | 1.1                               |

Mean values in each category are compared with control (all data excluding proteins from that category; not shown in this table). Those with a P-value (from a t-test) less than 0.2 (indicating 80% statistical significance) have been shown in bold. Many other differences are noteworthy but the data is too small (only a few proteins in each group) to return statistically significant P-values from t-test. Main comparison for the free proteins is done in DBP52 data only due to unreliable estimates in proteins with missing data (In most cases, coordinate data of is complete in the complex structure and DBP90 can be used for those values). All values are in atomic units.

Conformational Changes in DNA-Binding Proteins

provided in the PDB coordinate files and also by normal mode analysis (NMA) based estimates using an elastic network model (see methods). B-factors represent the intrinsic flexibility of atoms as measured during the X-ray diffraction experiments while determining the three-dimensional structure. We investigated the relationship between intrinsic flexibility and conformational changes by computing a protein-wise correlation coefficient between atomic displacements of each Cα-atom with the B-factors in the unbound protein structure. The average correlation coefficient between residue displacements and available B-factor values ranges from 0.19 (for inter-domain conformational change group) to 0.39 (for local loop group) (Table V) suggesting that B-factors contain...
some useful information about the residues likely to undergo conformational changes on complex formation.

To estimate the role of intrinsic flexibility in conformational changes, we used normal mode analysis. Normal modes have been successfully used to estimate the extent and direction of conformational changes in protein–protein and protein–ligand interactions and some individual classes of DNA-binding proteins.\(^4^8\)–\(^5^3\) and it has been shown that most of the protein movements can be approximated by a few low-frequency normal modes.\(^4^1\),\(^5^4\) We first computed a protein-wise correlation between the \(\text{C}_\alpha\) displacements observed by superimposing unbound and bound structures and those predicted by different normal modes. In this way corresponding to each of the 12 nontrivial lowest-frequency normal modes we obtained a correlation score for each protein. The normal mode whose displacement values were in the best agreement with the observed displacement values was selected to determine the best-case scenario as shown in Table V. We found that the average protein-wise correlation between observed and normal mode-predicted displacements in 52 proteins varied between 0.30 (for the open/close conformational group) and 0.67 (for the single arm movement group). Thus, normal modes contain potentially useful information about positions undergoing conformational changes.

Table IV

| Group | Number of proteins | Average \# of DNA contacting residues, contacting base atoms | Relative \# of DNA contacting residues, contacting backbone | Relative \# of DNA contacting residues, contacting with major groove | Relative \# of DNA contacting residues, contacting with minor groove |
|-------|-------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| DBP52 data | | | | | |
| rigid | 7 | 9.1 (0.0002) | 0.35 (0.07) | 0.8 (0.3) | 0.34 (0.0002) |
| LL | 34 | 19.9 (0.003) | 0.20 (0.5) | 0.92 (0.4) | 0.20 (0.8) |
| SS | 33 | 20.2 (0.002) | 0.21 (0.7) | 0.92 (0.5) | 0.16 (0.2) |
| OD | 10 | 52.1 (0.02) | 0.25 (0.9) | 0.95 (0.16) | 0.23 (0.4) |
| SA | 7 | 14.0 (0.8) | 0.07 (0.02) | 1.0 (0.2) | 0.21 (0.6) |
| OC | 1 | 23.3 (0.5) | 0.22 (0.9) | 0.86 (0.5) | 0.25 (0.2) |
| Overall | 52 | 17.3 | 0.21 | 0.91 | 0.20 |
| DBP90 data | | | | | |
| rigid | 8 | 9.5 (9.4e-05) | 0.35 (0.04) | 0.82 (0.2) | 0.35 (4.0e-05) |
| LL | 69 | 21.6 (7.8e-06) | 0.21 (0.6) | 0.91 (0.7) | 0.19 (0.30) |
| SS | 69 | 21.9 (1.3e-05) | 0.21 (0.8) | 0.91 (0.9) | 0.18 (0.5) |
| OD | 28 | 23.9 (0.01) | 0.20 (0.6) | 0.91 (0.9) | 0.18 (0.5) |
| SA | 10 | 29.4 (0.007) | 0.20 (0.91) | 0.94 (0.1) | 0.20 (0.8) |
| OC | 7 | 27.9 (0.02) | 0.17 (0.2) | 0.92 (0.6) | 0.23 (0.5) |
| ID | 10 | 21.5 (0.3) | 0.21 (0.9) | 0.90 (0.6) | 0.22 (0.6) |
| Overall | 90 | 19.6 | 0.22 | 0.90 | 0.20 |

Number of contacts within each category represents the mean of number of residues that are within 3.5 \(\AA\) of an atom of the DNA within that definition (e.g., nucleotide base or major groove) relative to the total number of residues of that protein in contact with any DNA atom. Although a statistical test of significance is difficult for such a small number of data points, \(P\)-values from a \(t\)-test for each group compared with the remaining proteins are anyway provided in brackets. Note that the four contact definitions are overlapping and do not include all protein-DNA contacts implying that the numbers may not add up to 100%. From the data, we observed that although the number of backbone contacts is similar in all categories, rigid group shows lower number of contacts with nucleic acid base atoms as well as with atoms in DNA major and minor grooves, suggesting lower target specificity of rigid proteins.

Table V

| Group | Number of proteins | Correlation with the cumulative displacement | Correlation with displacement in the best correlated normal mode | Correlation with B-factors in the unbound structure (proteins without B-factor values excluded) |
|-------|-------------------|-----------------------------------------------|---------------------------------------------------------------|----------------------------------------------------------------|
| LL | 33 | 0.48 | 0.60 | 0.39 |
| SS | 32 | 0.41 | 0.67 | 0.33 |
| SA | 5 | 0.53 | 0.56 | 0.39 |
| OC | 1 | 0.08 | 0.30 | NA |
| ID | 3 | 0.33 | 0.59 | 0.19 |

Cumulative displacement is computed by taking vector summations of amplitudes in first 12 nontrivial normal modes, as indicated by the corresponding eigen vectors. Best matching normal mode is selected as a single normal mode whose displacements correlate the best with the observed displacements.
observed displacements and a vector sum of the predicted amplitudes in the direction of eigen vectors using the first 12 normal modes. The correlation coefficients were lower than the best-matching criterion as expected but remained high enough to indicate that normal mode analysis provides important information about conformational changes.

To estimate whether the overall/global change that a protein undergoes upon complex formation can be well-estimated by a normal mode analysis, we computed the collectivity of each unbound protein, which measures the number of mobile atoms in a protein for a particular frequency mode under an elastic network model. These values, which we call the predicted collectivity, were then compared with observed collectivity values (see methods).

Using the collectivity of the best matching normal mode (defined by the maximum agreement between Cα displacements), the correlation between the predicted and observed collectivity values of proteins in DBP52 data set was found to be 0.36. These results suggest that despite a useful guide provided by normal mode analysis, even the best matching normal mode cannot accurately estimate the nature of collective motions expected on DNA-binding. Using the average collectivity from multiple modes did not improve the correlation coefficient (data not shown). This relatively low correlation was lower than the values reported earlier for protein-protein complexes, possibly because DNA interactions significantly alter the conformational ensemble populations on complex formation. Yet, the fact that at least one normal mode carries useful information about sites of conformational changes in proteins, will be helpful in narrowing down the search for estimating conformational changes expected on interface formation.

### Conformational changes and free energy of complex formation

In our data set, only 12 proteins had known thermodynamic data for free energy of protein-DNA complex formation, available from PRONIT. This included proteins which are co-clustered with DBP90 protein chains at 25% sequence identity threshold used in our original data set. We compared free energy values with the RMSD and GDT_TS5 of bound/unbound forms as shown in Table VI. The two measures of conformational changes are found to be consistently correlated with the free energy changes. The correlation coefficients between RMSD and GDT_TS5 on the one hand and free energy change on the other were observed to be −0.84 and 0.70 (Fig. 5). Since one protein BamHI (PDB 1bam) seems to contribute most significantly to this correlation, we tested the correlation coefficients after removing this protein. The correlation coefficients by excluding this protein dropped to 0.47 and 0.32, respectively. However, even these lower values of correlation seem to be sufficiently indicative of a role of conformational changes in determining the stability of protein-DNA complexes. Moreover the high correlation in the data when BamHI is present need not be discarded, because this protein has the highest degree of conformational changes among all the proteins considered. In the above sections, we found that proteins with larger conformational changes have higher numbers of DNA contacts and also large dipole moments and positive charges in their interface. All these results point out that proteins undergoing conformational changes may form more stable complexes compared to the ones undertaking rigid-body recognition.

### DISCUSSION

In this work, we analyzed the distribution of conformational changes across functional and structural classes of DNA-binding proteins and established that they are widely distributed across each category. In proteins with large conformational changes, such as open/close conformational switching and inter-domain movements, we observed that charges on the interface and bulk dipole moments play important roles, given the predominantly

| Serial No. | PDB ID | RMSD (Å) | GDT_TS5 (%) | ΔG (Kcal/mol) |
|------------|--------|----------|-------------|--------------|
| 1          | 1evx/leww | 0.91     | 96.6        | −11.3        |
| 2          | 3cro/2cro  | 0.6      | 98.8        | −9.13        |
| 3          | 1jey/1jeq  | 0.81     | 96.0        | −12.1        |
| 4          | 3ktu/1k09 | 0.89     | 95.5        | −3.34        |
| 5          | 2g1p/2ore  | 0.6      | 99.6        | −8.91        |
| 6          | 1b3t/1vhi  | 4.42     | 89.7        | −12.3        |
| 7          | 1bam/1bhmi | 6.07     | 83.9        | −39.9        |
| 8          | 1qc9/1ckq  | 2.11     | 93.3        | −15.7        |
| 9          | 1emh/1tkz  | 1.46     | 87.3        | −8.0         |
| 10         | 1axl/1axz  | 0.69     | 94.4        | −10.0        |
| 11         | 1up1/1h4l  | 1.07     | 98.2        | −8.8         |
| 12         | 1lau/1udh  | 0.46     | 99.6        | −2.8         |
| Overall correlation with ΔG | −0.83 | 0.84 |
electrostatic nature of protein-DNA interactions. We observed that residues with high surface propensities are likely to undergo larger displacements on complex formation. These residues also show higher packing density in the complex, with the exception of negatively charged residues, in which packing density remains preserved. Furthermore, the dependence of conformational changes on the size of the interface, and the relative number of contacts in the major and minor grooves, has been noticed. Many of these results are in agreement with those observed for proteins’ complexes with other molecules. However, the electrostatic nature of interactions and exclusion of negatively charged residues from the interface is special to protein-DNA complexes. Similarly the concept of major and minor grooves is exclusive to these interactions and our analysis of contacts in these regions provided us with clues not only to understanding conformational changes in DNA-binding proteins but also the process of DNA recognition by proteins in greater details.

Particularly, our interest was in (a) detecting whether the intrinsic flexibility of unbound proteins is enough to explain and predict the nature of conformational changes that are likely to occur on complex formation and (b) the impact of conformational changes on the stability and specificity of a complex. The first of these questions relates to the relevance of the theory of conformational selection to protein–DNA recognition. The degree of conformational changes that a protein undergoes upon complex formation was estimated from the intrinsic thermal motions in the unbound state. In the past, intrinsic fluctuations of individual proteins have been used to gain insights into the nature of conformational changes in protein–protein interactions\(^\text{4}\) and to predict binding sites in DNA-binding protein.\(^\text{5}\) Our results confirm the role of intrinsic flexibility in conformational changes in DNA-binding proteins as shown by (a) a significant correlation between temperature factors and observed movements of individual residues, (b) at least one of the lowest-frequency normal modes in the unbound protein carrying significant information about the site of conformational changes, and (c) a reasonable positive correlation between observed and predicted collectivity. However, the exact values of correlation coefficients are still low, indicating the limitations of intrinsic flexibility as a predictor for conformational changes.

Closer inspection of individual proteins (data not shown) revealed that the poorly correlated examples fell in two group’s, i.e., proteins in which (a) conformational changes were well-estimated by normal modes but the residue-wise displacements were not sufficient to account for the observed changes (e.g., a change in secondary structure may occur without large displacements in the main chain atoms of the protein) and (b) those in which conformational change occurred close to the DNA-interface, making their estimates from the unbound structure more challenging. First of these issues can be addressed by a case-by-case analysis of conformational changes. However, the second issue requires a priori knowledge of DNA-interface from the unbound structure. Fortunately, a number of successful methods to predict DNA-binding sites from amino acid sequences alone are available\(^\text{57,58}\) and the predicted binding site information from these methods is likely to aid in estimating conformational changes from the normal mode analysis. This approach has a distinct advantage in the case of DNA-binding proteins, compared with other complexes as it has been shown that DNA-binding residues can be estimated with a high degree of accuracy (from sequence information alone) (AUC of ROC ~80%), whereas the best performing methods for predicting protein–protein interaction sites lag behind, although the latter could also be significantly improved by including structural and evolutionary features.\(^\text{59,60}\)

Although intrinsic flexibility has been found to be useful in estimating the degree of conformational changes, it does not suggest a biological context under which the conformational change becomes necessary or the reasons why the nature selected flexible proteins to recognize some targets while rigid proteins do the job in other systems. Our results provide evidence that conformational changes aid in enhancing the stability of protein–DNA complexes and allow a DNA-binding protein to interact more specifically with its DNA targets. The evidence for the improved stability mediated by conformational changes was established by; (a) the number of DNA contacts in the major and minor groove, is higher for proteins undergoing larger conformational changes; (b) net charges and dipole moments of highly rigid proteins are lower than those undergoing any form of conformational changes and most importantly; (c) the experimentally observed free energy of protein-DNA complexes has a strong correlation with the degree of conformational changes observed on complex formation.
Our results also suggest that conformational changes, proposed to mediate a more stable protein–DNA complex, might themselves be mediated by the specificity of target recognition. Support for this argument comes from the fact that nonspecific DNA-binding proteins undergo much smaller conformational changes than the specific ones [Fig. 3(a)]. These observations imply that large conformational changes in high specificity proteins could be a general consequence of target recognition rather than its cause, as already pointed out by some investigators.61 For example, in the case of BstYI endonuclease, the only hemispecific complex in our data set, it has been reported that the structure of the complex with a noncognate (hemispecific target) DNA is closer to the unbound form rather than the one observed in its complex with the cognate fully specific complex.62 Our results also support the view that an unbound DNA-binding protein has sufficient signal to interact weakly with any DNA sequence (also evident from the high performance of sequence-based prediction methods, e.g., Refs. 57 and 63). However, specific target recognition and complex formation requires specific structures leading to the selection of a highly specific conformation from the conformational ensemble in the unbound state or even create a structure that is outside of the conformational ensemble. This part of the recognition presumably employs the induced fit mechanism to bind. The local folding hypothesis explaining the induced fit mechanism of specificity has been debated in the literature for long (e.g., Ref. 64). Recent studies suggest that conformational selection is the primary driver of molecular recognition, which is fine-tuned by induced fit, through which conformations not explicitly present in the unbound ensemble could be formed and lead to the eventual stabilization of the complex.8 The role of induced fit from the DNA perspective has also been confirmed recently by direct measurements of $k_{on}$ and $k_{off}$ rates in the p53 protein in complex with its cognate DNA sequences.65,66

The ability of proteins to bind noncognate DNA with small differences in free energy compared with cognate targets is well established (e.g., Ref. 65). This observation supports the opinion stated above that the unbound protein structure has sufficient information to bind any DNA, even if the final structure in the cognate complex is significantly different. The observed changes in structure probably arise for the purpose of bringing about additional stability to the complex and enabling downstream biological functions. There are several studies showing that DNA-binding may stabilize proteins and protect them against misfolding and aggregation, implying that stable structures could be created as a consequence of DNA-recognition.66 A conformational switch leading to a passage from nonspecific to specific target recognition in the SRY protein has been recently established by a molecular dynamics study,67 again illustrating a link between conformational changes and specificity. One wonders, if a specific conformation is most suitable to recognize specific DNA-targets, why the protein does not exist in this structure in the unbound form. Our results suggest that flexible and nonspecific interactions may be required to increase the number of chance encounters between proteins and DNA and of other important biological events such as sliding and steering preceding the specific recognition of targets, as reported in related studies.44,68,69

As discussed above, conformational changes enhance the stability of protein–DNA complexes, which apparently is the primary consequence of conformational changes. This conclusion, however, primarily relates to the conformational changes close to the DNA interface. Our results indicate that often conformational changes brought about by DNA-binding propagated to residues far from the interface (Fig. 4). In some of these proteins, we found that conformational changes stabilized protein structures rather than the complex, e.g., in the case of order/disorder transitions. Yet another critical consequence of DNA-induced conformational changes is the ability of proteins to recruit cofactors through allosteric changes. This phenomenon is in contrast to the cofactor binding preceding and triggering specific DNA-binding in some proteins (e.g., Ref. 70). Some of the experimental data as in the case of ChIP-seq experiments cannot clearly reveal the exact order of binding between these entities. However, the existence of crystal structures of one protein in complex with DNA alone and then with both a cofactor and DNA suggests that at least in such cases the cofactor binds after the stabilization by the conformational changes introduced upon DNA recognition. There are several reports of cofactor induction and allosterically initiated interactions of proteins with other molecules. For example, specific DNA-binding was shown to be an allosteric effector of functional interactions of the steroid receptor with the targets of transcriptional activation.71 Similarly HAP1-DNA interactions have been implicated for the allosteric effects on transcription activation events.72 Presumably, these allosteric events are mediated by conformational changes induced by DNA-binding, bringing forward another important consequence of conformational changes other than stabilizing protein-DNA complexes. The universality of the allosteric role of conformational changes can be investigated further in future studies.

**CONCLUSIONS**

Protein–DNA interactions involved in a diverse set of biological functions employ a range of conformational switches to fine-tune their interactions. Proteins with large conformational changes form more stable complexes and show greater specificity than their rigid counterparts. Intrinsic flexibility of the unbound proteins plays an...
important role in conformational changes but that alone cannot explain all aspects of this phenomenon, particularly in proteins with large interfaces. Improved stability and specificity are shown to be the primary consequences of conformational changes in DNA-binding proteins.

ACKNOWLEDGMENTS

This study was in part supported by the Industrial Technology Research Grant Program in 2007 from New Energy and Industrial Technology Development Organization (NEDO) of Japan, and also by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology (Grant Numbers 25430186, 25293079 and 22500277) and from the Ministry of Health, Labor, and Welfare.

REFERENCES

1. Reddy CK, Das A, Jayaram B. Do water molecules mediate protein-DNA recognition? J Mol Biol 2001;314:619–632.
2. Jayaram B, Jain T. The role of water in protein-DNA recognition. Annu Rev Biophys Biomol Struct 2004;33:343–361.
3. Betts MJ, Sternberg MJE. An analysis of conformational changes on protein–protein association: implications for predictive docking. Protein Eng Design Selection 1999;12:271–283.
4. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci USA 2005;102:18908–18913.
5. Dobbins SE, Lesk VL, Sternberg MJE. Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–DNA docking. Proc Natl Acad Sci USA 2008;105:10390–10395.
6. Tsai CJ, Ma B, Nussinov R. Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci USA 1999;96:9970–9972.
7. Martin G. How to mark off paths on the protein energy landscape. Proc Natl Acad Sci USA 2009;106:18879–18880.
8. Boehr DL, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 2009;5:789–796.
9. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein–protein interactions. Curr Opin Struct Biol 2004;14:104–109.
10. Zheng W, Brooks BR. Normal-modes-based prediction of protein conformational changes guided by distance constraints. Biophys J 2005;88:3109–3117.
11. Dickerson RE. DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res 1998;26:1906–1926.
12. Young MA, Ravishanker G, Beveridge DL, Berman HM. Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. Biophys J 1995;68:2454–2468.
13. Hagerman PJ. Sequence-directed curvature of DNA. Annu Rev Biochem 1990;59:753–781.
14. Suzuki M, Yagi N. Stereochemical analysis of DNA bending by transcription factors. Nucleic Acids Res 1995;23:2083–2091.
15. Travers AA. DNA conformation and configuration in protein-DNA complexes. Curr Opin Struct Biol 1992;2:71–77.
16. Olson WK, Gorin AA, Lu X, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci USA 1998;95:11163–11168.
17. Matsumoto A, Olson WK. Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. Biophys J 2002;83:22–41.
18. Nadassy K, Wodak SJ, Janin J. Structural features of protein-nucleic acid recognition sites. Biochemistry 1999;38:1999–2017.
19. van Dijk M, Bonvin AMJJ. A protein-DNA docking benchmark. Nucleic Acids Res 2008;36:388.
20. Pingoud A, Jeltsch A. Structure and function of type II restriction endonucleases. Nucleic Acids Res 2001;29:3705–3727.
21. Santosy O, Joyce CM, Potapova O, Le Reste L, Kohlbheim J, Torella JP, Grindle ND, Kapanidis AN. Conformational transitions in DNA polymerase I revealed by single-molecule FRET. Proc Natl Acad Sci USA 2010;107:715–720.
22. Xu C, Maxwell BA, Brown JA, Zhang L, Suo Z. Global conformational dynamics of a Y-family DNA polymerase during catalysis. PLoS Biol 2009;7:e1000225.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990 215:403–410.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Mates JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Gen 2000;25:25–29.
26. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Strinivasan AR, Schneider B. The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 1992;63:751–759.
27. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
28. McLachlan AD. Rapid comparison of protein structures. Acta Cryst 1982;A38:871–873.
29. Bramucci E, Piaiadiini A, Bossa F, Pasarella S, PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. BMC Bioinformatics 2012;13:52.
30. Zemla A. LGa: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
31. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
32. Zhang Y, Skolnick J, Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
33. Venclovas Ce, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;53:585–595.
34. Moult J, Fidelis K, Krystalfoyalh A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction—Round VIII. Proteins 2009;77:1–4.
35. Kabash W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
36. Andrabi M, Mizuguchi K, Sari A, Ahmad S. Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. BMC Struct Biol 2009;9:30.
37. Ahmad S, Keskin O, Sari A, Nussinov R. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Res 2008;36:5922–5932.
38. Go N, Noguti T, Nishikawa T. Dynamics of a small globular proteins in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 2010;107:715–720.
39. Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: trypsin inhibitor, crambin, tibonuclease and lysozyme. J Mol Biol 1985;181:423–447.
40. Suhre K, Sanchez JA. EINemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 2004;32:W610–W614.
Conformational Changes in DNA-Binding Proteins

Ahmad S, Sarai A. Moment-based prediction of DNA-binding proteins. J Mol Biol 2004;341:65–71.

Ahmad S, Sarai A. Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. BMC Struct Biol 2011;11:8.

Bhardwaj N, Gerstein M. Relating protein conformational changes to packing efficiency and disorder. Protein Sci 2009;18:1230–1240.

Gutfreund MY, Schueler O, Maragalit H. Comprehensive analysis of hydrogen bonds in regulatory protein-DNA complexes: in search of common principles. J Mol Biol 1995;253:370–382.

Brooks B, Karplus M. Harmonic dynamics of proteins: normal mode and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 1983;80:6571–6575.

Levitt M, Sander C, Stern PS. Protein normal-mode dynamics: trypsin inhibitor, cbramin, ribonuclease and lysozyme. J Mol Biol 1985;181:423–447.

Durand P, Trinquier G, Sanejouand YH. A new approach for determining low-frequency normal modes in macromolecules. Biopolymers 1994;34:759.

Tama F, Sanejouand YH. Conformational change of protein arising from normal mode calculations. Protein Eng 2001;14:1–6.

Delarue M, Sanejouand YH. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. J Mol Biol 2002;320:1011–1024.

Kreib W, Alexandrov V, Wilson C, Echols N, Yu H, Gerstein M. Normal mode analysis of macromolecular motions in a database framework: developing model concentration as a useful classifying statistic. Proteins 2002;48:682–695.

Sanejouand YH. Normal mode analysis suggests important flexibility between the two N-terminal domains of CD4 and supports the hypothesis of a conformational change in CD4 upon HIV binding. Protein Eng 1996;9:671–677.

Prabakaran P, An J, Gromiha MM, Selvaraj S, Uedaira H, Kono H, Sarai A. Thermodynamic database for protein-nucleic acid interactions (ProNIT). Bioinformatics 2001;17:1027–1034.

Zen A, de Chiara C, Pastore A, Micheletti C. Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. Bioinformatics 2009;25:1876–1883.

Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 2005;6:33.