PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2016, **18**, 21618

Conformational selection underpins recognition of multiple DNA sequences by proteins and consequent functional actions†

Gitashri Naiya,‡[a] Paromita Raha,‡[a] Manas Kumar Mondal,[b] Uttam Pal,[a] Rajesh Saha,[a] Susobhan Chaudhuri,[c] Subrata Batabyal,[c] Samir Kumar Pal,[c] Dhananjay Bhattacharyya,[b] Nakul C. Maiti[a] and Siddhartha Roy*[d]

Recognition of multiple functional DNA sequences by a DNA-binding protein occurs widely in nature. The physico-chemical basis of this phenomenon is not well-understood. The *E. coli* gal repressor, a gene regulatory protein, binds two homologous but non-identical sixteen basepair sequences in the gal operon and interacts by protein–protein interaction to regulate gene expression. The two sites have nearly equal affinities for the Gal repressor. Spectroscopic studies of the Gal repressor bound to these two different DNA sequences detected significant conformational differences between them. Comprehensive single base-substitution and binding measurements were carried out on the two sequences to understand the nature of the two protein–DNA interfaces. Magnitudes of basepair–protein interaction energy show significant variation between homologous positions of the two DNA sequences. Magnitudes of variation are such that when summed over the whole sequence they largely cancel each other out, thus producing nearly equal net affinity. Modeling suggests significant alterations in the protein–DNA interface in the two complexes, which are consistent with conformational adaptation of the protein to different DNA sequences. The functional role of the two sequences was studied by substitution of one site by the other and *vice versa*. In both cases, substitution reduces repression *in vivo*. This suggests that naturally occurring DNA sequence variations play functional roles beyond merely acting as high-affinity anchoring points. We propose that two different pre-existing conformations in the conformational ensemble of the free protein are selected by two different DNA sequences for efficient sequence read-out and the conformational difference of the bound proteins leads to different functional roles.

Received 14th May 2016,
Accepted 28th June 2016

DOI: 10.1039/c6cp03278h

www.rsc.org/pccp

Introduction

Specific molecular recognition underlies much of biology. Recognition of specific DNA sequences, embedded among a large excess of non-specific sequences, by DNA-binding proteins provides the most striking example of a highly specific intracellular molecular recognition. DNA-binding proteins recognize their target sequences (often more than one), with high specificity and affinity. An example of such promiscuous recognition is binding of more than 500 functional sites in the human genome by transcription factor p53. Elk-1, an ETS family transcription factor, also recognizes more than 200 sites in the genome. In general, these target sequences have significant sequence differences.[1–4] How such promiscuous recognition is achieved by a single transcription factor still remains inadequately understood at the level of molecular structure and energetics. The structural basis of how a transcription factor may recognize multiple sequences was revealed from NMR structures of the lac repressor headpiece with its three natural operators, O1, O2 and O3.[5] The structures clearly demonstrated that binding to different operators having different DNA sequences leads to rearrangements of protein side-chains and alterations in the protein–DNA interfaces beyond the mutation sites, while preserving the overall geometry (a detailed analysis is presented in Table S1, ESI†). Another example of such a kind of change in the protein–DNA interface may be found in the co-crystal

[a] Division of Structural Biology and Bioinformatics, CSIR-Indian Institute of Chemical Biology, 4, Raja S. C. Mullick Road, Kolkata 700 032, India
[b] Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India
[c] Department of Chemical, Biological & Macromolecular Sciences, S. N. Bose National Centre for Basic Sciences, Block JD, Sector III, Salt Lake, Kolkata 700 098, India
[d] Department of Biophysics, Bose Institute, P1/12 CIT Scheme VII M, Kolkata, 700054, India. E-mail: sidroykolkata@gmail.com

† Electronic supplementary information (ESI) available: Five tables and four figures. See DOI: 10.1039/c6cp03278h

‡ Contributed equally.

structures of the p53 DNA-binding domain bound to different natural target sequences.[5–7]

The view that macromolecular interactions are exquisitely specific and exclusive—like a lock and a key—has been replaced by a more dynamic view in which plasticity plays the key role.[8,9] It is now believed in many quarters that in many ligand–protein interactions, conformational adaptation of both the interacting partners occurs.[10] Two competing mechanisms have been put forward: (1) the induced-fit and (2) the conformational selection.[11] In the former case, the protein binds to its target and remodels itself to produce the correct fit. In the latter case, the bound conformation pre-exists in the rapidly equilibrating ensemble of the free protein and is selected by the ligand. Increasingly, NMR studies are pointing towards the latter mechanism being prevalent in many ligand–protein interactions.[12] The lac repressor and its complexes with O1 and non-specific DNA have been studied by NMR. It was concluded that specific binding to the operator involves conformational selection.[13] However, how the protein adapts to different target sites, is not known.

In this article, we have explored the conformations of a DNA-binding protein bound to different DNA sequences. We show that selection of different protein conformations by different DNA sequences plays a key role in promiscuous recognition of DNA sequences by the DNA-binding protein.

Results and discussion

We have chosen the gal repressor (GalR) as a model system for this study. GalR binds to two different sequences, operators O_E and O_I, in the gal operon (*E. coli* genome).

Binding affinities of two operators, O_E and O_I, are very similar

Relatively short synthetic oligonucleotide duplexes containing the operator sequences and labeled at the 5′ end with a fluorescence probe were used for obtaining binding isotherms by fluorescence anisotropy. The DNA sequences of the two operators are shown in Fig. S1 (ESI†). The binding isotherms obtained from such titrations were used to extract equilibrium dissociation constants. The measured dissociation constants of the O_E–GalR and O_I–GalR complexes are 4.28 ± 2.13 nM and 4.07 ± 1.31 nM, respectively (Fig. 1). Similar dissociation constants had been observed before in footprinting experiments.[14] Thus, despite multiple differences in their sequences, the binding affinities of the protein for the two operators are very similar.

Conformations of two protein–DNA complexes are different

To understand the nature of multiple sequence recognition, we have explored the conformation of the GalR/O_E and GalR/O_I complexes. The different nature of the O_E and O_I complexes with GalR has been shown before.[15] It is more strongly underlined here by time-resolved fluorescence spectroscopy of the single tryptophan present in the C-terminal domain, distant from the DNA binding site. The fluorescence intensity decay (Fig. 2) is significantly different for the free, and the two operator-bound states.

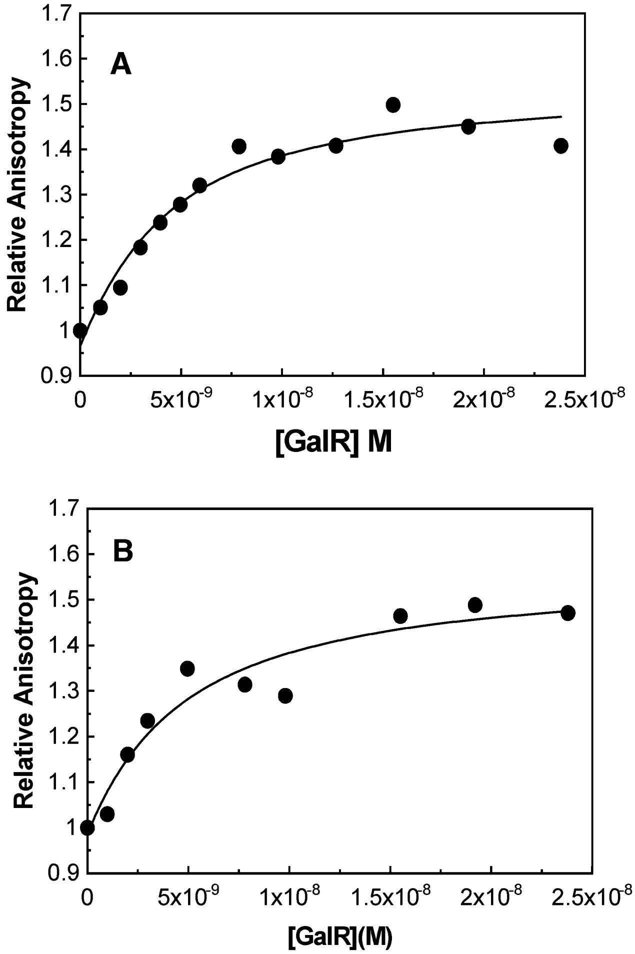

Fig. 1 Binding isotherms of (A) O_I and (B) O_E with GalR at 25 °C. Binding isotherms were determined by direct titration of fluorescein-end-labeled operators with unlabeled GalR. Measurements were carried out in 12.5 mM Tris-HCl buffer, pH 8.0 containing 300 mM KCl and 0.5 mM EDTA. Excitation and emission wavelengths used for each titration were 490 nm and 525 nm, respectively. The solid lines are best-fit lines to a single site binding equation.

The conformational difference has been studied in detail using different probes in a previous paper by our group.[16]

We have also measured DNA circular dichroism spectra of different complexes. DNA circular dichroism spectra are dominated by base-stacking geometries and are a good indicator of DNA conformational differences. The two complexes also show significant differences in DNA circular dichroism spectra reflecting conformational differences of the DNA in the two complexes (Fig. 3). Taken together, these results underline the difference in the conformation of the two complexes. Such differences have been noted in other systems as well.[17–19]

Single base-substitutions show differences in the interface

How the difference in the DNA sequence between the sites is transformed into different conformations is not well-understood. It is reasonable to expect that the nature of the protein–DNA interface plays a major role in transmitting the sequence information to the distant part of the protein, through which other

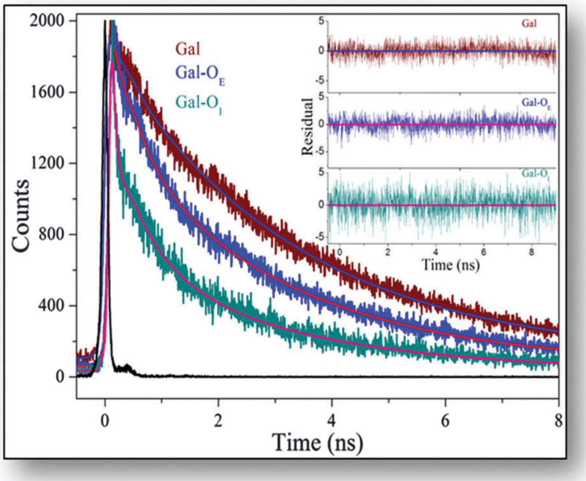

Fig. 2 Picosecond time-resolved fluorescence transients of the single tryptophan in Gal-repressor (t_{avg} = 2.4 ns), Gal repressor–O$_E$ (t_{avg} = 1.9 ns), Gal repressor–O$_I$ (t_{avg} = 0.7 ns) are shown. Excitation wave-length was at 287 nm. The decay was collected at 350 nm. The inset shows the corresponding residuals of the respective fits.

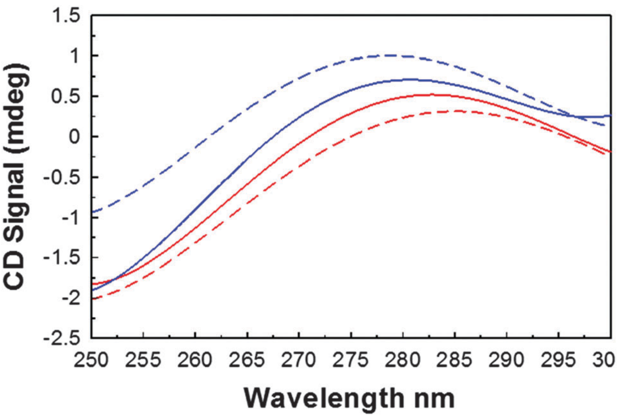

Fig. 3 Circular dichroism spectra of the operators in the presence of GalR (solid lines) and in its absence (broken lines). The red lines represent O$_E$, whereas the blue lines represent O$_I$. The experimental details are given in the Experimental section.

interactions often take place. Thus, we first explored the nature of the interfaces through exhaustive single basepair substitutions and equilibrium binding experiments.

Fluorescence anisotropy assay was used to accurately measure the loss of binding free energies in order to extract the energetic contribution of each basepair in the recognition process.[20] At first, the two operator sequences were aligned through the derivation of a consensus sequence based on the four eight-basepair half-sites of the two operators (Fig. S2, ESI†). The half-sites were chosen as templates because of the pseudo-symmetric nature of the DNA sequences. In the derived alignment, there are four basepairs that are different between the two operators. Fig. 4 shows the difference in free energy values ($\Delta\Delta G^\circ$) of all the substitution experiments performed on the target sequences, O$_E$ and O$_I$ (detailed in Tables S2 and S3, ESI†). It is clear that almost all the substitutions have a significant destabilizing effect on the respective complexes, but the magnitudes differ. The maximum variation in the binding energy is seen for position 7/7′, which differs significantly in all the four halves of the two sequences. An important aspect of these results is that substitutions in many of the homologous positions in O$_E$ and O$_I$ (even when the basepairs are identical in the two operators) have statistically significant differences in $\Delta\Delta G^\circ$ values (for example, positions 6, 8′, 6′, 4′ 3′, 2′). In some of these cases, the two flanking basepairs are also identical in the two operators. Thus, a global difference in the protein–DNA interface is indicated.

Base-specific interaction energy may be estimated from the $\Delta\Delta G^\circ$ of base-substitution

The free energy lost due to base-substitution is not a direct measure of the contribution of that basepair to the overall binding energy, as other perturbations may take place as a consequence of the substitution. The observed $\Delta\Delta G^\circ$ is a net value, which is the balance of both loss and gain of interactions. It is evident from the structures of lac repressor/O1 and lac repressor/O2 complexes that side-chains of the amino acid residues present in the proximity of the substituted site of the DNA may move to new spatial positions in response to binding to a variant DNA-sequence.[5] This may result in at least a partial compensation of the lost interaction energy due to the base substitution. To estimate the contribution of that particular base to the overall binding free energy, the experimentally observed loss of free energy upon base substitution ($\Delta\Delta G^\circ_{\text{exp}}$) was decomposed into three different components:

$$\Delta\Delta G^\circ_{\text{exp}} = \Delta\Delta G^\circ_{\text{loi}} + \Delta\Delta G^\circ_{\text{goi}} + \Delta\Delta G^\circ_{\text{pert}} \qquad (1)$$

or

$$\Delta\Delta G^\circ_{\text{loi}} = \Delta\Delta G^\circ_{\text{exp}} - \Delta\Delta G^\circ_{\text{goi}} - \Delta\Delta G^\circ_{\text{pert}} \qquad (2)$$

where $\Delta\Delta G^\circ_{\text{exp}}$ is the experimentally measured binding free energy difference between the wild-type and the mutant sequence; $\Delta\Delta G^\circ_{\text{loi}}$ (loi denotes the loss of interaction upon substitution of the new base) is the hypothetical loss of free energy, if no protein rearrangement has taken place; $\Delta\Delta G^\circ_{\text{goi}}$ (goi denotes the gain of interaction) is the free energy gain due to the formation of new interactions with the substituted basepair; $\Delta\Delta G^\circ_{\text{pert}}$ is the free energy lost or gained due to changes that occur in the other parts of the complex. Thus, $\Delta\Delta G^\circ_{\text{loi}}$ is taken as the base-specific interaction energy of that particular base.

For estimation of the base-specific interaction energy of a particular base, one has to estimate $\Delta\Delta G^\circ_{\text{loi}}$. As a first-order approximation, we assume that in comparison with the other two terms in eqn (1), $\Delta\Delta G^\circ_{\text{pert}}$ is small enough to be ignored for the purpose of estimation. To estimate $\Delta\Delta G^\circ_{\text{loi}}$, we carried out substitution of a basepair with the other three possible naturally occurring Watson–Crick basepairs. For the highest value of $\Delta\Delta G^\circ_{\text{exp}}$ among the three, the gain of interaction was assumed to be the least and negligible; thus, that particular $\Delta\Delta G^\circ_{\text{exp}}$ value was taken as an estimate of $\Delta\Delta G^\circ_{\text{loi}}$ for that basepair

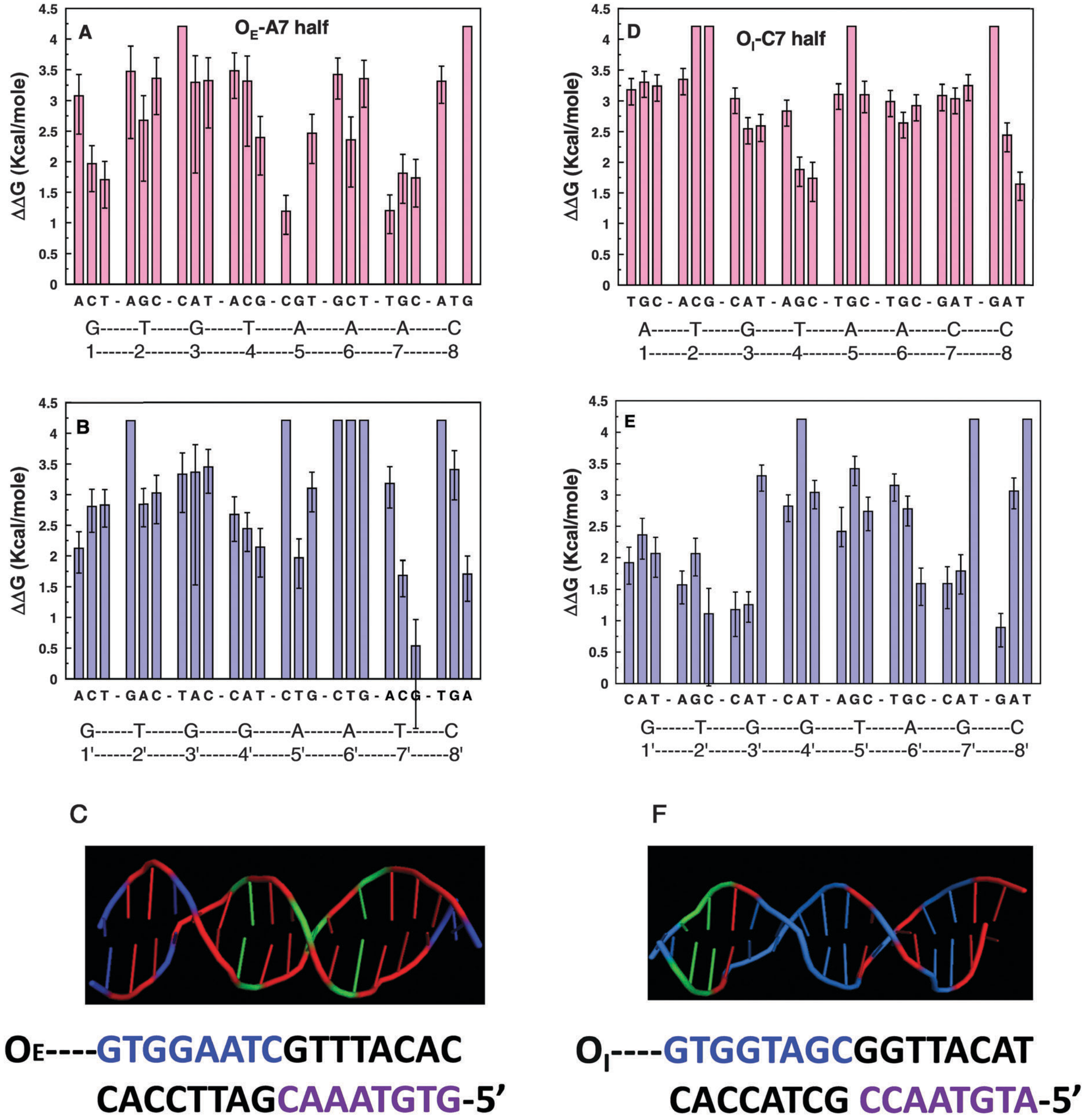

Fig. 4 Summary of free energy differences of single basepair mutants of (left panels) O_E from wild-type O_E and (right panels) O_I from wild-type O_I. (A) shows the $\Delta\Delta G^\circ$ values for the half that contains C at the 7th position. (B) shows the $\Delta\Delta G^\circ$ values for the other half. Numbered letters below indicate the actual sequence, whereas the smaller letters near the Y-axis represents the substituted base. (C) The figure shows the graphical representation of these data. The green color represents $\Delta\Delta G^\circ$ of less than 2 kcal mol^{-1},the blue color represents $\Delta\Delta G^\circ$ between 2–3 kcal mol^{-1} and the red color represents $\Delta\Delta G^\circ$ of greater than 4 kcal mol^{-1}. Details of experimental conditions are given in the Materials and methods section. The colors in the sequence denote two halves of each operator. (D) shows the $\Delta\Delta G^\circ$ values for the half that contains A at the 7th position. (E) shows the $\Delta\Delta G^\circ$ values for the other half. (F) The figure shows the graphical representation of these data as explained in (C).

and hence, the base-specific interaction energy. $\Delta\Delta G^\circ_{\rm loi}$ derived from base substitution experiments this way may be considered to be a good estimate for the specific binding energy contribution of that basepair to the overall binding free energy.

Base–protein interaction energy is modulated by structural perturbations around that basepair

In order to relate $\Delta\Delta G^\circ_{\rm loi}$ with structural changes, we have studied systems where both the high-resolution structure and the accurate base-substitution energies are known. Quantitative single basepair substitution effects have been measured in another prokaryotic system, the CI protein from bacteriophage λ and its target DNA binding site, O_R1.[21] O_R1 and its close homolog O_L1 (which differs by only one basepair), are 17 basepair pseudo-palindromic DNA sequences. There are six such binding sites, called operator-sites, in the λ genome and a consensus half-site can be derived from these sequences. In an operator-site, the half that resembles the consensus sequence more is termed the consensus-half and the other, the non-consensus-half.

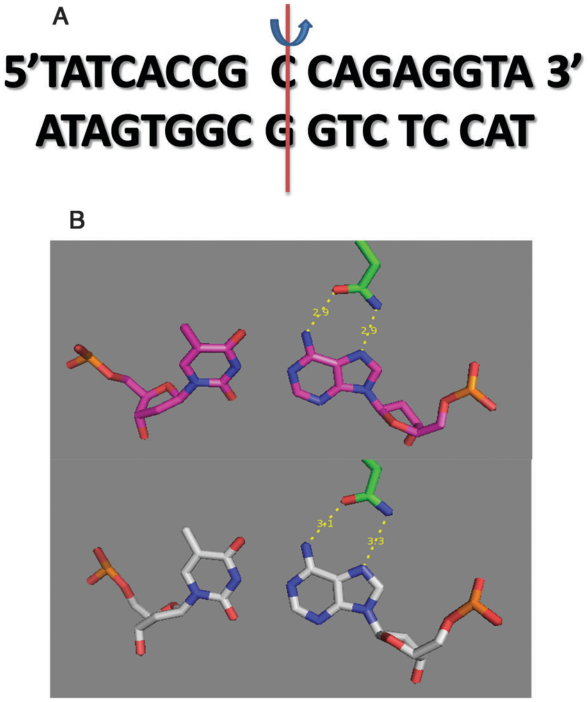

Fig. 5 (A) Sequence of O_R1 from bacteriophage λ. (B) Hydrogen bonding pattern of the AT basepair with the Gln44 side-chain in the consensus (upper structure) and non-consensus (lower structure) halves. The figure was generated by RASWIN, freeware from the Pdb structure 1LMB.

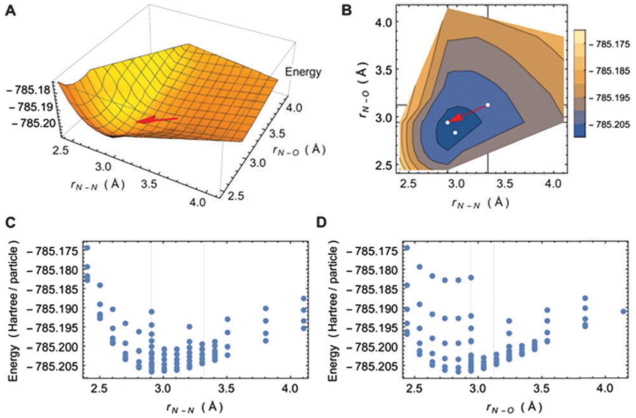

Fig. 6 Calculated potential energies for the H-bond formation between Gln44 and AT base pairs at two half sites. (A) The potential energy surface (PES) obtained by N–N and N–O distance perturbation. The red arrow indicates the energy gradient between the two crystal conformations. (B) The contour plot of the PES. The two connected white dots refer to the two crystal structure conformations. The other dot represents the optimum geometry. (C) Projection of the PES on the *xz* plane. The N–N distances in site I and site II are shown by the two vertical grid-lines. At site II, the distance corresponds to the equilibrium distance (r_e). (D) Projection of the PES on the *yz* plane. The N–O distances in site I and site II are shown by the two vertical grid-lines. At site II, the distance is more close to r_e.

The basepairs are numbered from 1–9 and then 8′–1′ starting from the consensus-half (Fig. 5A), with basepair 9 being the pseudo-symmetry axis. If one looks at the results of the single basepair substitution $\Delta\Delta G^\circ$s in this system, the highest magnitude of the $\Delta\Delta G^\circ$s (hence, the $\Delta\Delta G_{\mathrm{loi}}$) for symmetry-related AT basepairs 2 and 2′ differs by more than 100% (approximately 3 kcal mol^{-1} and 1.4 kcal mol^{-1}).[21] In the high-resolution structure of the N-terminal domain—λ-CI/O_L1 (1.8 Å resolution), these basepairs are brought into contact with the side-chain of Gln44.[22] In both halves, the side-chain amide group forms hydrogen bonds with the adenine at N7 and N6 positions (Fig. 5B). However, the distances between the N7 atom of the adenine and the nitrogen atom of the side-chain $CONH_2$ group of Gln44 are 2.9 Å and 3.1 Å in the consensus-half and the non-consensus-half, respectively. Similarly, the distances between the N6 atom of adenine and the oxygen atom of the side-chain $CONH_2$ group of Gln44 are 2.9 Å and 3.3 Å in the consensus-half and the non-consensus-half, respectively. When the hydrogen bond energy was calculated using the Hatree–Fock method,[23] a difference of about 5 kcal mol^{-1} was estimated for hydrogen bond energies in the two halves (Fig. 6). This weaker hydrogen bonding may account for the lower magnitude of $\Delta\Delta G^\circ_{\mathrm{loi}}$ of the base in the non-consensus-half. Thus, we may conclude that $\Delta\Delta G^\circ_{\mathrm{loi}}$ of a base may be a proxy for the strength of the base–protein interaction.

Modeled structures of O_E and O_I bound GalR show rearrangement of side-chains

The lac repressor forms high-affinity complexes with two operators O1 and O2. The binding affinities are very similar, just as in the case of gal repressor.[5,14] NMR structures of the lac repressor headpiece with O1 and O2 show readjustments of side-chains around the mutation sites, while retaining the overall similarity of the structures, suggesting that the modest side-chain rearrangement is the structural key to the recognition of different sequences.[5] Unfortunately, the structure of GalR as well as its operator complexes is not known. However, both NMR and X-ray structures of the closely related lac repressor and its complexes with multiple operators are known. We thus generated homology-modeled structures of GalR (1–61) with the two operators, O_E and O_I, followed by energy minimization and molecular dynamics refinement. Fig. 7 shows the two structures, which have a high degree of overall similarity as expected. However, there are significant differences in the protein–DNA interfaces. A detailed analysis of the differences is pointed out in the following paragraphs.

Environments of the three mutated residues were first compared (Fig. S3, S4 and S8, ESI†). In the O_E–TA 5′ basepair (basepair number), one of the Val-16 γ-methyl groups comes within the interacting distance of 6-NH_2 hydrogen atoms of A5′, with distances in the range of 2.6 Å. The methyl group of Ala-17 also comes within interacting distance of the methyl protons of T5′ with a distance of about 3 Å. In the O_I–AT 5′ basepair (changing from TA-to-AT) Val-16, methyl is about 5.6 Å away. However, the methyl group of Ala-17 stacks against the NH_2 group of A on the other strand at a distance of 2.7 Å (Fig. S3, ESI†). In the O_E–AT 7′ basepair, the residue that approaches the closest is Ala-56. The distance between the Ala methyl protons and A-2 and N3 is approximately 3.3 Å (Fig. S4, ESI†). In the O_I–CG 7′ basepair (AT to CG mutation), the Ala-56 methyl is stacked against G-2-NH_2 at a distance of 3.3 Å (Fig. S4, ESI†).

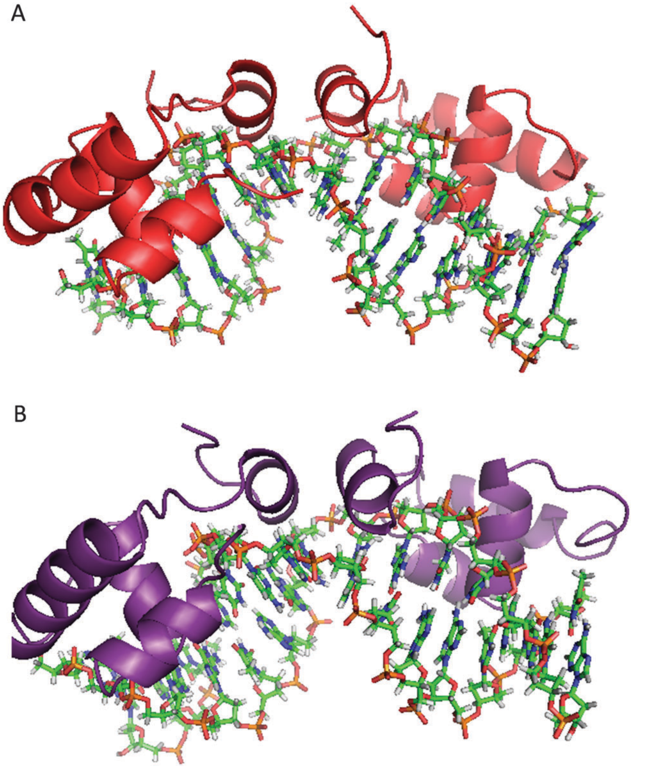

Fig. 7 (A) Homology modeled, molecular dynamics refined structure of O_E–GalR (1–62) complex. (B) Homology modeled, molecular dynamics refined structure of the O_I–GalR (1–62) complex.

In the symmetrically related O_E–AT 7 basepair, Ala-56 from the other subunit stacks against A2H and O4 of T at a distance of about 2.4 Å. For the O_I–CG 7 basepair (different from that of O_E), the Ala-56 methyl group stacks against G-2-NH_2 with a close approach of 2.4 Å (Fig. 8). Details are shown in Table S4 (ESI†). Thus, we may conclude that the change of basepairs has led to rearrangement of side-chains to accommodate these changes. It is evident that at least some of the lost energy due to base-substitution is recouped in the formation of new interactions in the substituted sequences, leading to partial compensation of the loss of free energy.

Basepair–protein interaction strengths are globally fine-tuned in the two complexes

Assuming that the highest among the three $\Delta\Delta G^\circ_{\text{exp}}$ values ($\Delta\Delta G^\circ_{\text{loi}}$) for a basepair is a proxy for the strength of that basepair–protein interaction, the value of $\Delta\Delta G^\circ_{\text{loi}}$ was used for the estimation of approximate strengths of individual basepair–protein interaction in the GalR/O_E and GalR/O_I complexes. Fig. 9 shows estimates of individual basepair–protein interaction strength (which is equal to $\Delta\Delta G^\circ_{\text{loi}}$) obtained with the method described above. When one compares the interaction strengths of basepairs across the two complexes for the same position, it becomes clear that the strength not only differs in the four positions that are different, but across the whole 16 basepair sequence. In some positions, interaction strengths are stronger in the O_E complex, whereas, in other positions they are weaker. The net balance favors O_E by a small amount (0.46 kcal mol^{-1}), indicating that the interactions fine-tune throughout the whole structure to generate an overall comparable affinity. If we focus on the structures of the protein in the two complexes, it will be clear that the protein conformations of the DNA-bound N-terminal domain are somewhat different in the two complexes, but closely related.

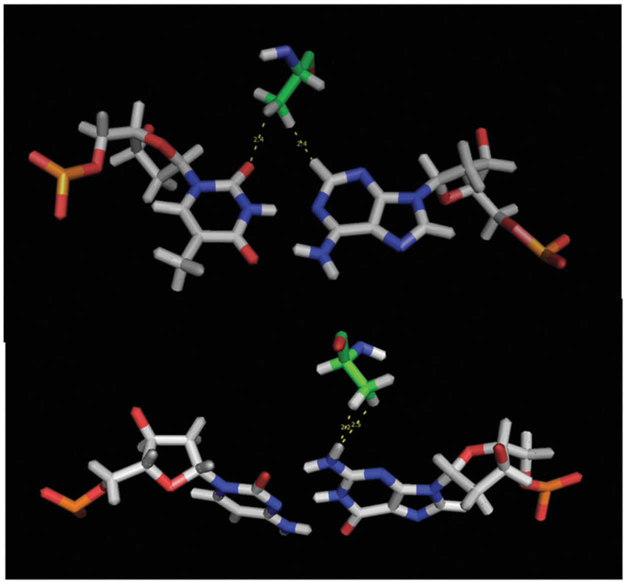

Fig. 8 Upper panel: Structure and interaction of basepair 7 of O_E. Lower panel: Structure and interaction of basepair 7 of O_I. The figures were generated using the freeware RASWIN.

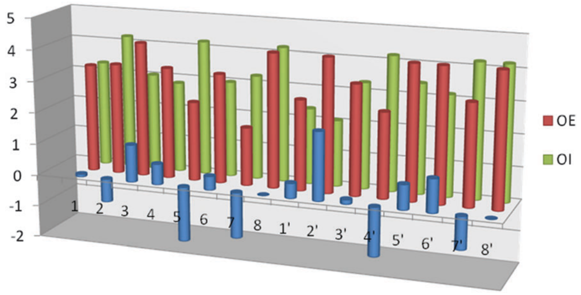

Fig. 9 Base specific interaction energies derived from $\Delta\Delta G^\circ_{\text{loi}}$. $\Delta\Delta G^\circ_{\text{loi}}$ was derived from base substitution data as described in the text. The blue bars represent the difference of $\Delta\Delta G^\circ_{\text{loi}}$ values for O_E and O_I ($\Delta\Delta G^\circ_{\text{loi}}/O_E - \Delta\Delta G^\circ_{\text{loi}}/O_I$). The basepair numbering protocol is as stated in the text.

Functional roles of different DNA sequences

Some studies suggest that a DNA-binding protein bound to different DNA sequences produces different DNA sequence-dependent functional outcomes. These sequences may induce a specific conformation in the protein that dictates that particular outcome.[24] Transcription factors, in general, have many binding sites in the genome and at least a significant fraction of these sites is likely to be functional. The sequences of functional sites often

Published on 29 June 2016. Downloaded on 14/08/2016 05:59:30.

vary significantly. However, in general, it is not known whether the actual sequence itself plays any functional role beyond just providing a binding site for the transcription factor. To understand the functional role of sequence variation in transcription regulation of the gal operon, a reporter plasmid was constructed in which the green fluorescent protein gene (GFP) was placed under the control of the Gal promoter (regulated by GalR). The gal promoter contains two binding sites for GalR, O_E and O_I, to which GalR binds, followed by GalR–GalR interaction and concomitant formation of a DNA loop.[25]

Formation of the loop is crucial for repression of one of the promoters, galP2, present.[26] Three plasmids were constructed, one with the wild-type O_E–O_I configuration and the other two in which the wild-type configuration was replaced with O_E–O_E and O_I–O_I (Table S5, ESI†). Table 1 shows the steady-state GFP expression levels of these plasmid-containing strains. Addition of the inducer, galactose, to strains bearing the plasmid having the O_E–O_I configuration enhances the fluorescence by 60–70% in 6 h (data not shown), indicating that the promoter on the plasmid is partially repressed in the absence of galactose (in the wild-type chromosomal gal promoter, higher levels of transcription induction is observed upon treatment with galactose). This is not unexpected as the multi-copy plasmid can only be partially occupied by low levels of indigenous GalR present in the strain.

In the same strain, both O_E–O_E and O_I–O_I plasmids have significantly higher fluorescence values in the absence of galactose, indicating that a significant fraction of repression in the O_E–O_I plasmid is relieved due to the substitution of one of the natural variants with the other sequence. These results clearly underline that in addition to the affinity, the sequence information is also important for the final outcome. Only a partial effect of substitution of one operator by the other on repression may be a reflection of the complexity of the gal operon regulation. The gal promoter is known to be regulated by the DNA loop formed by O_E and O_I bound GalR. The two promoters that are present in this region, galP1 and galP2, are differentially regulated by the loop formation.[14,26,27] When O_E is occupied alone, or the DNA loop is not formed, galP1 remains repressed to a far greater extent than galP2.[26] A possible explanation for the partial lifting of repression by substitution of O_I by O_E, or *vice versa*, is that both the sequence information are important for the loop formation and a proper regulation of galP2, thus affecting the overall repression partially. Thus, the conformational adaptation of GalR to different DNA sequences not only anchors the protein, but has been exploited by nature to fine-tune biological outcomes.

Molecular recognition is at the heart of every biological function. In protein–DNA interactions, conformational adaptation may be abundant and of central importance to biology.[28,29] We have demonstrated here that recognition of different sequences is a consequence of the alteration in side-chain positions in the complexes, reflecting conformational adaptation. As shown here and in other studies, recognition of different sequences, and consequent conformational adaptation involve conformational changes in the distant parts of the DNA-bound protein in addition to the protein–DNA interface as well.[17,24,30] These facts are compatible with a model that the gal repressor and perhaps other transcription factors exist as a rapidly inter-converting ensemble of multiple conformations from which different DNA sequences select different conformations. The existence of rapidly inter-converting conformations has been observed in the lac repressor.[13] A previous ultra-fast dynamical study of GalR also demonstrated significant ground-state heterogeneity, suggesting the existence of an ensemble of conformations.[16] We propose that different DNA sequences are recognized by different GalR conformers, resulting in freezing of one of the conformers in a particular protein–DNA complex. The altered conformations of the protein in different complexes may have been exploited by an organism for different gene regulatory outcomes (Fig. 10).

Table 1 GFP fluorescence with different O_E, O_I constructs

Construct	Relative fluorescence intensity (515 nm)
O_E–O_I (wild-type)	1
O_E–O_E	1.34
O_I–O_I	1.29

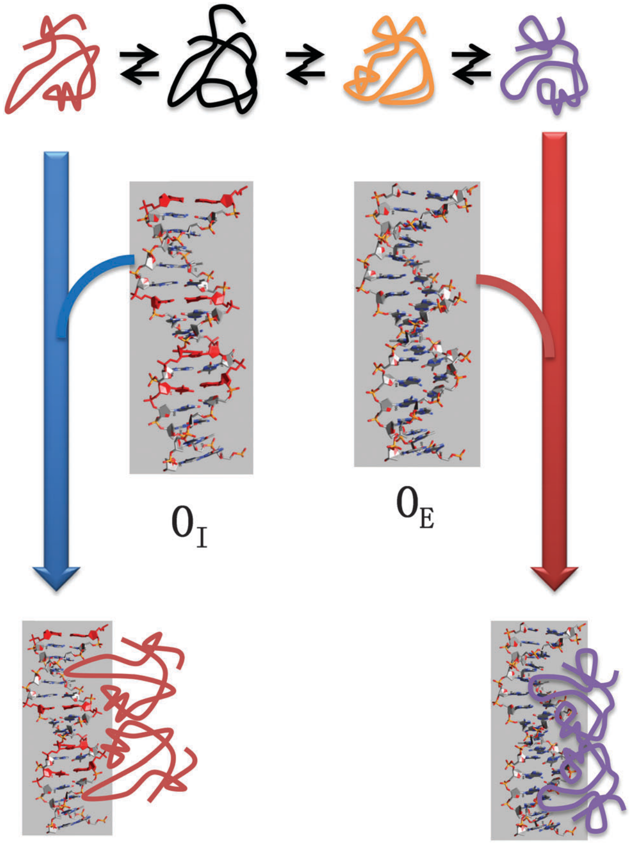

Fig. 10 A cartoon diagram of the conformational selection of different conformers of the DNA-binding proteins by different DNA sequences. The different colored wiggly chains are rapidly inter-converting different conformers of the DNA-binding protein. These conformations populate low-lying ground states. The two different DNA sequences, O_E and O_I are represented in stick form. The interface-distant changes in the bound conformations may provide interaction sites for other proteins, allowing different functional outcomes.

The differences in bound conformation of the protein between two operators may have implications for kinetics of the target search also. Recent work[31–33] in this area suggests that the presence of operators or operator-like sequences may have a significant effect on search kinetics. If the conformation of the protein depends on the bound DNA sequence, as was observed here, then the sliding process may involve switching of the protein between many conformations. How this will affect the search process is unclear without delving further into kinetic modeling.

Conclusions

DNA-binding proteins, including transcription factors, bind to many DNA sequences with variable affinity. Interactions with two classes of binding sites, specific and non-specific, have been studied in detail. The general consensus seems to be that the mode of interaction is different for these classes of binding sites. When bound to a specific site, the protein freezes into a particular conformation, usually forming hydrogen bonds and other weak interactions with DNA atoms, including those of the bases; thus establishing specificity for the target sequence. Binding to non-specific sites does not elicit freezing of conformation to the same extent and most of the interactions are with the sugar-phosphate backbone. However, little is known about the differences in the mode of recognition when the protein binds to two different specific sequences. In this article, we have demonstrated that when transcription factors recognize different DNA sequences, significant side-chain rearrangements are present in the protein–DNA interface, thus, creating a globally altered character of the interface. These rearrangements result in altered conformations of the protein as well. These changes in the interface regain some or all of the interaction energies lost due to changed bases present in different DNA sequences, thus producing comparable affinities. The dynamical character of the DNA-binding domain of this protein and its close homolog, the lac repressor, suggests that many conformations may pre-exist in the ground state. We suggest that the bound conformations are among those that pre-exist in the conformational ensemble of the free protein and are selected by different DNA sequences. The difference in bound conformations has been exploited by nature to create different functional outcomes.

Experimental methods

Materials

All the oligonucleotides used were either purchased from Trilink Biotechnologies (San Diego, CA, USA) or made in-house as described below. Sephadex G-25 was from Amer-sham Biosciences, GE Healthcare (Kolkata, India). Fluorescein-5-isothiocyanate (FITC) was purchased from Molecular Probes, Invitrogen (Bangalore, India). Ampicillin, IPTG and EDTA were purchased from Sigma-Aldrich Corporation (Bangalore, India). Bacto-tryptone, bacto-agar and yeast extract were purchased from Hi-Media Laboratories (Mumbai, India). Anhydrous glycerol was purchased from ACROS ORGANICS. Tris-HCl was purchased from Spectrochem (Mumbai, India). Ni-NTA agarose was purchased from Qiagen India (New Delhi, India). All other reagents used were of analytical grade.

Synthesis and purification of oligonucleotides

Oligonucleotides made in-house, were synthesized on an Applied Biosystems 3400 DNA Synthesizer. The oligonucleotides were then purified by reversed-phase HPLC (μ-BONDAPAK C-18 column) using a linear gradient of 100 mM tri-ethyl ammonium acetate, pH 7.0 in water to 100 mM tri-ethyl ammonium acetate, pH 7.0 in 100% acetonitrile in a WATERS HPLC instrument. This was followed by gel filtration of the respective oligonucleotides and the fractions containing the desired oligonucleotides were subjected to extensive dialysis.[34] The concentration of the oligonucleotides was calculated from A_{260} based on the base composition, using appropriate web servers.

Chemical modifications

All oligonucleotides used for end labeling were 5′-C6-amino linked. Oligonucleotides were labeled with FITC (dissolved in DMF) in a solution containing 1 M sodium carbonate/bicarbonate buffer, pH 9.0 : DMF : water in the ratio 5 : 2 : 3 as described before.[34] The sequence of the wild type operators used in our study was the 16 base gal operator core sequence, with GC overhangs added at both 5′ and 3′ ends. Single nucleotide substitutions were made at each of the 16 core positions of the operators; the base at each position was replaced with the 3 other Watson–Crick basepairs and the binding isotherm was determined using fluorescence anisotropy titration.

Purification of proteins

The *E. coli* strain DH5α bearing the plasmid pSEM1026, in the presence of the pBAD promoter, was grown in Luria broth containing 100 μg ml^{-1} of ampicillin at 37 °C. The plasmid was a generous gift from Dr Sankar Adhya's laboratory (NIH). After the optical density at 600 nm reached 0.4, the culture was induced with 0.2% arabinose for 5 hours. The cells were then harvested by centrifugation and re-suspended in 1/40 volume of lysis buffer I (50 mM potassium phosphate buffer pH 8.0 containing 0.5 mg ml^{-1} lysozyme) and stored on ice for 30 min. An equal volume of lysis II buffer (50 mM potassium phosphate buffer pH 8.0 containing 2 M KCl, 8 mM imidazole 20% glycerol and 1% Triton X-100) was added and incubated for 30 min on ice. The cell debris was removed by centrifugation at 10 000*g* for 1 h. The supernatant was added onto a Ni-NTA agarose column pre-equilibrated with lysis II buffer (excluding the detergent) and allowed to incubate for 1.5 to 2 h. Then 20 column volume of washing buffer (50 mM potassium phosphate pH 8.0 containing 600 mM KCl, 60 mM imidazole and 10% glycerol) was allowed to flow through the column. GalR was then eluted with 4 column volumes of elution buffer (50 mM potassium phosphate pH 8.0 containing 600 mM KCl, 250 mM imidazole and 10% glycerol). The eluted protein fractions were then pooled together and dialyzed against 50 mM potassium phosphate pH 8.0 containing 600 mM KCl, 30% glycerol and 0.5 mM EDTA. The purified protein was then divided into aliquots and stored at −80 °C.

Circular dichroism

Circular dichroism measurements for detection of DNA conformation were performed on a JASCO J850 spectropolarimeter using a one cm path length quartz cuvette. 2× stock solutions of the protein and the oligonucleotides were mixed in equal volumes by weighing in a microbalance to minimize pipetting errors. Separate oligonucleotide and protein solutions at the same concentrations were prepared by mixing equal volumes of buffer and previously prepared 2× solutions by weighing, as described above. The buffer-only spectrum was subtracted from the oligonucleotide spectra, and the protein-only spectrum was subtracted from the complex spectra. Spectral measurements of oligonucleotides and the oligonucleotide–GalR complexes were performed at oligonucleotide concentrations of 1 μM and GalR monomer concentrations of 4.0 μM, respectively. For GalR, the buffer used for CD measurements was 50 mM potassium phosphate, pH 8.0 containing 300 mM KCl, 0.5 mM EDTA and 5% glycerol. Protein and oligonucleotides were dialyzed overnight in the above-mentioned buffer prior to the CD measurements.

Fluorescence spectroscopy

All fluorescence studies were performed in either a PTI Quantimaster-6 T-geometry or Hitachi F3010 spectrofluorometer at 25 °C. The experiments were carried out either in a 1.0 cm or in a 0.5 cm path length quartz cuvette. For the gal repressor, binding of the wild type and different mutant operators was determined by direct titration of fluorescein-labeled operators with unlabeled GalR. All the anisotropy measurements were carried out in 12.5 mM Tris-HCl buffer, pH 8.0 containing 300 mM KCl and 0.5 mM EDTA at 25 °C. Excitation and emission wavelengths used for each titration were 490 nm and 525 nm, respectively, with a band pass of 5 nm in each channel. Dissociation constants were determined by fitting the anisotropy values to a single site binding equation.

Time-resolved fluorescence

All fluorescence decays were measured using a pico-second-resolved time-correlated single photon counting technique. A commercially available picosecond diode laser pumped time-resolved fluorescence spectrometer setup (Edinburgh Instrument, UK) was used. It has an instrument response function (IRF) of 50 ps. The pico-second excitation pulse from a Picoquant diode laser was used at 375 nm. A liquid scatterer was used to measure the FWHM of the IRF. Fluorescence from the sample was detected by a micro channel plate photo-multiplier tube (Hamamatsu) after dispersion through a grating monochromator.

Expression experiments. Table S5 (ESI†) contains the sequences of the regulatory region of the gal operon constructed in this study. The corresponding oligonucleotides were synthesized as described above. The above synthetic DNA fragments were cloned in the BamHI/XbaI cloning site of plasmid pSA11. In all constructs, the GFP gene was kept under the control of the cloned promoter regions. *E. coli* XL1B strain was used as the host. The plasmids containing the above genes were transformed in the XL1B strain and spread on a plate containing 2.5% Luria broth, 1.5% agar and 100 μg ml^{-1} ampicillin and incubated at 37 °C. A single colony of cells harboring each plasmid was then inoculated in 50 ml of LB medium and allowed to grow overnight at 37 °C. Then the concentration of cells was normalized to an A_{600} value of 2. The fluorescence of the cells was then measured using a PTI fluorometer. The excitation wavelength was 488 nm. The emission was scanned from 500 nm to 600 nm.

MD simulation and structural analysis

We have performed molecular dynamics (MD) simulation of the operator bound state using CHARMM[35] and considering an all *trans* conformation as the initial guess. The initial set up and minimization of the system was done using the AMBER-14[36] suite of programs with the ff14SB force field.[37] The system was solvated in an orthorhombic water box containing TIP3P[38] water molecules in such a manner that there was at least a 15 Å thick layer of water around the solute in all three directions. A required number of sodium ions was added to maintain the electro neutrality of the system. The system was then energy minimized for 20 000 cycles using a combination of steepest descent and conjugate gradient algorithms and applying periodic boundary conditions. Long-range electrostatic energy was calculated using the particle mesh Ewald summation[39,40] with 1 Å grid spacing and a 10^{-6} convergence criterion. Lennard-Jones and short-range electrostatic interactions were truncated at 10 Å. The MD simulation has been carried out using NAMD software[41] with the energy minimized structure. Initially heating to 300 K was carried out slowly over 30 ps. With 1 fs time steps. We continued the simulation for the system up to the desired time at constant temperature (300 K) and pressure (1 bar) using the Langevin–Piston algorithm.[42] Translational and rotational movements of the centre of mass were removed at an interval of 5 ps. SHAKE constraints were applied to all bonds involving hydrogen atoms and a 1 fs time step for the MD integrator. The extended trajectory was constructed by saving conformations after every 1 ps for further analysis. Root mean square displacement (RMSD) and root mean square fluctuation (RMSF) analyses of the DNA and protein sub-units were performed using CHARMM.[35] The RMSD values indicate that the structures equilibrate within 5 ns of MD run. The RMSF values also indicate reasonable fluctuations of the loop residues and the rather rigid nature of the secondary structural motifs. The study of hydrogen bonds between the protein side chain and operator DNA is done by using modified pyrHBfind[43] software, considering a H-bond distance cut off 3.5 Å and an angle greater than 120°.

Homology modeling

Homology modeled structures of the Gal repressor with the two operators, O_E and O_I were generated by mutating the lac repressor/O1 structure in Chimera;[44] followed by executing the above protocol and force field for energy minimization and 50 ns MD simulation refinement using NAMD.

Density functional theory analysis

The crystal structure of the lambda repressor–operator complex was obtained from the Protein Data Bank (PDB ID: 1LMB). Coordinates of the heavy atoms of Gln44 hydrogen bonded with the AT base pair were obtained from this crystal structure for the two half sites (site I and II). Hydrogen atoms were added to the complex in GaussView 5.0 followed by optimization in Gaussian 09. The heavy atoms were frozen during the optimization with the B3LYP/6-31G+(d,p) level of theory. The potential energy surface (PES) for the H-bond formation was scanned by perturbing the donor–acceptor distances (N–N and N–O). Leaving the scan coordinates, the geometry of the complex was optimized at each point with the B3LYP/6-31G+(d,p) level of theory. Very small step sizes (0.1 Å) were used near the equilibrium distances (r_e) and a 0.3 Å step size was used away from r_e. Donor–acceptor distances (r_{N-N} and r_{N-O}) were plotted against the energies to obtain the two/three dimensional potential energy diagrams.

Acknowledgements

We thank Prof. Gautam Basu for useful discussion. This work was supported by the Council of Scientific and Industrial Research, India and a JC Bose Fellowship, Department of Science and Technology, India to SR.

Notes and references

1 J. Boros, I. J. Donaldson, A. O'Donnell, Z. A. Odrowaz, L. Zeef, M. Lupien, C. A. Meyer, X. S. Liu, M. Brown and A. D. Sharrocks, *Genome Res.*, 2009, **19**, 1963–1973.

2 C.-L. Wei, Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, T. Zhang, A. Shahab, H. C. Yong, Y. Fu and Z. Weng, *Cell*, 2006, **124**, 207–219.

3 R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig and R. S. Mann, *Annu. Rev. Biochem.*, 2010, **79**, 233.

4 R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann and B. Honig, *Nature*, 2009, **461**, 1248–1253.

5 J. Romanuka, G. E. Folkers, N. Biris, E. Tishchenko, H. Wienk, A. M. Bonvin, R. Kaptein and R. Boelens, *J. Mol. Biol.*, 2009, **390**, 478–489.

6 M. Kitayner, H. Rozenberg, N. Kessler, D. Rabinovich, L. Shaulov, T. E. Haran and Z. Shakked, *Mol. Cell*, 2006, **22**, 741–753.

7 M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig and Z. Shakked, *Nat. Struct. Mol. Biol.*, 2010, **17**, 423–429.

8 G. Schreiber and A. E. Keating, *Curr. Opin. Struct. Biol.*, 2011, **21**, 50–61.

9 A. P. Yamniuk and H. J. Vogel, *Mol. Biotechnol.*, 2004, **27**, 33–57.

10 D. D. Boehr, R. Nussinov and P. E. Wright, *Nat. Chem. Biol.*, 2009, **5**, 789–796.

11 G. Wei, W. Xi, R. Nussinov and B. Ma, *Chem. Rev.*, 2016, **116**, 6516–6551.

12 P. Csermely, R. Palotai and R. Nussinov, *Trends Biochem. Sci.*, 2010, **35**, 539–546.

13 C. G. Kalodimos, N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens and R. Kaptein, *Science*, 2004, **305**, 386–389.

14 M. Geanacopoulos and S. Adhya, *J. Bacteriol.*, 1997, **179**, 228–234.

15 S. Chatterjee, Y.-N. Zhou, S. Roy and S. Adhya, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 2957–2962.

16 S. Choudhury, G. Naiya, P. Singh, P. Lemmens, S. Roy and S. K. Pal, *ChemBioChem*, 2016, **17**, 605–613.

17 S. Deb, S. Bandyopadhyay and S. Roy, *Biochemistry*, 2000, **39**, 3377–3383.

18 J. M. Petersen, J. J. Skalicky, L. W. Donaldson, L. P. McIntosh, T. Alber and B. J. Graves, *Science*, 1995, **269**, 1866–1869.

19 S. M. Holmbeck, H. J. Dyson and P. E. Wright, *J. Mol. Biol.*, 1998, **284**, 533–539.

20 T. Heyduk, Y. Ma, H. Tang and R. H. Ebright, *Methods Enzymol.*, 1996, **274**, 492.

21 A. Sarai and Y. Takeda, *Proc. Natl. Acad. Sci. U. S. A.*, 1989, **86**, 6513–6517.

22 L. J. Beamer and C. O. Pabo, *J. Mol. Biol.*, 1992, **227**, 177–196.

23 S. J. Grabowski, W. A. Sokalski, E. Dyguda and J. Leszczynski, *J. Phys. Chem. B*, 2006, **110**, 6444–6446.

24 S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen and K. R. Yamamoto, *Science*, 2009, **324**, 407–410.

25 S. Adhya, *Annu. Rev. Genet.*, 1989, **23**, 227–250.

26 H. E. Choy and S. Adhya, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**, 11264–11268.

27 T. Aki, H. E. Choy and S. Adhya, *Genes Cells*, 1996, **1**, 179–188.

28 J. L. Vaughn, V. A. Feher, C. Bracken and J. Cavanagh, *J. Mol. Biol.*, 2001, **305**, 429–439.

29 Y. Pan, C.-J. Tsai, B. Ma and R. Nussinov, *Trends Genet.*, 2010, **26**, 75–83.

30 S. Tan and T. J. Richmond, *Cell*, 1990, **62**, 367–377.

31 D. Gomez and S. Klumpp, *Phys. Chem. Chem. Phys.*, 2016, **18**, 11184–11192.

32 A. B. Kolomeisky, *Phys. Chem. Chem. Phys.*, 2011, **13**, 2088–2095.

33 M. Bauer, E. S. Rasmussen, M. A. Lomholt and R. Metzler, *Sci. Rep.*, 2015, **5**, DOI: 10.1038/srep10072.

34 S. Debnath, N. S. Roy, I. Bera, N. Ghoshal and S. Roy, *Nucleic Acids Res.*, 2013, **41**, 366–377.

35 B. R. Brooks, C. L. Brooks, A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels and S. Boresch, *J. Comput. Chem.*, 2009, **30**, 1545–1614.

36 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, *Comput. Phys. Commun.*, 1995, **91**, 1–41.

37 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, 2006, **65**, 712–725.

38 W. L. Jorgensen and J. D. Madura, *J. Am. Chem. Soc.*, 1983, **105**, 1407–1413.

39 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

40 D. M. York, T. A. Darden and L. G. Pedersen, *J. Chem. Phys.*, 1993, **99**, 8345–8348.

41 L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, *J. Comput. Phys.*, 1999, **151**, 283–312.

42 S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, 1995, **103**, 4613–4621.

43 S. Mukherjee, S. Majumdar and D. Bhattacharyya, *J. Phys. Chem. B*, 2005, **109**, 10484–10492.

44 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.