# Weak Operator Binding Enhances Simulated Lac Repressor-mediated DNA Looping

Andrew V. Colasanti, Michael A. Grosner, Pamela J. Perez, Nicolas Clauvelin, Xiang-Jun Lu,[*] Wilma K. Olson[†]

*Department of Chemistry & Chemical Biology, BioMaPS Institute for Quantitative Biology, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA*

[*] *Current address*: Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[†] *Correspondence to*: Wilma K. Olson; email wilma.olson@rutgers.edu

**This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/bip.22336**

**ABSTRACT**

The 50th anniversary of *Biopolymers* coincides closely with the like celebration of the discovery of the *Escherichia coli* (*lac*) lactose operon, a classic genetic system long used to illustrate the influence of biomolecular structure on function. The looping of DNA induced by the binding of the Lac repressor protein to sequentially distant operator sites on DNA continues to serve as a paradigm for understanding long-range genomic communication. Advances in analyses of DNA structures and in incorporation of proteins in computer simulations of DNA looping allow us to address long-standing questions about the role of protein-mediated DNA loop formation in transcriptional control. Here we report insights gained from studies of the sequence-dependent contributions of the natural *lac* operators to Lac repressor-mediated DNA looping. Novel superposition of the ensembles of protein-bound operator structures derived from NMR measurements reveals variations in DNA folding missed in conventional structural alignments. The changes in folding affect the predicted ease with which the repressor induces loop formation and the ways that DNA closes between the protein headpieces. The peeling of the auxiliary operators away from the repressor enhances the formation of loops with the 92-bp wildtype spacing and hints of a structural reason behind their weak binding.

## INTRODUCTION

The *Escherichia coli* Lac repressor protein suppresses the formation of enzymes involved in the metabolism of lactose by simultaneously binding to two sequentially distant binding sites, or operators, on a DNA molecule and forcing the intervening residues into a loop.[1] In addition to the principal operator site $O_1$ found within the region of the genome that promotes the expression of the DNA coding for β-galactosidase, there are two weaker auxiliary binding sites, $O_3$ and $O_2$, located respectively 92 base pairs (bp) upstream and 401 bp downstream of the primary operator.[2] The DNA loops formed upon binding the repressor to the primary operator and to one of the auxiliary operators are required for maximal repression of transcription,[3] and, in the case of the smaller $O_3 \cdots O_1$ loop, impede access of the transcriptional machinery to the promoter sites.[4] Formation of the larger $O_1 \cdots O_2$ loop precludes complete formation, *i.e.*, synthesis, of the RNA that codes for β-galactosidase.

The looping of DNA induced by the binding of the Lac repressor to the $O_3$ and $O_1$ operators occurs on a length scale shorter than the natural scale of DNA deformation. The distortions of local double-helical structure found at ambient temperatures — bends of 5-6° between adjacent base pairs, 4-5° fluctuations in twist, and 0.2-0.6-Å displacements of successive base pairs[5] — lead to spatial arrangements more extended on average than the spatial pathways needed to form the $O_3 \cdots O_1$ loop. The natural occurrence of such a small DNA loop has stimulated interest in factors that may enhance its formation, including large sequence-dependent and protein-induced distortions of the double-helical structure.[6-8] The DNA comprising the $O_1 \cdots O_2$ loop, by contrast, lies in the range of chain lengths predicted to close most readily into circular or looped forms.[9-11]

A number of research groups have established sensitive *in vivo* assays of the looping of DNA mediated by the Lac repressor protein, basing the ease of loop formation on the expression of reporter genes controlled by the repressor.[7,12,13] The measured levels of gene expression depend, among other things, upon the distance between *lac* operator sites. The variation in gene-product levels with chain length exhibits a complex oscillatory pattern, with extremes in production occurring every 10-12 bp, in approximate phase with the DNA helical repeat. The expression levels also depend upon operator identity. For example, the chain-length-dependent patterns of repression are several base pairs out of phase on gene constructs flanked at the 5′-end by an ideal, fully symmetric, high-affinity operator sequence called $O_{sym}$ and at the 3′-end by $O_1$ or $O_2$.[12,13] That is, the precise shapes of the plots of gene expression versus operator spacing, including the exact positions of peaks and troughs in the repression

profiles, differ when $O_1$ is substituted by $O_2$. The repression levels also vary when the natural *lac* operators replace the auxiliary $O_{sym}$ operator in a 92-bp $O_{sym}\cdots O_1$ construct.[12]

NMR solution studies of the *lac* operators bound to the N-terminal headpieces of the repressor protein point to subtle differences in molecular structure at the binding sites.[14,15] Whereas the reported complexes containing $O_1$, $O_2$, and $O_{sym}$ assume similar spatial arrangements with comparable numbers of close protein-DNA contacts, consistent with their similar binding affinities, there is a significant loss of intermolecular contacts in the complexes containing the more weakly bound $O_3$ operator. The loss of contacts perturbs the three-dimensional states adopted by the $O_3$-containing structure compared to those observed in the $O_1$, $O_2$, and $O_{sym}$ complexes. The 5′-half of $O_3$, which bears close sequence similarity to the stronger operators, makes more contacts with the protein headpieces than the 3′-half of $O_3$. One of the two identical proteins that comprise the DNA recognition element seemingly penetrates more deeply into the major groove than its partner.

The precise pathways of the operators determine the ways in which the intervening DNA fits between the two halves of the repressor. Previous computational studies of Lac repressor-mediated DNA looping[8,16,17] have assumed that the natural operators adopt identical, rigid, symmetric structural folds on the protein assembly similar to those reported[18] in the low-resolution crystal structure of the tetramer with two bound operators. The three-dimensional arrangement of the full repressor-operator complex must be inferred from the overlap of corresponding protein atoms in the high-resolution crystal structures of the dimer binding $O_{sym}$ and the tetramer without DNA binding headpieces.[18,19] The two halves of the assembled structure form a V and make contact with DNA at the ends of the V. The asymmetric, sequence-dependent DNA pathways extracted from NMR studies of the natural operators with the Lac repressor headpieces could have profound effects on the configurations and supercoiled states of the intervening loops. The arrangement of operator DNA on the full tetrameric assembly determines the positions and orientations of the DNA at the two ends of the tethered loops. These anchoring conditions, in turn, dictate the preferred pathways of the loops.[20-22]

Here we investigate how changes in operator sequence and anchoring conditions influence the ease of DNA loop formation and possibly contribute to known effects of operator identity on *lac* gene repression. We construct models of the operator-bound Lac tetramer that take account of the sequence-dependent structures of the bound DNA and the features of protein shared among known high-resolution structures. We take advantage of new technology that makes it possible to examine ensembles of structures from a common perspective. We also consider the various orientations in which the operator sequences may bind to the repressor complex. We use the ends of the many reported protein-bound operator fragments as anchoring points in

Monte-Carlo simulations of the likelihood of DNA loop formation and a new optimization technique to identify the energetically preferred spatial pathways. We focus on loops with the wildtype 92- and 401-bp spacing and examine how substitutions of different operators and variations in the operator pathways, suggested by the NMR solution structures, affect the ease of loop formation and the preferred configurations of the simulated loops.

For simplicity, we ignore the large-scale opening of the Lac repressor detected in low-resolution structural studies,[6,23-26] *i.e.*, we treat the protein is a rigid scaffold. Opening of the tetramer along the lines suggested by the experiments alters the simulated configurations of short loops of certain chain lengths but does not significantly affect the predicted ease of DNA loop formation.[8,17] We also omit consideration of the effects on looping of non-specific architectural proteins, such as HU, which occur in abundance in *Escherichia coli* and distort the structure of DNA.[27-29] Random binding of HU at the stoichiometric levels observed *in vivo* increases both the variety of simulated looped structures and the computed likelihood of loop formation.[8,17] We further ignore known sequence-dependent structural features of the DNA loops, such as the intrinsic deformability of the promoter elements.[7] Omission of the latter contributions allows us to focus on the roles of the sequence-dependent operator structures in DNA loop formation.

We find wide variability in overall operator structure missed in conventional alignments of related structures based on root□mean□square fits of corresponding atoms. We also find that some of the operator pathways substantially enhance and others appreciably suppress the simulated formation of Lac repressor-mediated DNA loops. The changes in operator structure on the protein headpieces change the pathway of the intervening DNA loop and hint of a potential structural role of the weaker auxiliary operators in Lac repressor-mediated DNA looping.

## MATERIALS AND METHODS

### Repressor-operator Models

Three-dimensional models of the Lac repressor bearing natural, asymmetric operators on the two dimeric halves of the protein were generated in two stages. First, selected amino-acid atoms from the crystal complex of the dimer with the $O_1$ operator (Protein Data Bank[30] (PDB) entry 1jwl[31]) were superimposed on the corresponding atoms in the V-shaped model of the Lac repressor-operator assembly previously constructed from high-resolution crystal structures.[16] The superimposed atoms, which lie in the N subdomain of the dimer (residues 63-334 of chains A and B), overlap closely in the complexes of the dimer with $O_1$ and $O_{sym}$ (0.38-Å root-mean-

square deviation of protein backbone atoms in the specified residues).[31] Alignment of these protein atoms makes it possible to take account of the spatial displacement of the natural operator and its binding headpieces with respect to the symmetric operator and its headpieces. The different lengths of the two operators preclude direct superposition of DNA. The central G·C base pair of the natural operator is offset from the CG base-pair step at the center of the symmetric operator.

The superposition of the $O_1$-dimer complex also takes account of apparent errors in the positions of atoms in the non-template DNA so that complementary bases associate via Watson-Crick pairing. A 180°-rotation of the thymine in residue 10 of chain E, from the conformationally unlikely *syn* glycosyl conformation to the *anti* form, converts the noncanonical A·T pair in the deposited structure to the desired Watson-Crick arrangement. Given the uncertainty in the orientation of the $O_1$-bound dimer on the tetramer, the complex is placed in two directions on each arm of the V-shaped complex, with the 5′-3′ direction of the operator pointing toward either the inside or the outside of the protein structure.

The second stage of model building entailed superposition of the structures of the natural operators deduced from NMR measurements on the DNA pathways found in the solid state. We take advantage of a new feature in the 3DNA suite of programs that allows a user to look at multiple structures from a common perspective.[32] The software aligns the individual DNA pathways in the multi-model NMR structure files on a common reference frame. Here we position the G·C base pair at the centers of the three natural operators on the corresponding base pair in the $O_1$-containing crystal structure and the highly kinked CG base-pair step at the centers of the symmetric operator models on the corresponding base pair in the $O_{sym}$ crystal complex (PDB entry 1efa[18]).

We characterize the various operator-repressor constructs in terms of the bending and twisting of the bound DNA and the six rigid-body parameters between base pairs at the ends of the two binding sites.[8] The latter parameters also describe the arrangement of the first and last base pairs of the intervening DNA loop. The degree of bending of the bound operators is taken as the angle between the helical axes of the DNA at the two ends of each protein-DNA structure. The directions of the axes are computed with the 3DNA software[33,34] over the three terminal base-pair steps. We take advantage of another new feature of 3DNA that facilitates the analysis of ensembles of related structures in collecting the components of the local helical axes.[32] The twisting of DNA is based on a discrete ribbon constructed from the origins and reference frames of successive base pairs.[35] In contrast to the twist angle included in the six rigid-body parameters specifying the relative spatial arrangements of successive base pairs,[36] the twist reported here, the so-called twist of supercoiling, can be combined with the writhing number of a

closed structure to obtain the correct linking number.[35,37]

**Looping Simulations**

The various combinations of operator alignments and models lead to a variety of starting points for simulations of DNA looping. For example, the 10 structures of the $O_3$ operator consistent with NMR distance constraints and the 20 such structures of the $O_1$ operator lead to 800 possible ways to anchor the ends of $O_3 \cdots O_1$ loops. The cited number reflects the $10 \times 20$ possible arrangements of bound operator structures in each of four possible orientations on the Lac repressor. As noted above, in the absence of knowledge of the directions in which the natural operators attach to the dimeric halves of the tetramer, we place each operator in two orientations on the protein-binding headpieces. The operators thus run in roughly parallel directions in two of the constructs and in approximately antiparallel directions in the other two.[16,38] The combinations of known operator structures and orientations similarly produce 1600 sets of looping constraints for $O_1 \cdots O_2$ loops and 1936 constraints for $O_{sym} \cdots O_{sym}$ loops. The latter figure also takes account of the asymmetry of the $O_{sym}$-protein complexes, *i.e.*, the identity of the leading sequence-bearing strand is not known.

For simplicity, we delete extraneous base pairs at the ends of the NMR structures, namely the four non-consensus base pairs at the termini of the 23-bp constructs containing the natural operators and the four corresponding pairs in the 22-bp constructs containing the symmetric operator. Loop length is measured in the conventional sense in terms of the center-to-center base-pair spacing $N$ between bound operators. The loops anchored to the natural operator pathways thus include 16 protein-bound base pairs and $N-16$ unbound base pairs, *e.g.*, 76 bp in loops with the 92-bp wildtype spacing. Loops of the same total length anchored to the symmetric operators include two more base pairs, *e.g.*, 78 bp of protein-free DNA in loops of 92 bp.

We treat the double helix at the level of base-pair steps, using six rigid-body parameters to specify the arrangements of successive base pairs and a potential that allows for elastic deformations of the long, thin molecule from the canonical B-DNA structure.[11] The pathway of protein-free DNA is constructed, one base-pair step at a time, from randomly sampled sets of rigid-body parameters subject to the elastic potential. The allowed fluctuations are compatible with the known bending and twisting of mixed-sequence DNA.[11] The DNA attached to the Lac repressor is described in terms of the rigid-body parameters of the DNA base pairs found in the different operator-bound complexes. The repressor is included as a 'side group' of the DNA, *i.e.*, the atoms of protein are expressed in the reference frame of one of the bases in the molecular complexes.

The ease of DNA loop formation is estimated from the fraction of simulated configurations of a linear molecule with terminal base pairs positioned so as to overlap the base pairs at the ends of the operators attached to the Lac repressor. The formation of a successfully closed loop is detected by adding a virtual base pair to the 3´-end of the DNA and checking its coincidence with the first base pair of the chain.[8,11] The probability of DNA looping is reported in terms of the Jacobson-Stockmayer *J* factor, the well-known ratio of the equilibrium constant for polymer ring closure compared to the bimolecular association of a linear molecule of the same length and composition.[39] The greater the value of *J* is, the lower the free energy of DNA is and the greater the likelihood of looping.

We incorporate fluctuations in operator structure indirectly by considering the effects of the associated sets of anchoring conditions on loop formation. That is, for each combination of operator structures and loop orientations we find the proportion of simulated loops with terminal base pairs arranged so as to coincide with the base pairs at the appropriate edges of the bound operators. The range of rigid-body parameters between the ends of the DNA brackets the conformational freedom of both the bound operators and the looped DNA. The detailed treatment of protein and DNA structure makes it possible to visualize the spatial arrangements of looped DNA and to gain insight into the possible effects of operator sequence and structure on gene organization and expression.

Samples of $10^{17}$ random chains, collected in 5-8 hours on a high performance computer cluster for each of ~1000 combinations of loop anchoring conditions, typically yield about 500 possible arrangements of the less easily formed loops (*J* factors of ~$10^{-14}$) and over 500,000 examples of the more readily closed structures (*J* factors of ~$10^{-11}$). The successfully generated configurations are subsequently used as starting points in a new energy-optimization procedure (described below). The combination of approaches is valuable in identifying competing configurational states, *i.e.*, allowing for so-called basin hopping on the configurational energy landscape, and in obtaining a reliable estimate of DNA loop stability. Direct minimization from an arbitrary DNA configuration, while also possible, may introduce serious problems related to the stability of the solution and also lead to unphysical states, *e.g.*, self-intersecting spatial pathways, during the course of minimization. We draw the reader's attention to published examples of Monte-Carlo generated Lac repressor-mediated DNA loops that are illustrative of the starting states used in the loop optimization procedure.[8] The changes in three-dimensional structure effected by minimization are described in the Results and Discussion.

**Loop Optimization**

Minimum-energy looped structures are obtained with a procedure that makes it possible to optimize the potential energy of elastic deformation of a collection of base pairs, in which the positions and orientations of the first and last pairs are held fixed. That is, we solve a constrained optimization problem where the potential energy of elastic deformation is the objective function. Our approach differs from earlier DNA loop-optimization schemes, which are based primarily on the equations of motion of a collection of base pairs and entail direct determination of the forces and moments that are needed to place the end base pairs in the desired arrangement.[10,16,40-43] Here we employ an alternate representation of the geometry of a collection of base pairs that makes it possible to take the spatial constraints into direct account, *i.e.*, the constraints can be solved analytically. In other words, we simplify a constrained optimization problem into an unconstrained one. Moreover, the minimization of the potential energy of elastic deformation always yields the optimally stable solution, another methodological difference from earlier approaches for which the stability of solutions must be computed separately. External forces and torques applied to the base pairs can be accounted for through the addition of appropriate potential energy functions, *i.e.*, expressions of the work produced by the external loadings.

We consider a collection of $N$ base pairs with imposed positions and orientations of the first and last base pairs, *i.e.*, we specify the three components of the end-to-end vector and the three rotational degrees of freedom that describe the relative orientation of coordinate frames on the first and last base pairs. We replace the set of base-pair step parameters, a $6(N–1)$-dimensional vector, with an equivalent set of independent variables referred to as the step degrees of freedom, also a $6(N–1)$-dimensional vector. The dependence of these degrees of freedom on the traditional base-pair step parameters can be obtained analytically. In addition, the imposed end-to-end vector and end-to-end rotational constraints can be expressed such that the step degrees of freedom for the last base-pair step are functions of the degrees of freedom of all the other steps. In other words, we reduce the dimensionality of the problem by taking into account the boundary conditions (the dimension is reduced from $6(N–1)$ to $6(N–2)$). The gradient of the potential energy of elastic deformation for the collection of base pairs as well as the Hessian matrix are also obtained analytically. The minimization procedure is implemented as a gradient-based optimization, *e.g.*, conjugate gradient or BFGS (Broyden-Fletcher-Goldfarb-Shanno) methods, in the reduced space of the step degrees of freedom. Full details will be presented elsewhere.

## RESULTS AND DISCUSSION

### Sequence-dependent Deformations of DNA Operators

The reorientation of Lac repressor-bound operator structures in a common reference frame reveals features of overall DNA folding hidden in the coordinate files stored in the Protein Data Bank. The conventional alignment of related structures on the basis of a root-mean-square fit of corresponding atoms produces a series of similar spatial pathways that roughly superimpose upon one another.[32] The superposition of the same models in a common nucleotide coordinate frame reveals sizable distortions of global structure, in which the DNA molecules flex to different degrees and in different directions (Figure 1). Here the natural operators are superimposed such that the G·C base pair at the center of each 23-bp NMR-based model[15] coincides with the G·C base pair at the center of the $O_1$ operator crystallized in the presence of the Lac dimer.[31] The 22-bp NMR models of the symmetric operator[14] are aligned somewhat differently on the crystallized $O_{sym}$-dimer complex,[19] with the coordinate frame of the central CG·CG base-pair step used as the reference for superpositioning. Only the protein headpieces of the crystalline complexes are depicted in the images since the variation of the protein backbone in the superimposed models is relatively small. The average distances between the $C^{\alpha}$ atoms of the protein fragments in the NMR models and those in the related X-ray structure (residues 4-58 of individually superimposed strands A and B) are respectively 0.9, 1.2, 1.1, and 1.5 Å in the $O_1$, $O_2$, $O_3$, and $O_{sym}$ complexes. By contrast, the corresponding displacements of the centers of the central 15 bp of the natural operators are 3.4, 4.9, and 5.5 Å and those of the central 14-bp of the symmetric operator are 2.0 Å. The differences in the DNA pathways are even greater if expressed in terms of backbone atoms.

The reorientation of the NMR models with respect to the crystalline DNA pathways, the changes in global bending with operator sequence, and the variation in apparent DNA deformability stand out in the similarly oriented molecular images. Interestingly, the degree of protein-induced DNA bending decreases and the variability among double-helical models increases with loss of operator strength. The DNA structures fitted to the NMR distances associated with the weak and notably less curved $O_3$ operator vary especially widely at the 3′-end of the sequence where the differences in chemical identity with $O_1$ and $O_2$ are greatest. A few of the $O_3$ models stand out as highly deformed but others more closely match the pathway of the primary operator, *e.g.*, the centers of the 3′-terminal base pairs in models 4 and 9 of $O_3$ deviate from the X-ray reference structure by 9 Å or more but those in model 2 deviate by 5 Å or less. Although not as widely dispersed, the $O_2$ operator structures also show enhanced variability over the $O_1$ operator at the 3′-end of the recognition sequence.

The molecular images also hint of the changes in twist that accompany the deformations of repressor-bound DNA. Compared to the $O_1$ operator, the $O_2$ and $O_3$ structures are undertwisted, particularly at their 3′-ends (Figure 2). The net bending of the operator structures ranges from an average value of $35_{\pm 16}$° over the central 15 bp of the $O_3$ models to mean values of $50_{\pm 6}$° and $70_{\pm 6}$° over the corresponding stretches of $O_2$ and $O_1$. The bending over the central 14 bp of the synthetic $O_{sym}$ operator models is $41_{\pm 10}$°. The differences in overall spatial configuration fall within the bounds of certainty of the experimentally measured interactions, *i.e*., the changes in DNA operator structure are sufficiently large to be distinguished by the set of interatomic distances measured in the NMR experiments.

**Spatial Constraints on DNA Looping**

The variation in local operator structure introduces even greater variability in the spatial constraints placed on the ends of the DNA loops formed upon Lac repressor binding. The control that the repressor has on the intervening DNA manifests itself in the way in which it positions the bound operators. Small changes in the orientations of the operators can potentially lead to large changes in the overall shape of the loop itself.[20-22]

The superimposed images of 15-bp $O_3$ and $O_1$ operators in Figure 3 illustrate the relatively loose hold of the Lac repressor on $O_3 \cdots O_1$ loops. Here we draw attention to the variability in the spatial disposition of the operators in the four possible orientations of the DNA on the protein assembly. We use the letters P and A to designate operators that run in approximately parallel and antiparallel directions and the numerals 1 and 2 to distinguish whether the operator at the 5′-end of the loop, *i.e*., $O_3$ on the left half of the complex, points respectively toward the inside or outside of the tetramer. The distances between the ends of the $O_3$ and $O_1$ operators that would abut the intervening DNA loop (blue and red space-filled atomic representations in the figure) span slightly broader ranges than the distances between the corresponding ends of two bound $O_{sym}$ operators (respective A1, A2, P1, P2 end-to-end separations of 67-78, 83-90, 50-60,112-119 Å for the depicted $O_3 \cdots O_1$ termini compared to values of 73-80, 74-80, 48-57, 108-113 Å for $O_{sym} \cdots O_{sym}$ termini). Moreover, the peeling of the 3′-end of $O_3$ away from protein leads to an asymmetry in the displacement and orientation of the loop-anchoring base pairs not present in the DNA attached to $O_{sym}$ operators. Whereas the average angles between the planes of the bases at the ends of $O_{sym} \cdots O_{sym}$ loops are fairly similar in A1 and A2 orientations ($160_{\pm 5}$° and $159_{\pm 5}$°, respectively), those at the ends of $O_3 \cdots O_1$ loops differ appreciably ($159_{\pm 10}$° and $140_{\pm 8}$°, respectively). The changes in $O_3$ geometry also increase the angles between the anchoring base pairs in loops attached to the repressor in parallel orientations (respective values of $59_{\pm 11}$° and $18_{\pm 9}$° for P1 and P2 arrangements of $O_3$ and $O_1$ compared to values of $45_{\pm 6}$° and $12_{\pm 5}$° for the corresponding configurations of $O_{sym}$ operators).

The presence of the $O_2$ recognition elements on the complementary strand of the *Escherichia coli* genome changes the asymmetry of the looping constraints. The peeling of DNA away from the repressor at the 5′-end of the leading sequence alters the relative angles between the base pairs at the ends of antiparallel $O_1 \cdots O_2$ loops compared to those at the ends of $O_3 \cdots O_1$ loops (respective A1, A2 values of $133_{\pm 5}°$ and $167_{\pm 5}°$ between the $O_1$ and $O_2$ terminal base pairs). The angles between the base pairs at the ends of parallel $O_1 \cdots O_2$ loops also exceed those at the ends of the corresponding $O_3 \cdots O_1$ loops (respective P1, P2 values of $72_{\pm 7}°$ and $21_{\pm 5}°$ between the ends of $O_1$ and $O_2$). The distance constraints placed on the terminal base pairs of the $O_1 \cdots O_2$ loops (respective A1, A2, P1, P2 end-to-end distances of 77-86, 78-85, 46-56, 117-122 Å) are slightly more stringent than those placed on $O_3 \cdots O_1$ loops but less restrictive than those placed on $O_{sym} \cdots O_{sym}$ loops.

**Operator-induced Changes in Loop Formation**

The changes in the pathways of the natural *lac* operators have striking effects on the predicted ease of Lac repressor-mediated DNA loop formation (Figure 4). Although the computed *J* factors of the DNA loops anchored to most of the NMR-based operator models roughly match the values previously reported for 92-bp loops formed between identical rigid operator sites,[8] some of the operator pathways substantially enhance and others appreciably suppress the looping propensities. For example, the added deformability of $O_{sym}$ operators in the NMR models increases the likelihood of closing 92-bp loops in A1 orientations by up to an order of magnitude over that determined with static anchoring conditions and decreases the chances of closing DNA in a P1 orientation by over an order of magnitude on average and over two orders of magnitude for certain combinations of operator models (compared the color-coded images with the previously reported *J* factors cited in the figure legend). The looping probabilities are sensitive to the orientation of the anchoring base pairs, with higher *J* factors found when the planes of terminal base pairs describe strong ‘bend’ angles, *i.e.*, values less than the 180° angle formed between precisely oriented antiparallel base pairs and greater than the 0° angle formed between perfectly aligned parallel base pairs. That is, the ends of the more easily formed loops (attached to the model combinations denoted by yellow and orange as opposed to blue and green colors) are less perfectly aligned than the ends of the less readily closed structures.

The effect of anchoring conditions on loop formation is even more striking for 92-bp ‘wildtype’ loops formed between $O_3$ and $O_1$ operators. The *J* factors of $O_3 \cdots O_1$ loops attached to the Lac repressor in a P1 orientation increase by over two orders of magnitude as the angle between the anchoring base pairs varies from approximately 40° to 90°. Moreover, the ease of looping approaches that found for DNA closed between antiparallel operators as the bend angle reaches the upper limit. The variation in the *J* factor with base-pair orientation is somewhat less

pronounced for loops anchored by operators in antiparallel orientations, increasing by roughly 60-fold over an approximately 40° reduction in the bend angle between terminal base pairs in A1 orientations and decreasing by about 30-fold over a similar change in the bend angle in A2 orientations. As evident from the plaid-like patterns in the *J* factor mosaics (Figure 4), the enhancement in looping is tied to specific operator pathways. The peeling of DNA away from the repressor, such as found with $O_3$ model 4, enhances the computed looping propensities and, as described below, alters the pathway of the intervening DNA.

The likelihood of generating 401-bp 'wildtype' $O_1 \cdots O_2$ loops is substantially greater than forming 92-bp loops anchored to the same operator pathways (Figure 5). The enhancement in looping follows from the known ring-closure properties of DNA, which adopts cyclic forms with greatest ease when two to three times the persistence length.[9,44,45] The increase in chain length raises the *J* factors of looping by four or more orders of magnitude compared to the values found for the shorter chain fragment. Moreover, unlike their 92-bp counterparts, the *J* factors of the 401-bp loops are roughly independent of anchoring conditions with values spanning similar narrow ranges. The differences among the various operator structures are minor compared to the spacing between $O_1$ and $O_2$, *i.e.*, the through-space separation distances, equivalent to a 20-30-bp stretch of double-helical DNA, are much smaller than the chain contour length. Unlike the shorter loops, which close somewhat more readily when the operators are placed in one of the antiparallel orientations, the longer loops attach to the repressor with nearly equivalent ease in all four orientations. The slightly lesser likelihood of forming 401-bp loops between operators bound in a P2 orientation follows from the near impossibility of forming shorter loops of the same type. The few successfully closed 92-bp loops of this type, which are not included in Figure 5, adopt tight, semicircular pathways much higher in energy than the DNA rest state. Interestingly, the dip in the *J* factor seen in a subset of the 92-bp $O_1 \cdots O_2$ loops anchored in A2 orientations is confined to operator models where $O_2$ comes in closest contact with the repressor.

Representative structures of 92-bp energy-minimized $O_3 \cdots O_1$ loops (Figure 6) reveal some of the subtle changes in DNA folding that accompany the variations in operator structure and looping propensities. Here we present four representative repressor-bound loops derived from states captured in Monte Carlo sampling, two with operators in antiparallel orientations and two with operators in parallel orientations. The more easily formed loops, in which $O_3$ peels away from the repressor, follow somewhat different pathways from the less readily closed loops, where the operator remains in tighter contact with protein. Configurations in the former category are less strongly deformed and thus of lower energy than those in the latter category. In particular, the DNA in the vicinity of $O_3$ tends to straighten in the low-energy structures and the

loops appear to open compared to their high-energy counterparts. The lower energy antiparallel loop makes a more gradual U-turn than the higher energy form, and the parallel loop 'relaxes' to the lower energy configuration through an approximately 90º rotation about the pseudodyad axis of the repressor, *i.e*., the axis that bisects the V-shaped protein assembly through its base. The parallel structures resemble conformational states that might be encountered in the flipping of a loop from a crossed to an uncrossed configuration. The root-mean-square deviation between corresponding base-pair centers on the two depicted antiparallel loops is 16 Å and that on the parallel loops is 22 Å. The optimization reduces the energy of the illustrated states by factors of 6-9, from values in the range of 2-3 *kT* per base pair to values of a few tenths *kT*. The change in energy is not surprising given that the initial Monte Carlo configurations are highly improbable states with expectedly large deformations. The minimization procedure translates the configuration as a whole while smoothing the spatial pathway. Interestingly, the global displacement between initial and final states is somewhat greater for the two antiparallel loops than for the parallel loops (respective root-mean-square deviations between base-pair centers of 23-29 Å and 13-14 Å). The corresponding deviations in the rigid-body parameters between successive base pairs are similar in these examples, *i.e*., angular deviations of 4.1-4.7° and translational differences of 0.14 Å.

**Concluding Remarks**

Classical equilibrium expressions describe the binding of ligands to protein in terms of the relative amounts of ligand in the bound and unbound state. From this perspective the weak auxiliary $O_3$ *lac* operator spends more time off the Lac repressor than its $O_1$ and $O_2$ counterparts. Here we find that this weakness in binding facilitates *in silico* formation of short Lac repressor-mediated DNA loops with the wildtype 92-bp spacing. The subtle, sequence-dependent changes in repressor-operator interactions, found to account for the relative affinities of the natural *lac* operators for the Lac repressor,[15] lead to deformations in DNA double-helical structure that affect the ease with which a short chain segment can be looped between the two halves of a rigid protein scaffold. The broader range of states available to the weak $O_3$ operator provides more 'landing sites' for successful loop formation. By contrast, the more highly curved and closely convergent set of structures adopted by the stronger $O_2$ operator result in lesser chances of forming loops of the same length. The relative ease of $O_3$ dissociation coupled with the apparent enhancement of $O_3 \cdots O_1$ loop formation could potentially facilitate control of gene expression, *e.g*., allow for the ready uptake of competing proteins on DNA and efficient loop closure.

One must be careful in extrapolating the predicted effects of operator sequence on the formation of DNA loops *in silico* to effects detected *in vivo*. The *J* factors reported here for 92-bp

are orders of magnitude smaller than the DNA looping propensities deduced from gene expression studies. The computed values, nevertheless, fall within the same range of values as the *J* factors determined in recent tethered particle motion studies of 92-bp loops anchored to the Lac repressor by $O_{sym}$ and $O_1$ operators.[7] As we have reported recently,[8] various cellular factors can account for the unexpectedly high looping propensities deduced from *in vivo* measurements. For example, addition of the architectural protein HU significantly enhances the formation of short Lac repressor-mediated loops. The repressor may also undergo different levels of deformation at different chain lengths and enhance loop formation. Thus, many molecular features beyond DNA chain length and operator identity must be considered in the design of gene-expression experiments that test the ideas proposed here.

Finally, given the uncertainty in the structures of DNA employed in the present simulations and the approximate treatment of protein, a more thorough investigation of the Lac repressor-operator energy landscape is needed to confirm our ideas. Small deformations in protein structure, such as the twisting of the DNA-binding headpieces suggested by atomic-level simulations and the 'kinking' of the N- and C-terminal subdomains associated with the uptake of the IPTG inducer molecule,[46,47] are likely to modulate the simulated looping. The peeling of DNA operators, in combination with likely fluctuations of the protein complex, may allow for the long-anticipated looping of sufficiently long DNA around the Lac repressor assembly[48] and may help to rationalize the opposing chain-length-dependent patterns of Lac repressor-induced repression detected on promoters flanked at their 3′- and 5′-ends by different operators.[7,12,13]

**ACKNOWLEDGEMENTS**

The U.S. Public Health Service under research grants GM34809 and GM096889 has generously supported this work. MAG and PJP gratefully acknowledge fellowship support from the US Department of Education Graduate Assistance in Areas of National Need program.

**REFERENCES**

1. Krämer, H.; Niemöller, M.; Amouyal, M.; Revêt, B.; von Wilcken-Bergmann, B.; Müller-Hill, B. EMBO J 1987, 6, 1481-1491.
2. Reznikoff, W. S.; Winter, R. B.; Hurley, C. K. Proc Natl Acad Sci, USA 1974, 71, 2314-2318.
3. Oehler, S.; Elisabeth R, E.; Krämer, H.; Müller-Hill, B. EMBO J 1990, 9, 973-979.
4. Becker, N. A.; Peters, J. P.; Lionberger, T. A.; Maher III, L. J. Nucleic Acids Res 2013, 41, 155-166.
5. Olson, W. K.; Gorin, A. A.; Lu, X.-J.; Hock, L. M.; Zhurkin, V. B. Proc Natl Acad Sci, USA 1998, 95, 11163-11168.
6. Mehta, R. A.; Kahn, J. D. J Mol Biol 1999, 294, 67-77.
7. Johnson, S.; Lindé, M.; Phillips, R. Nucleic Acids Res 2012, 40, 7728–7738.
8. Czapla, L.; Grosner, M. A.; Swigon, D.; Olson, W. K. PLoS ONE 2013, 8, e56548.
9. Shimada, J.; Yamakawa, H. Macromolecules 1984, 17, 689.

10. Zhang, Y. L.; Crothers, D. M. Biophys J 2003, 84, 136-153.
11. Czapla, L.; Swigon, D.; Olson, W. K. J Chem Theor Comp 2006, 2, 685-695.
12. Müller, J.; Oehler, S.; Müller-Hill, B. J Mol Biol 1996, 257, 21-29.
13. Becker, N. A.; Kahn, J. D.; Maher 3rd, L. J. J Mol Biol 2005, 349, 716-730.
14. Spronk, C. A.; Bonvin, A. M.; Radha, P. K.; Melacini, G.; Boelens, R.; Kaptein, R. Structure 1999, 7, 1483-1492.
15. Romanuka, J.; Folkers, G. E.; Biris, N.; Tishchenko, E.; Wienk, H.; Bonvin, A. M. J. J.; Kaptein, R.; Boelens, R. J Mol Biol 2009, 390, 478-489.
16. Swigon, D.; Coleman, B. D.; Olson, W. K. Proc Natl Acad Sci, USA 2006, 103, 9879-9884.
17. Olson, W. K.; Grosner, M. A.; Czapla, L.; Swigon, D. Biochem Soc Trans 2013, 41, 559–564.
18. Lewis, M.; Chang, G.; Horton, N. C.; Kercher, M. A.; Pace, H. C.; Schumacher, M. A.; Brennan, R. G.; Lu, P. Science 1996, 271, 1247-1254.
19. Bell, C. E.; Lewis, M. Nature Structural Biology 2000, 7, 209-214.
20. Tobias, I.; Coleman, B.; Olson, W. K. J Chem Phys 1994, 101, 10990-10996.
21. Zhang, P.; Tobias, I.; Olson, W. K. J Mol Biol 1994, 242, 271-290.
22. Westcott, T. P.; Tobias, I.; Olson, W. K. J Phys Chem 1995, 99, 17926-17935.
23. Friedman, A. M.; Fischmann, T. O.; Steitz, T. A. Science 1995, 268, 1721-1727.
24. Ruben, G. C.; Roos, T. B. Microscopy Research and Technique 1997, 36, 400-416.
25. Taraban, M.; Zhan, H.; Whitten, A. E.; Langley, D. B.; Matthews, K. S.; Swint-Kruse, L.; Trewhella, J. J Mol Biol 2008, 376, 466-481.
26. Virnik, K.; Lyubchenko, Y. L.; Karymov, M. A.; Dahlgren, P.; Tolstorukov, M. Y.; Semsey, S.; Zhurkin, V. B.; Adhya, S. J Mol Biol 2003, 334, 53-63.
27. Ali Azam, T.; Iwata, A.; Nishimura, A.; Ueda, S.; Ishihama, A. J Bacteriol 1999, 181, 6361-6370.
28. Swinger, K. K.; Lemberg, K. M.; Zhang, Y.; Rice, P. A. EMBO J 2003, 22, 3749-3760.
29. Sagi, D.; Friedman, N.; Vorgias, C.; Oppenheim, A. B.; Stavans, J. J Mol Biol 2004, 341, 419-428.
30. Berman, H. M.; Westbrook, J.; Feng, Z., Gilliland; G., B., T.N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235-242.
31. Bell, C. E.; Lewis, M. J Mol Biol 2001, 312, 921-926.
32. Colasanti, A. V.; Lu, X. J.; Olson, W. K. J Vis Exp 2013, (74), e4401.
33. Lu, X. J.; Olson, W. K. Nucleic Acids Res 2003, 31, 5108-5121.
34. Lu, X.-J.; Olson, W. K. Nature Protoc 2008, 3, 1213-1227.
35. Clauvelin, N.; Tobias, I.; Olson, W. K. J Chem Theor Comp 2012, 8, 1092-1107.
36. Dickerson, R. E.; Bansal, M.; Calladine, C. R.; Diekmann, S.; Hunter, W. N.; Kennard, O.; von Kitzing, E.; Lavery, R.; Nelson, H. C. M.; Olson, W. K.; Saenger, W.; Shakked, Z.; Sklenar, H.; Soumpasis, D. M.; Tung, C.-S.; Wang, A. H.-J.; Zhurkin, V. B. J Mol Biol 1989, 208, 787-791.
37. Britton, L.; Olson, W. K.; Tobias, I. J Chem Phys 2009, 131, 245101.
38. Geanacopoulos, M.; Vasmatzis, G.; Zhurkin, V. B.; Adhya, S. Nature Structural Biology 2001, 8, 432-436.
39. Jacobson, H.; Stockmayer, W. H. J Chem Phys 1950, 18, 1600-1606.
40. Coleman, B. D.; Olson, W. K.; Swigon, D. J Chem Phys 2003, 118, 7127-7140.
41. Biton, Y.; Coleman, B.; Swigon, D. J Elasticity 2007, 87, 187-210.
42. Becker, N. B.; Everaers, R. Structure 2009, 17, 579-589.
43. Koslover, E. F.; Fuller, C. J.; Straight, A. F.; Spakowitz, A. J. Biophys J 2010, 99, 3941-3950.
44. Kratky, O.; Porod, G. Rec Trav Chim Pays-Bas 1949, 68, 1106-1122.
45. Livshits, M. A. Mol Biol 1996, 30, 85-90.
46. Villa, E.; Balaeff, A.; Schulten, K. Proc Natl Acad Sci, USA 2005, 102, 6783-6788.
47. Dabera, R.; Stayrooka, S.; Rosenberga, A.; Lewis, M. J Mol Biol 2007, 370, 609-619.

48. Tsodikov, O. V.; Saecker, R. M.; Melcher, S. E.; Levandoski, M. M.; Frank, D. E.; Capp, M. W.; Record Jr., M. T. J Mol Biol 1999, 294, 639-655.

## LEGENDS TO FIGURES

Figure 1. Cartoon images of NMR-derived Lac repressor-operator models[14,15] superimposed on the central base-pair elements of related crystal complexes[19,31] (see text). Images rendered in PyMOL (www.pymol.org) with DNA backbones shown as gold tubes, DNA bases as gold sticks, and protein backbones (chains A and B) as pink and cyan ribbons. Note the wide structural differences among the natural $O_1$, $O_2$, and $O_3$ operators when placed in a common reference frame on the central (deep blue) G·C base pair and the similar spatial pathways of the artificial symmetric operator $O_{sym}$ (centered on its central CG·CG base-pair step) and $O_1$.[14,15] DNA is depicted with the sequence-bearing (recognition) strand running from the lower left to upper right and making contacts at its 5′-end with chain A and its 3′-end with chain B. All views looking down the long axis of the kinked CG·CG step at the centers of the reference crystal structures.

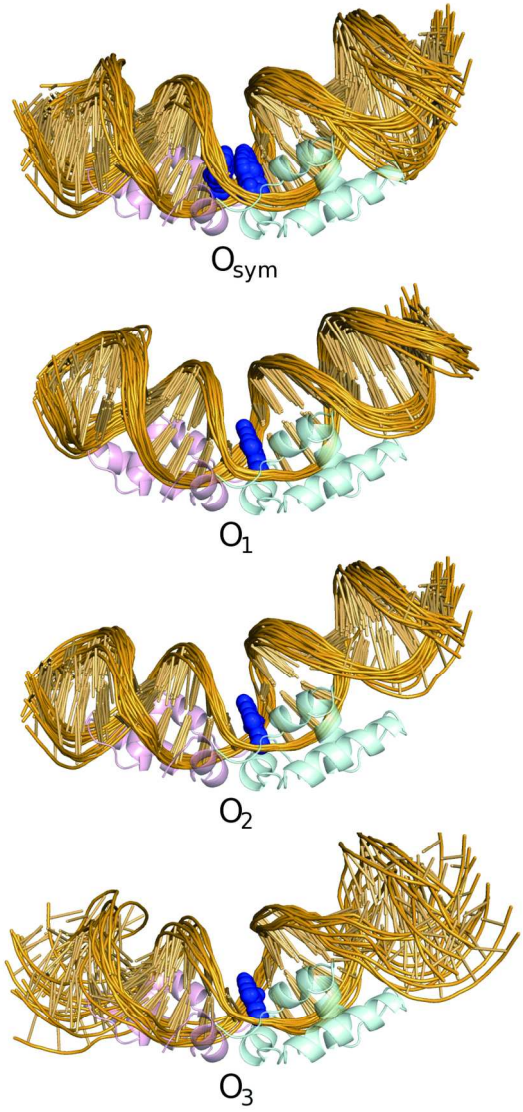
Figure 2. Mosaic images of the DNA twist in NMR-derived structures[14,15] of different Lac repressor-operator complexes. Values of twist, expressed as the number of base pairs per helical turn, are color-coded from light blue to yellow to rose over the range 8-12 bp/turn. The color-coded plots are aligned by sequence with the recognition elements of operators denoted below the relevant images. Values outside the common consensus region, *i.e.*, the three base-pair steps at the ends of the NMR models, are not depicted. The heights of the different mosaics reflect the number of associated model structures. The break in the twist along $O_{sym}$ emphasizes the missing base pair in the symmetric structure. Note the greater unwinding of DNA, *i.e.*, greater number of base pairs per turn, in the weaker protein-operator complexes ($O_2$, $O_3$) and the more pronounced unwinding in the 3′-halves of these duplexes.

Figure 3. Cartoon representations of $O_3$ and $O_1$ operator segments in different orientations on the full Lac repressor assembly. Atomic coordinates obtained by composition of crystal and NMR data (see text) and images rendered in PyMOL (www.pymol.org). Protein chains are depicted in pink (chains A and C near the 5′-ends of the $O_3$ and $O_1$ models) and cyan (chains B and D near the 3′-ends of the DNA). The base pairs that anchor the intervening $O_3 \cdots O_1$ loop, *i.e.*, the A·T base pair at the 3′-terminus of $O_3$ and the T·A base pair at the 5′-terminus of $O_1$, are depicted respectively in deep blue and red. The characters A and P specify the antiparallel or parallel orientations of the 15-bp operators, and the numerals 1 and 2 distinguish whether $O_3$ points toward the inside or outside of the protein assembly. Note the wide variability in loop attachment sites associated with the different combinations of operator models.

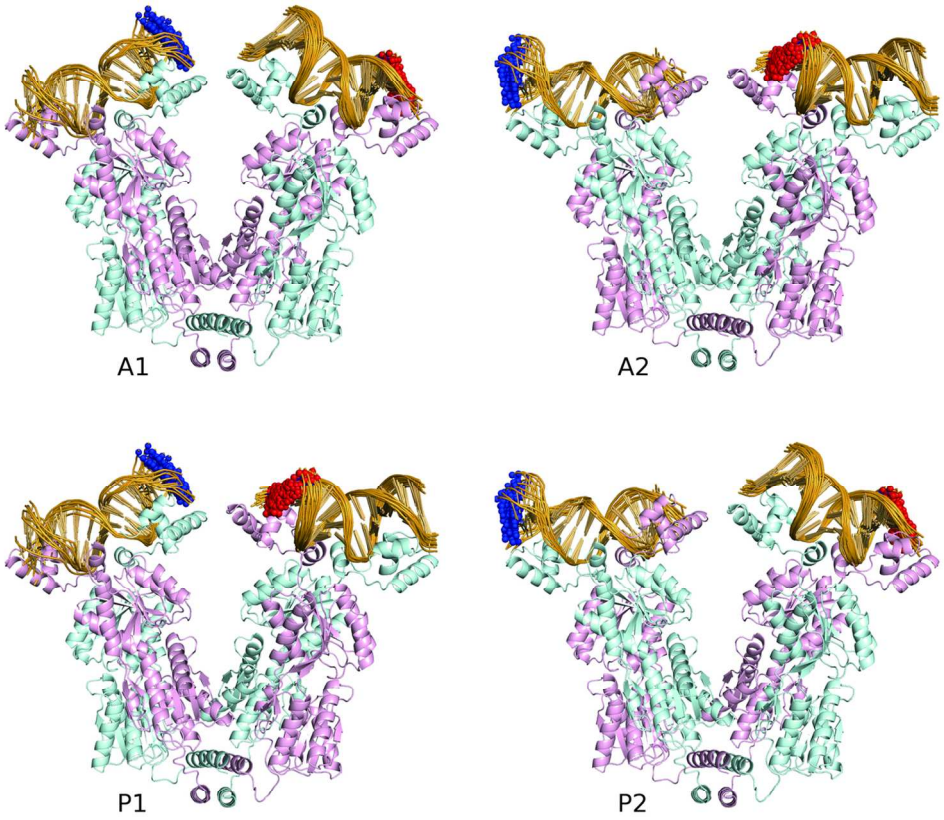
Figure 4. Color-coded mosaic of the computed ease of forming 92-bp DNA loops anchored by Lac repressor-bound $O_3$ and $O_1$ operators and by $O_{sym}$ operators (right and left panels). The blue to red scale on the right denotes the probability of loop formation for the specified combinations of operator models at the 5′- and 3′-ends of the loops. Red denotes the more easily formed loops with higher *J* factors and blue the loops with lower *J* factors. Operator models are depicted by the numerical values used in the Protein Data Bank files (see text) and the looping propensities are expressed in terms of the logarithmic values of the *J* factor. Note (i) the differences in magnitude associated with the formation of antiparallel (A1, A2) *vs*. parallel (P1) loops, (ii) the absence of feasible configurations that meet the stringent constraints imposed on loops anchored in parallel (P2) orientations on the outer edges of the bound operators, (iii) the enhancement of loop formation associated with particular operator models, and (iv) the broad range of values of the *J* factor compared to those determined for the formation of 92-bp loops on static $O_{sym}$ operator structures, where the corresponding values for loops attached in A1, A2, P1, P2 orientations are –11.9, –12.0, –12.1, –13.2.[8]

Figure 5. Distributions of the *J* factors of 92- and 401-bp DNA loops anchored in different orientations on the Lac repressor with all combinations of $O_1$ and $O_2$ operator models. Note (i) the enhancement of loop formation and the decreased sensitivity to operator models at the longer chain length and (ii) the different ordering of looping propensities with operator orientation for the shorter and longer loops.

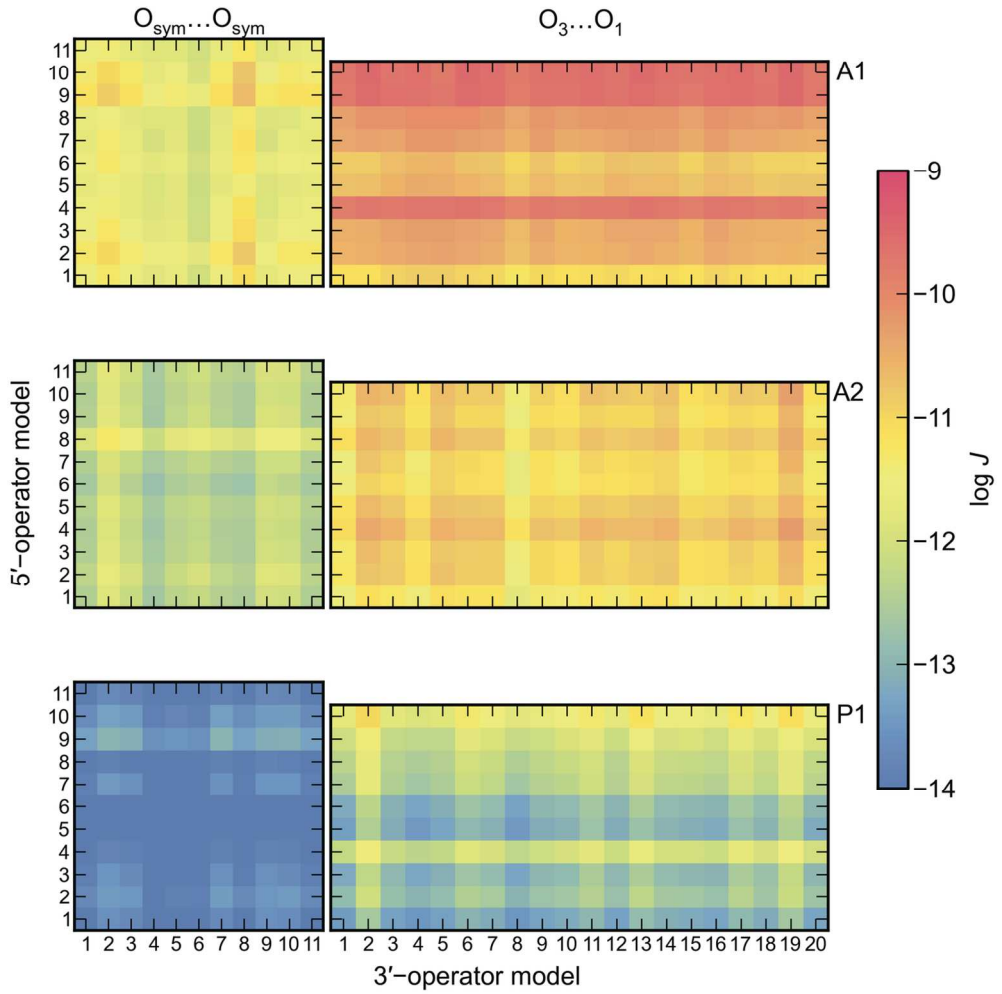
Figure 6. Minimum-energy configurations of 92-bp DNA fragments anchored to different $O_3$ and $O_1$ operator structures on the Lac repressor assembly. The examples are illustrative of the anchoring conditions that enhance (+) and suppress (–) the formation of loops between operators in antiparallel A1 and parallel P1 orientations: (top) $A1^+$, $A1^-$ loops of 20.4, 24.6 *kT* anchored by $O_3\cdots O_1$ models (4···2), (1···10); (bottom) $P1^+$, $P1^-$ loops of 23.8, 30.2 *kT* anchored by models (10···2), (1···8). Images rendered with PyMOL (www.pymol.org). See the legend to Figure 3 for atomic color-coding.



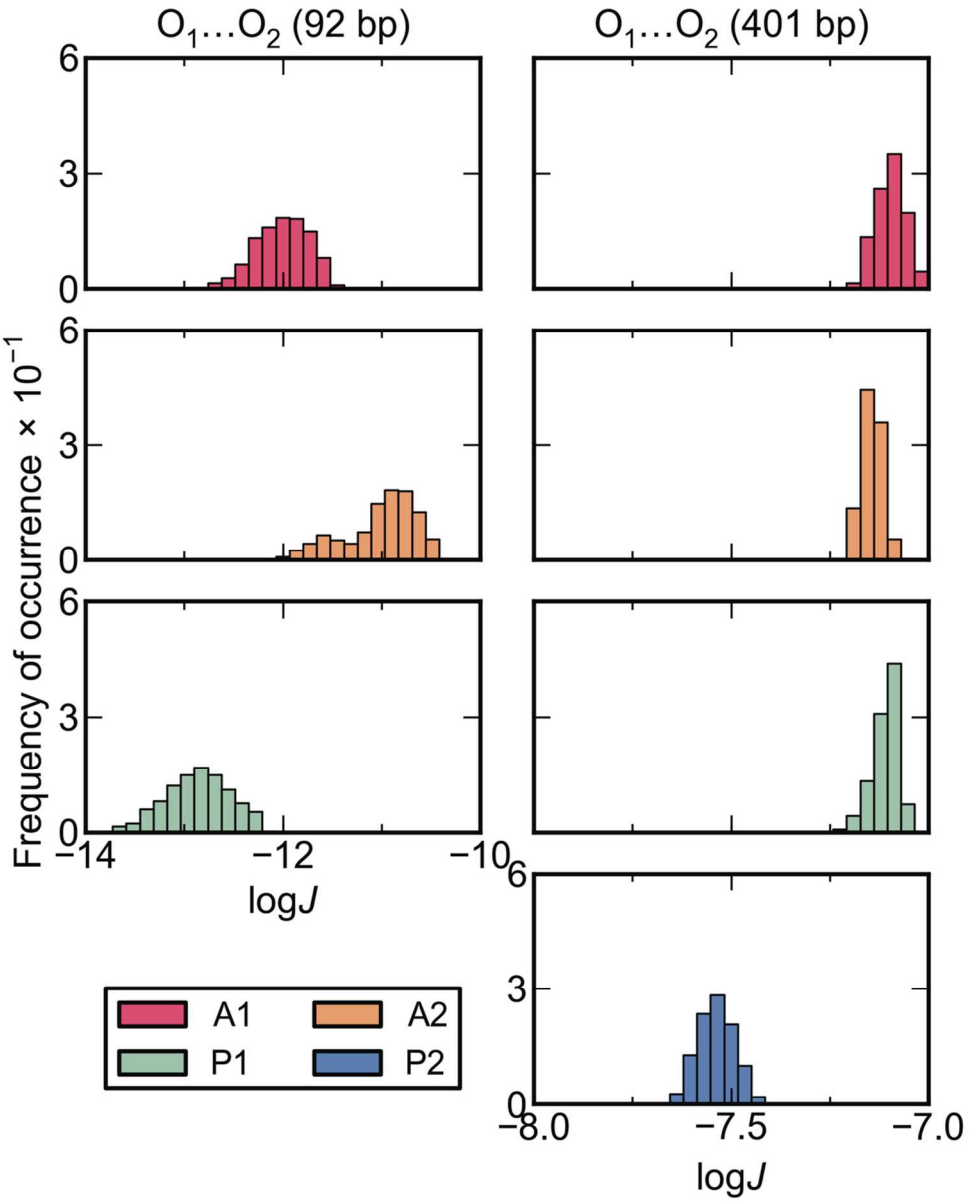

Cartoon images of NMR-derived Lac repressor-operator models[14,15] superimposed on the central base-pair elements of related crystal complexes[19,31] (see text). Images rendered in PyMOL (www.pymol.org) with DNA backbones shown as gold tubes, DNA bases as gold sticks, and protein backbones (chains A and B) as pink and cyan ribbons. Note the wide structural differences among the natural $O_1$, $O_2$, and $O_3$ operators when placed in a common reference frame on the central (deep blue) G·C base pair and the similar spatial pathways of the artificial symmetric operator $O_{sym}$ (centered on its central CG·CG base-pair step) and $O_1$.[14,15] DNA is depicted with the sequence-bearing (recognition) strand running from the lower left to upper right and making contacts at its 5′-end with chain A and its 3′-end with chain B. All views looking down the long axis of the kinked CG·CG step at the centers of the reference crystal structures.
125x272mm (300 x 300 DPI)



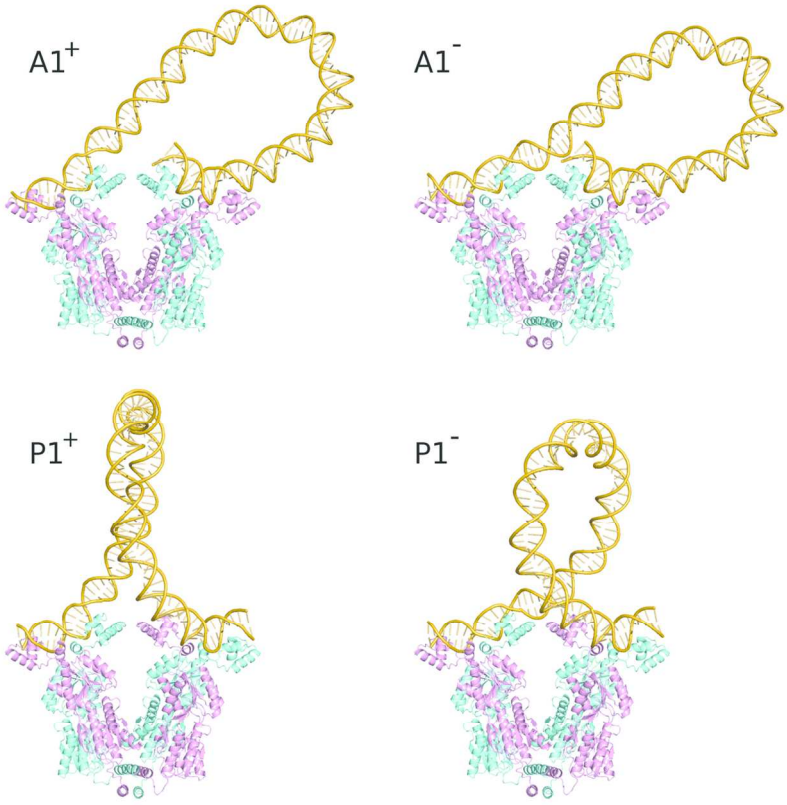

Cartoon representations of $O_3$ and $O_1$ operator segments in different orientations on the full Lac repressor assembly. Atomic coordinates obtained by composition of crystal and NMR data (see text) and images rendered in PyMOL (www.pymol.org). Protein chains are depicted in pink (chains A and C near the 5′-ends of the $O_3$ and $O_1$ models) and cyan (chains B and D near the 3′-ends of the DNA). The base pairs that anchor the intervening $O_3 \cdots O_1$ loop, i.e., the A·T base pair at the 3′-terminus of $O_3$ and the T·A base pair at the 5′-terminus of $O_1$, are depicted respectively in deep blue and red. The characters A and P specify the antiparallel or parallel orientations of the 15-bp operators, and the numerals 1 and 2 distinguish whether $O_3$ points toward the inside or outside of the protein assembly. Note the wide variability in loop attachment sites associated with the different combinations of operator models.
127x115mm (300 x 300 DPI)




Color-coded mosaic of the computed ease of forming 92-bp DNA loops anchored by Lac repressor-bound $O_3$ and $O_1$, operators and by $O_{sym}$ operators (right and left panels). The blue to red scale on the right denotes the probability of loop formation for the specified combinations of operator models at the 5′- and 3′-ends of the loops. Red denotes the more easily formed loops with higher *J* factors and blue the loops with lower *J* factors. Operator models are depicted by the numerical values used in the Protein Data Bank files (see text) and the looping propensities are expressed in terms of the logarithmic values of the *J* factor. Note (i) the differences in magnitude associated with the formation of antiparallel (A1, A2) *vs*. parallel (P1) loops, (ii) the enhancement of loop formation associated with particular operator models, and (iii) the broad range of values of the *J* factor compared to those determined for the formation of 92-bp loops on static $O_{sym}$ operator structures, where the corresponding values for loops attached in A1, A2, P1, P2 orientations are −11.9, −12.0, −12.1, −13.2.[8]
127x130mm (300 x 300 DPI)




Distributions of the *J* factors of 92- and 401-bp DNA loops anchored in different orientations on the Lac repressor with all combinations of $O_1$ and $O_2$ operator models. Note (i) the enhancement of loop formation and the decreased sensitivity to operator models at the longer chain length and (ii) the different ordering of looping propensities with operator orientation for the shorter and longer loops.
101x124mm (300 x 300 DPI)




Minimum-energy configurations of 92-bp DNA fragments anchored to different $O_3$ and $O_1$ operator structures on the Lac repressor assembly. The examples are illustrative of the anchoring conditions that enhance (+) and suppress (–) the formation of loops between operators in antiparallel A1 and parallel P1 orientations: (top) $A1^+$, $A1^-$ loops of 20.4, 24.6 *kT* anchored by $O_3 \cdots O_1$ models (4···2), (1···10); (bottom) $P1^+$, $P1^-$ loops of 23.8, 30.2 *kT* anchored by models (10···2), (1···8). Images rendered with PyMOL (www.pymol.org). See the legend to Figure 3 for atomic color-coding.

133x133mm (300 x 300 DPI)