PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Understanding the basis of a class of paradoxical mutations in AraC through simulations

Ana Damjanovic,[1,2] Benjamin T. Miller,[2] and Robert Schleif[3]*

[1] Department of Biophysics, Johns Hopkins University, Baltimore, Maryland

[2] Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland

[3] Department of Biology, Johns Hopkins University, Baltimore, Maryland

ABSTRACT

Most mutations at position 15 in the N-terminal arm of the regulatory protein AraC leave the protein incapable of responding to arabinose and inducing the proteins required for arabinose catabolism. Mutations at other positions of the arm do not have this behavior. Simple energetic analysis of the interactions between the arm and bound arabinose do not explain the uninducibility of AraC with mutations at position 15. Extensive molecular dynamics (MD) simulations, carried out largely on the Open Science Grid, were done of the wild-type protein with and without bound arabinose and of all possible mutations at position 15, many of which were constructed and measured for this work. Good correlation was found for deviation of arm position during the simulations and inducibility as measured *in vivo* of the same mutant proteins. Analysis of the MD trajectories revealed that preservation of the shape of the arm is critical to inducibility. To maintain the correct shape of the arm, the strengths of three interactions observed to be strong in simulations of the wild-type AraC protein need to be preserved. These interactions are between arabinose and residue 15, arabinose and residues 8–9, and residue 13 and residue 15. The latter interaction is notable because residues L9, Y13, F15, W95, and Y97 form a hydrophobic cluster which needs to be preserved for retention of the correct shape.

Proteins 2013; 81:490–498.

Key words: Langevin dynamics; molecular dynamics; hydrophobic cluster; *in vivo* measurements; AraC protein; gene regulation.

INTRODUCTION

The physical basis for the effects of mutations that drastically alter the stability of a protein or that alter enzymatic activity sometimes can easily be understood. This is not the case, however, for mutations that interfere with the activity of a protein by altering an allosteric regulatory mechanism. Understanding such more subtle properties requires incisive experimental studies and/or computational analysis. Mutations at position 15 in the N-terminal arm of the gene regulatory protein, AraC, are one such class.

The dimeric AraC protein is the arabinose-responsive regulator of the genes in *Escherichia coli* that are required for the uptake and catabolism of arabinose.[1–7] AraC protein has been very well studied, initially because it appeared to be an unusual activator of gene expression, later because it was found to repress expression via a DNA looping mechanism,[8] and continuing into the present because the system is one of a few that is suitable for a deep, but cost effective, analysis that includes genetic, molecular genetic, biochemical, biophysical, and computational studies.[9] The intense and prolonged study has provided the structures of the two domains of the protein as well as a proposal for its mechanism of action, called the light switch mechanism.[10–12] Although the mechanism explains a substantial body of experimental data, much about the protein's regulatory activities remains to be understood. The behavior of mutations at position 15 of the N-terminal arm of AraC differs from the behavior of mutations elsewhere in the 20-residue arm.[13] Mutations at this position do not activate transcription in the presence of arabinose. Almost all alterations in other arm positions render the protein constitutive.[13,14] That is,

Grant sponsor: NSF; Grant number: 1021031 (R. S.); Grant sponsors: NIH, NHLBI, Open Science Grid; Grant sponsors: National Science Foundation and the U.S. Department of Energy's Office of Science
*Correspondence to: Robert Schleif, Biology Department, Johns Hopkins University, 3400 N. Charles St. Baltimore, MD 21218. E-mail: schleif@jhu.edu
Received 17 July 2012; Revised 15 September 2012; Accepted 2 October 2012
Published online 14 November 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24207

instead of repressing transcription from the p_{BAD} promoter in the absence of arabinose and inducing transcription in the presence of arabinose, as is the case for wild-type AraC protein, constitutive AraC mutants induce transcription both in the absence and presence of arabinose.

In addition to the concentration of constitutive mutations in the N-terminal arm of AraC, the role of the arm in the protein's response to arabinose was highlighted by the crystal structures of the domain obtained in the presence and absence of arabinose. In the presence of arabinose, the arm was found to be folded over the bound arabinose,[10] and in the absence of arabinose, the arm was found to be folded in a completely different structure and was not positioned over the arabinose binding pocket.[11,14,15] The remainder of the dimerization domain retained the same structure with and without bound arabinose.

When this study was begun, eight mutations in residue 15 had been isolated.[13] All failed to induce the *ara* p_{BAD} operon in the presence of arabinose. Superficially, the behavior of mutations at position 15 could be easily understood. In the light switch mechanism, the N-terminal arm of AraC is the primary response element to the presence of arabinose.[12] The mechanism postulates that, in the absence of arabinose, the arm occupies a position that immobilizes the DNA binding domains such that looping and repression are energetically favored. In the presence of arabinose, the arm repositions over arabinose that is bound in a pocket of the dimerization domain. This relocation terminates the arms' role in immobilizing the DNA binding domains. The resulting freedom of the domains allows them to occupy adjacent direct repeat DNA sites in the promoter and from there to stimulate transcription. Residue F15 makes significant direct contact with bound arabinose. Therefore, a simple explanation for the behavior of AraC with mutations at position 15 would be that the altered residues' interaction with arabinose is too weak to reposition the arm. Preliminary *in silico* studies with the then known mutants involving energy minimization of the full dimerization domain followed by calculation of the strength of the substituted residue's interaction with arabinose did not yield energy differences sufficient to account for the mutant protein's drastically reduced abilities to induce transcription. Therefore, this is an ideal case for a deeper computational analysis. Such an analysis could be expected to provide an explanation for the behavior of the mutations at position 15, thereby improving our understanding of the mechanism of AraC action, and at the same time, aiding further development of computational analysis of protein function.

Molecular dynamics (MD) has the potential to fully simulate protein folding and conformational changes,[16,17] but often in unrealistic times and at unrealistic computational costs. Nonetheless, even without simulating the full system for periods sufficiently long to include multiple transitions between relevant states, MD can provide valuable information. We, therefore, chose to simulate the wild-type AraC dimerization domain and 18 of the 19 possible substitutions at residue 15 with MD simulations starting from the crystal structure of the arabinose-bound state. We also chose to construct and measure the *in vivo* arabinose responses of the same 18 possible substitutions at position 15 in full-length AraC. With this set of numeric and experimental data, the mechanistic basis of the mutations could become more clear. Substitutions at position 15 appear to have little effect on the repression state, that is, the behavior of the mutations indicates that residue 15 is not critically involved in the repression structure.[13] Thus, the free energy of the repression state of AraC should be relatively independent on the identity of residue 15. Hence, it could be sufficient to examine and compare the properties of the inducing conformations of the arm.

Preliminary simulations of AraC have shown that in case of a removal of arabinose, or introduction of a destabilizing mutation in position 15, the unfolding of the arm from the inducing conformation occurs on a relatively short timescales, that is, hundredths of ps to a ns. To obtain a meaningful sampling of the unfolding events, we chose to perform multiple MD simulations that have been started from different initial velocities. It has been shown previously that better conformational sampling can be achieved by running a large number of short simulations rather than one or a few long simulations.[18–21] In a previous study of regulatory interactions in NtrC, it was found that 25 independent simulations were sufficient to obtain well converged root mean square deviations of the polypeptide chain backbone atoms from the initial structure even though some of the individual simulations deviated significantly.[22] Thus, for simulations of ArC, we chose to run a sample of 25 independent simulations. For the wild type and 18 variants of AraC this amounts to $19 \times 25 = 475$ individual simulations.

To speed the conformational sampling further, we used self-guided Langevin dynamics, SGLD, simulations instead of the regular MD simulations. The self-guided dynamics simulation method differs from Langevin dynamics simulations in the use of additional guiding forces determined on the fly.[23] These forces are proportional to the momentum of the particle averaged over a predetermined time window. The addition of the guiding forces increases search efficiency by enhancing systematic, low-frequency conformational changes. This increase has allowed SGLD to be used to examine conformational changes induced by removal of a phosphate groups in NtrC,[22] charging of internal ionizable groups in mutants of staphylococcal nuclease,[24] and protonation combined with binding of a sugar in lactose permease.[25] The method has recently been reviewed.[26]

Because the large amount of computation required for the study, that is, 475 individual simulations, and because these simulations can be naturally parallelized, a major

portion of the computation was performed over a 2-year interval on the Open Science Grid (OSG) resource for distributed computing.[27] This was possible because of the recent adaptation of CHARMM to run on the Grid.[28]

MATRIALS AND METHODS

Simulated proteins and system setup

The structure of a monomer of the wild-type AraC dimerization domain in the inducing state, PDB code 2ARC,[29] was used for modeling of the variants of AraC. The program SCWRL was used to model the initial structures of the altered side chains.[30] The coordinates of all other side chains were taken from the structure of the wild-type dimerization domain. In total, 18 variants with a substitution of residue F15 were constructed. The F15R variant was not simulated because it was expected that its structure would deviate strongly from the structure of the wild-type protein. Histidine residues were modeled as uncharged and protonated at the Nδ atom with the exception of H80, which was modeled as protonated at the Nϵ atom. Other ionizable residues were modeled as charged except for K15 in the F15K variant, because it was located close to the Arg-38 side chain. The program PROPKA suggested that the pK_a of this lysine residue is shifted to 7.00.[31] We note that because MD simulations showed that this lysine relaxes into a conformation that is more water exposed, its actual pK_a is likely closer to 10.4, suggesting that it is likely charged at pH 7. However, we expect that simulations with a charged Lys-15 would deviate more from the structure of the wild-type protein (due to a presence of Arg-38 in its vicinity), justifying thus a choice of a neutral Lys-15 which will provide a lower bound to structural relaxation.

Crystallographic water molecules and arabinose were included in the initial system setup. These systems were first minimized for 500 steps. For the minimization, system setup and subsequent MD and SGLD runs, the program CHARMM was used[32,33] with the CHARMM force field, version 22.[34] Arabinose parameters were obtained from the glucose parameters.

The briefly minimized protein systems were embedded in a water box and water molecules within 2.5 Å of the protein or crystallographic water molecules were removed. The protein was centered at the coordinate origin, and all water molecules further than 36 Å from the origin were removed. A total of 12 Na^+ ions, and 9 Cl^- ions were added for neutralization. For the Asp or Glu variants, only 8 Cl^- ions were added. The systems were subjected to minimizations under rhombic dodecahedral symmetry. We will refer to these as the starting structures.

MD and SGLD simulations

The systems were heated from 100 to 300 K in steps of 2 K. Equilibration for 100 ps in an NPT ensemble (constant number of particles, pressure, and temperature) followed. The extended system formalism was used to maintain constant pressure and temperature via the Hoover thermostat[35] with a thermostat coupling constant of 1000 kcal/mol/ps, whereas the normal pressure was maintained using a barostat with a piston mass of 500 amu, and piston collision frequency of 20/ps.[36] Rhombic dodecahedral periodic boundary conditions and the particle mesh Ewald method[37] for electrostatic interactions were used, with the following parameters for Ewald simulations: $\kappa = 0.45$, interpolation order of 6, grid spacing of ~1 Å, and real-space interaction cutoff of 10 Å. Lennard-Jones interactions were shifted to zero after 10 Å. The leapfrog Verlet algorithm was used with a time step of 1 fs. For each of the simulated systems, 25 different heating and equilibration runs were initiated with 25 different seed numbers for the random number generator that was used for assigning initial velocities. The simulations were performed using self-guided Langevin dynamics,[23] using as guiding parameters $\lambda = 1$ and $t_L = 0.1$ ps and a friction coefficient γ of 1/ps. SGLD runs were performed in an NVT ensemble (constant number of particles, volume, and temperature) at 300 K. Each SGLD simulation was run for 3 ns, for a total of 75 ns of simulation time for each of the variants.

Open science grid runs

The simulation of the mutant systems was performed using computational resources made available by the OSG.[27] Grid computing projects such as the OSG provide large amounts of computing capacity to researchers. The entire resource, which may span numerous sites, is made available through a consistent interface, which aids in running a large number of very similar computational jobs. In previous work,[28] a workflow management system for CHARMM scripts running on the OSG was developed. This infrastructure was re-used for this study, with each mutant having a separate but identical workflow. Because of time limits on OSG jobs, the 3-ns simulation of each mutant was broken down into 120 simulations of 50 ps apiece. Each of these simulations was run as a single job on one grid processor; the wall clock time was generally 8–12 h per job.

Each workflow was managed by a separate instance of the workflow system's management daemon, which is responsible for submitting jobs to the grid, checking that results have been returned, and detecting when jobs have failed. Each mutant workflow consisted of 2160 separate jobs. Thirty simulations of each mutant were run, with each simulation having two heating jobs, 10 equilibration jobs, and 60 production run jobs. Several mutants were run in parallel at any given time; the exact number was adjusted to give maximum throughput (in general, four to five independent simulations at any one time). Relatively few job failures were noticed, indicating that the grid is a stable platform for simulation science.

RMSD values

For each of the simulations, the root mean square deviations, RMSD, from the energy minimized starting structure of the backbone atoms (atom names C, N, CA, and O) during the third ns of the production run of the simulations were determined. The calculations were performed on snapshots recorded every ps during the third nanosecond of simulation time. The averaging was performed over each of the snapshots and each of the 25 different simulations.

Calculation of interaction energies

Interaction energies were calculated with the INTEraction command in CHARMM. With this command, interaction energy between selected groups of atoms is determined. The calculations are effectively performed in vacuum, and no explicit consideration of solvent effects is taken into account. The calculations were performed on snapshots recorded every 10 ps during the third nanosecond of simulation time.

Generation of the new mutants and measurement of their inducibility

Substitutions W, Y, P, I, T, V, K, and S were not represented in the set of mutations previously studied at position 15.[13] Quik-Change® mutagenesis (Stratagene) was used to introduce them into the vector pWR03.[38] Candidates were verified by DNA sequencing and their DNA was transformed into strain SH321.[39] Induction was quantified as the level of arabinose isomerase in SH321 cells grown overnight to stationary phase in YT broth medium containing 0.2% L-arabinose.[40]

RESULTS

The paradoxical F15L variant

This work was initiated because, as mentioned in the introduction, eight of the mutations then existing at position 15 in the N-terminal arm of AraC eliminated the protein's ability to respond to arabinose, an atypical property of arm mutations.[13] Residue F15 is located within a small hydrophobic cluster formed by residues Y13, P8, V20, and Y97 in which the hydroxyl groups of the two tyrosine residues lie outside the hydrophobic cluster and point away from it (Fig. 1). On this basis, it seemed plausible that the replacement of F15 in AraC with another large hydrophobic residue would retain the cluster and the protein would display the normal wild-type *in vivo* behavior. This was not the case however, as the F15L variant was among the original eight, and it was less than 1% as inducible as the wild type.

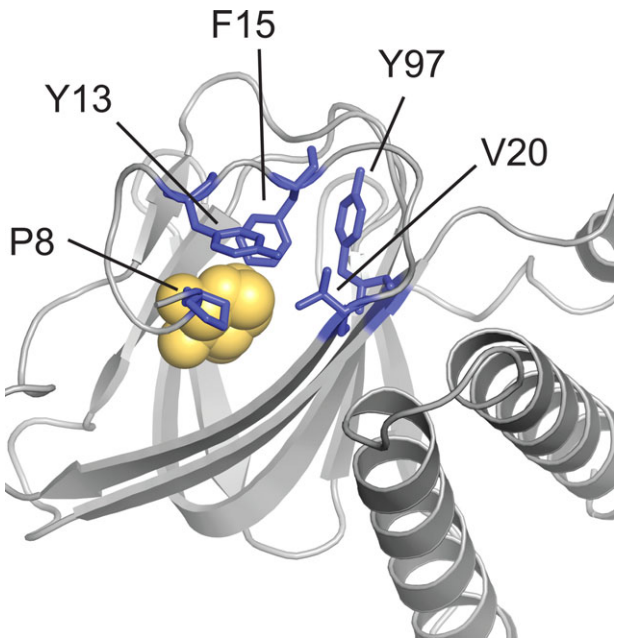

Figure 1

A dimerization domain of AraC in cartoon form with containing a bound arabinose molecule in Van der Waals representation showing in stick form residues P8, Y13, F15, V20, and Y97. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Because the side chain of F15 directly contacts arabinose, it had also seemed possible that the strength of the F15-arabinose interaction is critical to inducibility and perhaps the L15-arabinose interaction was too weak. Substituting the best fitting leucine rotamer, energy minimizing the domain with CHARMM, and then using CHARMM to calculate the interaction energy yielded an interaction energy of −0.41 kcal/mol compared to −1.18 kcal/mol for phenylalanine calculated in the same way. This crude calculation yielded a difference of 0.77 kcal/mol between the two interaction energies. This is not sufficient to explain the experimentally observed difference in inducibilities.

Molecular dynamics simulations of the wild type and the F15L variant

As neither the retention of hydrophobicity in the region nor the strength of residue 15′s interaction with arabinose provided an explanation for the absence of inducibility, we turned to MD simulations for an explanation of the mutants' behaviors. Because the wild-type protein adopts the inducing conformation only in the presence of arabinose, comparing the plus and minus arabinose MD trajectories of the wild-type protein should reveal differences indicative of inducibility. Such a difference would appear to be highly diagnostic if both

the plus and minus arabinose simulations of the uninducible F15L variant behaved like the minus arabinose simulations of wild type.

The average RMSD profiles of the backbone atoms of the two proteins in the presence and in the absence of arabinose were averaged over 25 independent simulations as described in the Methods section. The behavior of the WT protein simulated with arabinose present was different from the simulation without arabinose present. For the plus arabinose trajectory, the averaged RMSD profile of the arm region from the starting structure was small; for the minus arabinose trajectory, the averaged RMSD profile was large (Fig. 2). The averaged RMSD profiles of the residues beyond 20 were the same for the two simulations. In the case of the F15L mutant, both the plus and minus arabinose simulations displayed the large averaged RMSD values in the arm region. Thus, large values of the RMSD in the arm region appeared to be indicative of lack of induction activity of the protein.

Comparison of other F15 variants with the wild type

If the correlation between large averaged RMSD profiles of the N-terminal arm and the absence of inducibility as noted in the previous section were true for all variants at position 15, it would then be sensible to look for the mechanistic basis for the large variability. Therefore, with molecular genetics we constructed the remainder (except for F15R) of the variants at position 15, measured their *in vivo* inducibilities, and compared these to the RMSD profiles found in MD simulations, with arabinose present, of the same mutants. The RMSD profiles in the arm region displayed considerable variability among the different variants while all the profiles beyond residue 20 were highly similar.

We used several methods to quantify the differences in the behavior of the arm region. First, we determined the averaged backbone RMSD values for residues 7–18 from the 25 profiles (Table I). The wild-type protein still exhibited the smallest arm RMSD value. The RMSD of the arm in the F15W mutant was found to be closest to the wild type. Notably, only this mutant and wild-type protein are significantly inducible *in vivo*.

The RMSD average over 25 simulation runs for a mutant could be dominated by the behavior of the arm in a small number of runs in which the arm is unfolded. Therefore, we also characterized the behavior of each mutant by counting the number of simulation runs in which the average of the arm RMSD during the third nanosecond was less than 2 Å, (Table I, column 3). By this criterion, wild-type and the F15W mutant, both of which are inducible, are even more sharply distinguished from the remainder, all of which are uninducible.

In view of the role proposed for the arm in the light switch mechanism of AraC, another logical measure of arm behavior would be the fraction of the time, which is approximated by the number of the simulations of the 25, in which the arm is significantly dissociated from arabinose over the third nanosecond of the simulation. Because of the dynamic nature of the system, no distance between any single atom of the arm and an atom of the arabinose was truly indicative. Therefore, the average over the third nanosecond of the minimal distance between any backbone nitrogen or oxygen atom of residues 8 and 9, and any oxygen atom of arabinose, (Table I, column 4) was used as a criterion for dissociation. The numbers of

Table I
Molecular Dynamics and In Vivo Properties of Wild Type and the Mutants at Position 15

Residue 15	Average arm RMSD, Å	No. of correctly folded (RMSD<2 Å) out of 25[a]	No. of dissociated (R_{min} > 4 Å) out of 25[b]	Inducibility[c]
F(WT)	1.83	19	5	1
W	2.39	17	7	0.8
M	2.48	13	7	<0.1
P	2.91	9	5	<0.1
H	3.03	4	14	<0.1
C	3.15	2	6	<0.1
L	3.16	3	10	<0.01
I	3.29	3	5	<0.1
Y	3.48	3	17	Ambiguous[d]
T	3.51	3	15	<0.1
V	3.71	2	10	<0.1
A	4.05	5	12	<0.1
K	4.05	6	16	<0.1
Q	4.13	1	17	<0.1
N	4.38	1	19	–
D	4.42	0	19	<0.1
G	4.46	3	14	<0.1
E	4.49	2	16	<0.1
S	4.57	3	19	<0.1

[a]Arm is considered correctly folded in a simulation if the average RMSD <2 Å.
[b]Arm is considered dissociated in a simulation if the average minimal distance between oxygen atoms of arabinose and the backbone N and C atoms of residues 8 or 9, R_{min}, is more than 4 Å.
[c]Inducibility is arabinose isomerase level relative to that measured in wild-type cells.
[d]Inducibility could not be ascertained because this mutant is constitutive whereas all others possessed wild-type expression levels in the absence of arabinose.

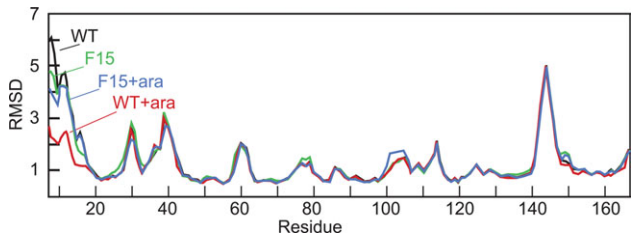

Figure 2
RMSD values in Å of the wild-type AraC dimerization domain with and without arabinose, red and black respectively, and the F15L variant with and without arabinose, blue and green, respectively, averaged over 25 independent simulations.

Table II
Various Interaction Energies in Wild Type and Residue 15 Mutants[a]

Residue 15	Arabinose residues 7–18	Arabinose residues 8–9	Arabinose residue 15	Residue Y13 residue 15
F(WT)	−7.15	−4.65	−1.18	−4.98
W	−7.12	−4.68	−1.31	−6.04
M	−5.87	−4.00	−0.89	−2.52
P	−6.29	−4.89	−0.01	−3.08
H	−8.83	−4.69	−2.97	−2.77
C	−5.47	−4.04	−0.70	−0.79
L	−5.90	−4.55	−0.39	−1.62
I	−6.79	−4.87	−0.22	−2.41
Y	−7.25	−3.21	−2.82	−4.67
T	−5.61	−4.87	0.16	−1.09
V	−5.5	−3.87	−0.11	−2.22
A	−5.08	−4.56	0.03	−1.08
K	−6.98	−4.85	−0.86	−3.09
Q	−3.59	−3.18	−0.09	−1.62
N	−5.01	−3.7	−0.49	−1.54
D	–	–	–	–
G	−4.52	−3.76	0.00	−0.64
E	−7.32	−5.91	2.00	−7.22
S	−7.73	−5.73	−0.02	−1.80

[a]Averages over the full third nanosecond for those simulations in which the RMSD of the arm remained less than 2 Å as in Table I. Energies in kcal/mol.

dissociated trajectories strongly anticorrelates with number of the trajectories in which the arm exhibited an average RMSD of less than 2 Å, which we will refer to as correctly folded, (Table I, column 3). Indeed, visual inspection of the trajectories of the mutants with program VMD[41] that most markedly deviated from the anticorrelation, M, P, C, I, and V revealed that in many of the simulation runs for these mutants the arm remained folded, but incorrectly because it is in a position that is different from that observed for the wild-type arm. Overall, the RMSD criterion provides a closer correspondence to the *in vivo* inducibility than dissociation.

Interaction energies

To analyze why the wild type and the tryptophan substitution, but none of the other substitutions at position 15, are inducible, we determined average interaction energies involving the arm over the third nanosecond of those simulations for which the arm exhibited an average RMSD of less than 2 Å, (Table II). In addition to the wild-type arm and the F15W substituted arm, residues 7–18 of the arm from several other variants, F15H, F15Y, F15E, and F15S, also exhibited strong interaction energies with arabinose. Thus, the strength of this interaction alone does not determine the behavior.

The interaction energies were calculated as described in the Methods section. To test the quality of the calculated interaction energies, for two proteins, WT and the F15H variant, we performed a calculation of interaction energies based on the difference in the energy of a dimer (protein + arabinose) and of the two monomers. The dimer and the monomers were modeled to be solvated. Solvation energies were determined by using the aspenr routine in CHARMM, which is based on the atomic solvation parameters of Wesson and Eisenberg.[42] This simple test calculation showed an excellent correlation (correlation coefficient 0.98) of per-residue interaction energies determined with the two methods. The magnitude of interaction energies between the two methods (in vacuum and by using aspenr) differed, however, suggesting that we can deduce meaningful conclusions about which protein residues are important in arabinose–protein interactions, but that we cannot determine the absolute binding energies correctly.

In the wild-type protein, the strongest individual residue contributions to the interaction energy between arabinose and the arm arose from residues Pro-8 (−2.64 kcal/mol) and Leu-9 (−2.01 kcal/mol), apparently because the backbone atoms of these two residues are well positioned to make hydrogen bonds with arabinose. The interaction of arabinose with F15 amounts to −1.18 kcal/mol and the contribution of all arm residues except residues 8, 9, and 15 amounts to only -1.32 kcal/mol. Neither the interaction strength of residues 8 and 9 with arabinose, nor that of residue 15 with arabinose, nor the strengths of the remaining interactions with arabinose fully explain why only the wild type and F15W AraC proteins are inducible (Table II).

Because it is residue 15 that changes in these experiments, we examined its interaction energy with its neighbors, residues P8, Y13, V20, M42, R38, W95, and Y97. Its interactions with residue 13 are of greatest interest because they were the strongest for the wild type (Table II). Overall, only the F15W mutant arm exhibited the same approximate interaction energies as the wild-type arm for all three of the key interactions, that is, the interactions of residues 8, 9, and 15 with arabinose, and the interaction of residues 13 and 15.

Details of several specific cases

Here, we discuss in detail the interactions and the structure of the arm for the uninducible mutants that, nonetheless, exhibited strong interactions with arabinose, mutants F15H, F15Y, F15E, and F15S. This discussion is meant to highlight that when it comes to stabilization of a given shape, having a correct distribution of interaction energies is more important than having a high total interaction energy.

In simulations of the F15H variant, the arm exhibited an average RMSD of less than 2 Å in only four simulations. The interaction energy between arabinose and residue 15 for those four simulation runs was −2.97 kcal/mol, larger than in the wild type; however, the interaction energy between residues 13 and 15 was reduced compared to the wild type. Thus, the distribution of interaction energies in the F15H is different from the distribution of interaction energies for the wild type. Not

surprisingly, in a large number of simulations, residue 13 became solvent exposed and left the hydrophobic cluster.

In simulations of the F15Y variant, the arm exhibited an average RMSD of less than 2 Å in only three simulations. For those three simulations however, the interaction energy between arabinose and residues 8 and 9 was reduced by about 1.5 kcal/mol compared to the wild type. Inspection of trajectories revealed a variety of different patterns of deviation from the correct structure. Some revealed that the OH group of Tyr-15 or a water molecule made hydrogen bonds with residue 8 or with arabinose, thus disturbing the wild-type hydrogen bonding pattern.

The arm remained correctly folded, that is, exhibited an average RMSD of less than 2 Å, in only two trajectories in simulations of the F15E variant. These two showed an ion pair interaction between Glu-15 and Arg-35. Inspection of the two trajectories revealed a strong interaction between residues 13 and 15 involving the backbone of residue 13 rather than the side chain. In this variant, the interaction between Glu-15 and arabinose was repulsive. In most other trajectories, the Glu-15-Arg-38 ion-pair remained; however, the arm was either unfolded or moved into a position different from that of the wild-type arm.

In simulations of the F15S variant, the arm remained correctly folded in only three trajectories. The strong interaction energy of arabinose with the arm arose through the stronger than usual interactions with residues 8–11. The interaction with residue 15, and between residues 13 and 15, was strongly diminished. In a majority of the simulations residue 15 became solvent exposed.

Comparison of SGLD and MD simulations

Previous studies of protein conformational changes have shown that SGLD can sample conformational space better than conventional MD.[22,24,25] An interesting question is whether this is also the case for this protein. To compare the two forms of dynamics, we performed 25 runs of 5 ns each of regular MD simulations of wild type, F15W, F15L, and F15Y. The average RMSD values of the arm and the number of runs during which the arm remained correctly folded that were obtained with the two types of simulations were determined (Table III). The MD simulations exhibit roughly the same trends as the SGLD simulations in that the F15W variant is the most similar to the wild type. Experiments have indicated that F15L and F15Y variants are however quite different from the WT and the F15W variant. These differences are much more pronounced in SGLD simulations than in MD simulations, in agreement with experimental results. These results indicate that SGLD simulations are more appropriate than regular MD simulations for quick screening of the structural relaxation induced by mutation.

Table III
Comparison of Regular and SGLD Dynamics of AraC

Residue 15	SGLD[a]		MD[c]	
	Average RMSD residues 7–18	No. of folded out of 25[b]	Average RMSD residues 7–18	No. of folded out of 25[b]
F(WT)	1.83	19	1.69	19
W	2.39	17	1.81	19
L	3.16	3	2.31	10
Y	3.48	3	1.98	15

[a]Self guided Langevin dynamics of 3 ns.
[b]Number of simulations in which the arm remained correctly folded over the third ns of simulation.
[c]Standard molecular dynamics simulation of 5 ns.

DISCUSSION

This investigation was initiated by the surprising finding that, while mutations at most positions in the N-terminal arm of AraC protein resulted in constitutive regulatory behavior, the eight known mutations then known at residue 15 left the protein not constitutive and, unlike wild-type protein, unable to respond to arabinose.[13,14] Simple explanations based on hydrophobicity or contacts between arabinose and residue 15 did not explain the anomalous behavior. If the protein, wild type and mutant, could be accurately simulated, both would display the actual behavior that is observed for their real counterparts and the simulations could be dissected to determine the basis of the mutant's behavior. While extraordinary computational efforts in simulating proteins with MD have recently revealed previously unattainable details of protein folding and allosteric transitions,[16,17] it seemed possible that less heroic computational efforts in simulating AraC with MD might also provide substantial understanding.

A major problem in MD simulations is that important conformational transitions in proteins may take place on the millisecond or longer time scale, but the simulations must be performed in femtosecond time steps. Recently it has been found that the rate of conformational transitions in MD simulations can be significantly increased by the use of a variant that is termed self-guided Langevin dynamics,[23] and this approach was used in the work described here. The cost of the lengthy computations required for this work was also greatly alleviated by the use of the OSG,[27] on which it has recently become possible to run the MD program CHARMM.[28,32,33]

Our MD simulations of the arabinose binding/dimerization domain of AraC were performed both with arabinose present and bound to the protein and absent. The simulations are consistent with what is known of the genetics, biochemistry, and biophysics of the protein.[8,12,38,42–46] The simulations indicate that with arabinose bound to the wild-type protein, the 20 residue N-terminal arm relocates from another position to a position directly over the bound arabinose, and that it is the arm's removal from the former position that allows

the protein to adopt the inducing structure. Hence, the failure of a mutant to respond to arabinose likely results from a failure of the arm to bind over arabinose. In our simulations with the wild-type protein, we found that if the simulation were begun with the arm in the plus arabinose position and if arabinose were present, the arm largely stayed close to its starting position. When arabinose was not present in the simulation, the arm changed structure and often moved away from the bound arabinose. Hence, at least with wild-type protein, retaining the starting structure was indicative of induction. We then found that this correlation was consistent with the inducibility of the remainder of the eight mutants at position 15 of the arm that were known at the time. Simulations predicted however, that yet another substitution at position 15, tryptophan, might exhibit the wild-type behavior. It was constructed and found to do so. Therefore, the rest of the possible substitutions at position 15 were constructed and measured as well as simulated. Strong correlation is found between inducibility and the ability of the arm to retain a shape similar to that of the wild-type arm in the presence of arabinose (Table I). For some of the mutants with hydrophobic substitutions, the arm did remain folded in the simulations, but it deviated from the shape of the wild-type arm. Experimentally, these mutants were still uninducible.

Can the basis of the retention of the correct arm shape be found in the relevant MD trajectories? Analysis of the interaction energies between various residues of the N-terminal arm and arabinose and among the residues themselves revealed that no single interaction energy is the determining factor in the arm's behavior. Instead, three strong interactions were identified as important: the strength of the arabinose interaction with residue 15, the strength of the interaction between arabinose and the residue 8 plus 9 pair, and the strength of the interaction between residues 13 and 15. The latter interaction is notable because residues L9, Y13, F15, W95, and Y97 form a hydrophobic cluster. Apparently alteration of either the hydrophobicity or the shape of residue 15 interferes with the integrity or shape of this cluster.

CONCLUSIONS

In summary, complete correspondence has been found between the *in vivo* regulatory behavior of all 19 different variants at position 15 in AraC and the behavior of the N-terminal regulatory arm in self-guided Langevin molecular dynamics simulations. The basis for the behavior appears to be retention or not of a correct shape of the hydrophobic cluster of residues in the protein.

ACKNOWLEDGMENTS

The authors thank Michael Rodgers and Bernard R. Brooks for discussions and comments on the manuscript, as well as Petar Maksimovic for help with the Open Science Grid operations.

REFERENCES

1. Sheppard DE, Englesberg E. Further evidence for positive control of the L-arabinose system by gene *araC*. J Mol Biol 1967;25:443–454.
2. Englesberg E, Irr J, Power J, Lee N. Positive control of enzyme synthesis by gene C in the L-arabinose system. J Bacteriol 1965;90: 946–957.
3. Steffen D, Schleif R. Overproducing *araC* protein with lambda-arabinose transducing phage. Mol Gen Genet 1977;157:333–339.
4. Schleif R. An L-arabinose binding protein and arabinose permeation in *Escherichia coli*. J Mol Biol 1969;46:185–196.
5. Brown CE, Hogg RW. A second transport system for L-arabinose in *Escherichia coli* B-r controlled by the *araC* gene. J Bacteriol 1972;111:606–613.
6. Greenblatt J, Schleif R. Arabinose C protein: regulation of the arabinose operon *in vitro*. Nat New Biol 1971;233:166–170.
7. Kolodrubetz D, Schleif R. Regulation of the L-arabinose transport operons in *Escherichia coli*. J Mol Biol 1981;151:215–227.
8. Dunn TM, Hahn S, Ogden S, Schleif RF. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proc Natl Acad Sci USA 1984;81: 5017–5020.
9. Schleif R. AraC protein, regulation of the L-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. FEMS Microbiol Rev 2010;34:779–796.
10. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. Structural basis for ligand-regulated oligomerization of AraC. Science 1997;276:421–425.
11. Weldon JE, Rodgers ME, Larkin C, Schleif RF. Structure and properties of a truely apo form of AraC dimerization domain. Proteins 2007;66:646–654.
12. Saviola B, Seabold R, Schleif RF. Arm-domain interactions in AraC. J Mol Biol 1998;278:539–548.
13. Ross JJ, Gryczynski U, Schleif R. Mutational analysis of residue roles in AraC function. J Mol Biol 2003;328:85–93.
14. Dirla S, Chien JY, Schleif R. Constitutive mutations in the *Escherichia coli* AraC protein. J Bacteriol 2009;191:2668–2674.
15. Rodgers ME, Holder ND, Dirla S, Schleif R. Functional modes of the regulatory arm of AraC. Proteins 2009;74:81–91.
16. Dror RO, Arlow DH, Borhani DW, Jensen MØ, Piana S, Shaw DE. Identification of two distinct inactive conformations of the β2-adrenergic receptor reconciles structural and biochemical observations. Proc Natl Acad Sci USA 2009;106:4689–4694.
17. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science 2010;330:341–346.
18. Elofsson A, Nilsson L. How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized *Escherichia coli* thioredoxin. J Mol Biol 1993;233:766–780.
19. Caves LSD, Evanseck JD, Karplus M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Protein Sci 1998;7:649–666.
20. Damjanović A, Schlessman JL, Fitch CA, García AE, García-Moreno EB. Role of flexibility and polarity as determinants of the hydration of internal cavities and pockets in proteins. Biophys J 2007;93:2791–2804.
21. Auffinger P, Louise-May S, Westhof E. Multiple molecular dynamics simulations of the anticodon loop of tRNAAsp in aqueous solution with counterions. J Am Chem Soc 1995;117:6720–6726.
22. Damjanovic A, García-Moreno EB, Brooks BR. Self-guided Langevin dynamics study of regulatory interactions in NtrC. Proteins: Struct Funct Bioinform 2009;76:1007–1019.

23. Wu X, Brooks BR. Self-guided Langevin dynamics simulation method. Chem Phys Lett 2003;381:512–518.
24. Damjanovic A, Wu X, García-Moreno EB, Brooks BR. Backbone relaxation coupled to the ionization of internal groups in proteins: a self-guided Langevin dynamics study. Biophys J 2008;95:4091–4101.
25. Pendse PY, Brooks BR, Klauda JB. Probing the periplasmic-open state of lactose permease in response to sugar binding and proton translocation. J Mol Biol 2010;404:506–521.
26. Wu X, Damjanovic A, Brooks BR. Efficient and unbiased sampling of biomolecular systems in the canonical ensemble: A review of self-guided Langevin dynamics. In: Rice SA, Dinner AR, eds. Advances in chemical physics volume 150. Hoboken: John Wiley & Sons; 2012. p. 255–326.
27. Pordes R, Petravick D, Kramer B, Olson D, Livny M, Roy A, Avery P, Blackburn K, Wenaus T, Würthwein F, Foster I, Gardner R, Wilde M, Blatecky A, McGee J, Quick R. The open science grid. J Phys: Conf Ser 2007;78:012057.
28. Damjanovic A, Miller BT, Wenaus TJ, Maksimovic P, Garcia-Moreno EB, Brooks BR. Open science grid study of the coupling between conformation and water content in the interior of a protein. J Chem Inf Model 2008;48:2021–2029.
29. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. J Mol Biol 1997;273:226–237.
30. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Struct Funct Bioinform 2009;77:778–795.
31. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins: Struct Funct Bioinform 2005;61:704–721.
32. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
33. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: The biomolecular simulation program. J Comput Chem 2009;30:1545–1614.
34. MacKerell AD, Bashford D, Bellott , Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wirkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
35. Hoover WG. Canonical dynamics: equilibrium phase–space distributions. Phys Rev A 1985;31:1695.
36. Feller SE, Zhang Y, Pastor RW, Brooks BR. Constant pressure molecular dynamics simulation: The Langevin piston method. J Chem Phys 1995;103:4613–4621.
37. Darden TA, York DM, Pedersen LG. Particle mesh Ewald: an $N.\log(N)$ method for Ewald sums in large systems. J Chem Phys 1993;98:10089.
38. Reed WL, Schleif RF. Hemiplegic mutations in AraC protein. J Mol Biol 1999;294:417–425.
39. Hahn S, Dunn T, Schleif R. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. J Mol Biol 1984;180:61–72.
40. Schleif RWP. Methods in Molecular Biology. New York: Springer-Verlag; 1981.
41. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–38.
42. Wesson D, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
43. Wu M, Schleif R. Mapping arm-DNA-binding domain interactions in AraC. J Mol Biol 2001;307:1001–1009.
44. Harmer T, Wu M, Schleif R. The role of rigidity in DNA looping-unlooping by AraC. Proc Natl Acad Sci USA 2001;98:427–431.
45. Seabold RR, Schleif RF. Apo-AraC actively seeks to loop. J Mol Biol 1998;278:529–538.
46. Lobell RB, Schleif RF. DNA looping and unlooping by AraC protein. Science 1990;250:528–532.
47. Martin K, Huo L, Schleif RF. The DNA loop model for ara repression: AraC protein occupies the proposed loop sites *in vivo* and repression-negative mutations lie in these same sites. Proc Natl Acad Sci USA 1986;83:3654–3658.