Published in final edited form as:

J Mol Biol. 2010 May 7; 398(3): 462–470. doi:10.1016/j.jmb.2010.03.021.

Computational predictions of the mutant behavior of AraC

Monica Berrondo¹, Jeffrey J. Gray¹, and Robert Schleif²

¹Chemical and Biomolecular Engineering

²Department of Biology Johns Hopkins University

Abstract

An algorithm implemented in Rosetta correctly predicts the folding capabilities of the 17 residue N-terminal arm of the AraC gene regulatory protein when arabinose is bound to the protein and the dramatically different structure of this arm when arabinose is absent. The transcriptional activity of 43 mutant AraC proteins with alterations in the arm sequences was measured *in vivo* and compared with their predicted folding properties. Seventeen of the mutants possessed regulatory properties that could be directly compared to folding predictions. Sixteen of the seventeen were correctly predicted. The algorithm predicts that the N-terminal arm sequences of AraC homologs fold to the *E. coli* AraC arm structure. In contrast it predicts that random sequences of the same length and many partially randomized *E. coli* arm sequences do not fold to the *E. coli* arm structure. The high success shows that relatively “simple” computational methods can in some cases of predict the behavior of mutant proteins with good reliability.

Introduction

In the past few years, computational methods have become increasingly successful in designing and redesigning proteins.^{1–6} Such success raises the question of whether the computational methods that have been developed to predict the folded structure of an amino acid sequence can be used for related problems. For example, if the structure and activity of a wild type protein is known, can the activities of mutants be predicted? While prior knowledge of the structure and activity of the native protein is helpful, the problem is particularly difficult for two reasons. First, the relevant energy differences between the wild type and mutant protein in general will be much less than the overall stability of a folded protein—a criterion by which correct folding is often assessed.⁷ Second, the structural changes induced by most mutations are relatively small, and at present it is not possible to relate such changes to changes in enzymatic or regulatory protein activities.^{8–10}

One of the major hurdles in relating structural changes with changes in activity is the lack of suitable datasets from experimental studies on which computational methods can be trained and tested. The necessary experimental studies would include a considerable set of mutations whose activity has been characterized systematically.

© 2009 Elsevier Ltd. All rights reserved.

CORRESPONDING AUTHOR: Robert Schleif Department of Biology Johns Hopkins University 3400 North Charles Street Baltimore, MD 21218 USA Office Telephone: 410-516-5206 Fax Number: 410-516-5213 schleif@jhu.edu .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

AraC protein, the regulator of the L-arabinose operon in *Escherichia coli*, provides a unique opportunity to evaluate the effectiveness of computation in a problem intermediate between that of predicting the folding of a protein and predicting the change in activity due to a minor structural change induced by a mutation. AraC protein regulates the synthesis of the arabinose uptake and catabolic enzymes in response to arabinose. A considerable body of evidence has accumulated for a particularly simple on and off mechanism of the protein known as the light switch mechanism.^{5,11-15} In this, in the absence of arabinose, the N-terminal 20 or so amino acids possess one conformation and the protein acts to repress. In the presence of arabinose, the arm, and only the arm, dramatically restructures, and the protein acts as an inducer. In this case, and as far as is known to the authors, this is the only case thus far known, it should be possible to predict the activity of mutant AraC proteins by predicting whether mutant N-terminal arms of the protein can fold to the repressing structure in the absence of arabinose, and the inducing structure in the presence of arabinose. Further recommending AraC for such an analysis is the fact that the relevant structures are known and there exists a very considerable body of knowledge about the regulatory properties of wild type and many mutant proteins.

The basic mechanism of the AraC protein is shown in Fig. 1. In the absence of arabinose, the protein's two DNA binding domains are rigidly held to the dimerization domain,¹⁶ an interaction in which the 23 N-terminal amino acids play a key role.^{5,14,17} This conformation of the protein favors its binding to two nonadjacent DNA half-sites and the formation of a DNA loop that interferes with RNA polymerase access to the promoter.¹⁸ When arabinose is present, its binding to the dimerization domain of AraC makes it more favorable for the N-terminal arm of the protein to fold over the arabinose than to mediate the interaction between the dimerization and DNA binding domains.^{5,19} This relocation of the N-terminal arm frees the DNA binding domains from their constraints in orientation and position and allows them, instead of forming a DNA loop, to bind to the two adjacent DNA half-sites alongside the RNA polymerase binding site.¹¹ The absence of the DNA loop and the presence of AraC bound to the two DNA half-sites beside the RNA polymerase binding site activates the initiation of transcription by RNA polymerase, leading to a 100-fold induction of expression of the adjacent *ara* genes.^{20,21}

The available evidence suggests that the restructuring of the arm from its apo structure in the absence of arabinose to its holo structure in the presence of arabinose is the triggering effect of the binding of arabinose.^{22,23} Any arabinose-induced structural changes in the core of the dimerization domain are thought to play almost no role in the protein's response to arabinose.

In the work reported here we developed a straightforward computational protocol for refolding the regulatory arm of wild type AraC protein in the presence and absence of arabinose. We applied the same protocol to mutant arm sequences and found that the protocol correctly predicted the behavior of 16 out of 17 mutants.

Results

Structure prediction the N-terminal arm

Biochemical experiments have shown that folding of the arm to its apo structure in the absence of arabinose and holo structure in the presence of arabinose does not require the presence of the DNA binding domain.^{5,22,23} Therefore, the DNA binding domain was omitted from all the calculations described here. Further, residues 1-6 of the protein may be deleted without effect on the protein's regulatory properties¹⁴ and these residues generate little or no electron density in X-ray crystallography.^{22,23} Therefore, these residues were also omitted from the computations. In the computations, residues 24-166 of the dimerization domain were retained in their native conformation, and only residues 7-23 were allowed to vary in conformation through unfolding and refolding.

The nonregular structure of the N-terminal arm and the paucity of its contacts with the dimerization domain pose a substantial problem in structure prediction quite similar to that of predicting the structure of a large loop. Initial experiments showed that de novo folding of the N-terminal arm from completely random structures did not yield correct structures within reasonable amounts of computation. Therefore, we used only partially unfolded structures of the arm as the starting points of folding attempts.

Each structure prediction attempt starts by partially unfolding the N-terminal arm from the starting holo or apo structure to form semi-randomized structures. The “unfolding” was performed by randomly changing the values of all 34 of the ϕ and ψ polypeptide backbone angles of the arm residues from their native values by Gaussian distributed values with a standard deviation of 20 degrees. Refolding followed the same protocol for both the holo and apo arm structures (see *SI text*). Rosetta was used to predict wild type and mutant apo and holo arm structures. The structure prediction algorithm was run 1000 times, creating 1000 candidate structures. For each of the mutant sequences, the energy score is plotted against the root-mean squared deviation (RMSD) for the arm from the starting (wild-type) structure.

Structure prediction of the mutants

We applied the unfolding-refolding protocol as described in *SI text* to 43 mutants, in each case starting from semi-randomized apo and semi-randomized holo arm structures. Table 1 shows the mutants studied, their *in vivo* regulatory properties, and the final structure prediction results. Fig. 2 shows examples of “good” and “bad” predicted arm structures of refolded Leu10 mutated to Ile overlaid with the wild type arm structure. Experiments show that this mutation does not have an effect on the arabinose response of AraC, and therefore we expect it to fold to the holo structure, as was found. Here the “good”, low energy prediction is within 0.5 Å RMSD to the native and the “bad”, high energy prediction is more than 3 Å RMSD from the native.

Comparison of *in vivo* regulatory properties and computational structure prediction results

Two basic *in vivo* regulatory properties have been measured for each of the mutants used in this study. The first is the *in vivo* expression level controlled by AraC protein of the *ara* genes in the absence of arabinose. When this level is significantly above the normal uninduced level found in wild type cells, the mutation is called constitutive. A constitutive mutation interferes with the interactions necessary to hold the DNA binding domains such that DNA looping and repression is preferred. As shown in Fig. 3, a mutant arm can produce constitutivity in two different ways. First, the mutant arm may fail to fold on the dimerization domain to the apo structure that is necessary for interacting with the DNA binding domain. Second, the arm may properly fold to the apo structure, but an altered side chain may interfere with the binding of the DNA binding domain to the dimerization domain-apo arm structure.

The second property that has been measured for the mutants is whether cells harboring the mutant AraC further induce the *ara* operon when arabinose is added. This induction by arabinose is masked however in the case of high-level constitutives because the level of *ara* enzymes is high already, and greater induction is not possible or cannot be discerned. The key point for the work being reported here is that any mutant that *in vivo* displays an arabinose response, that is, which detectably further induces the *ara* enzymes when arabinose is added, must upon calculation, be capable of folding to the holo structure. As explained above, the converse, “a mutant capable of folding to the holo structure must display an arabinose response *in vivo*”, need not be true. Such a mutant could be a high level constitutive because the arm is unstructured in the absence of arabinose and the arm could remain unstructured in the presence of arabinose.

Since the response to arabinose should exist both *in vivo* and computationally, we have compared the two. It is important to understand that only those mutants with arabinose response shown *in vivo* can be directly compared with the computational predictions. In these cases, the arabinose response is predicted computationally by the ability of the arm to refold to its native holo structure in the presence of arabinose. Reference to Table 1 shows that the mutants displaying such an arabinose response *in vivo* are D7A, P8A, P8H, P8T, L10I, L10C, P11Y, P11D, G12A, G12F, G12T, S14T, N16R, N16A, A17S, H18R, and L19Q. All but one of the seventeen mutants that display an *in vivo* arabinose response, P11Y, were predicted computationally to refold to the correct holo arm structure, (foldability as defined in Methods ≥ 7). As explained above, only these mutants can be directly compared to computational results.

In addition to computationally testing whether each of the seventeen mutants can fold to the holo crystal structure, we performed similar computational folding tests on the remaining 26 arm mutants. Out of the 43 mutant arm sequences analyzed, including those not compared to experimental activity (Fig. S5), 28 were found to refold to the holo arm structure. If there were only a random connection between displaying an arabinose response and folding to the native holo structure, the probability that out of 17 with an arabinose response, 16 of 17 would refold is 0.005 (See Supplementary Material). Thus, the high success ratio is significant.

The foldability of the arms of AraC homologs and random arm sequences

The existence of homologs of the *E. coli* AraC gene provides an additional test of arm folding. A search of the NCBI database for homologs of AraC yields more than 65,000 entries with sufficiently high similarity as to ensure highly similar tertiary structure in the regions of homology. For most, the homology covers only the DNA binding domain of AraC. There remains however, a very large number possessing significant homology to the dimerization domain as well. Clearly, some, i.e. RhaR, RhaS, MelR, XylS, do not regulate in response to arabinose. Those homologs however, which are located in their genomes immediately adjacent to genes that are also homologous to the genes coding for the *E. coli* proteins for the uptake and catabolism of arabinose almost surely are regulators that respond to arabinose. The plus arabinose structures of the N-terminal arms of many of these proteins are likely to be highly similar to the structure of the *E. coli* arm.

Table 2 shows those presumptive arabinose-responding AraC homologs whose N-terminal arm is different from the *E. coli* arm, their sequences, and their foldability.

As can be seen, most fold very well to the *E. coli* plus arabinose arm structure. To support our contention that foldability to the native AraC arm structure is biologically relevant, we also tested, Table 3, the foldability of random sequences and pseudo arm sequences based on the *E. coli* arm sequence. The pseudo arm sequences were generated by replacing each residue with a probability of 0.225 (the frequency of sequence variation in Table 2) with a randomly chosen residue. None of the random sequences show any tendency to fold to the AraC arm structure, and even the slightly randomized AraC arm sequences show a dramatically reduced foldability. We thus conclude that it is highly likely that foldability to the arm structure as calculated by our Rosetta-based algorithm is biologically relevant.

Discussion

A major challenge in the biological sciences is predicting the folded structure of an arbitrary amino acid sequence.²⁴⁻²⁸ When extraordinary computational and intellectual efforts are expended, “successful” structure predictions seem currently to be made for about 50% of the “new” sequences attempted.²⁹⁻³¹ Can comparatively modest efforts with the same machinery as is employed in structure prediction challenges be usefully applied to any other important problems? One possibility is the functional analysis of mutant proteins. This problem is related

to protein structure prediction, but is different in important ways. Many mutations produce only small and local effects and do not dramatically alter the overall structure of a protein.³² Nonetheless, some of these mutant proteins retain enzymatic activity and others lose a significant fraction of their normal activity. At present, it is very difficult in most cases to relate structural or stability changes to changes in enzymatic activity.⁸⁻¹⁰ Not surprisingly then, the few successful computational analysis of mutant proteins has been confined to calculations of the stability of the mutant proteins, not their activities.^{8,33-38}

The AraC gene regulator from *Escherichia coli* provides a unique test case intermediate between *de novo* protein structure prediction and predicting the enzymatic activity of a mutant protein. As described in the introduction, the N-terminal 23 amino acids of the AraC gene regulator from *E. coli* play a key role in determining whether AraC acts as a repressor or whether it acts as a transcriptional activator of transcription of the *ara* genes.^{5,14,17} In the repressing state, when arabinose is absent, the arm possesses one structure, the apo structure.²³ When arabinose is bound to the dimerization domain, the arm adopts another structure, the holo structure,²² and AraC shifts from repressing to inducing.

Several reasons therefore make a computational study of AraC feasible for comparison to experiment. First, it appears that only the arm, and not the rest of the dimerization domain, shifts conformation in response to arabinose. Thus, the computations required may not be excessive. Second, it is the ability of the arm to fold to the apo or holo structure that determines whether the protein represses or induces. That is, a calculated structure of the arm can be used to predict the protein's function. Third, the difference between the apo and holo structures is substantial. Fourth, a large number of arm mutants have been isolated and their *in vivo* regulatory properties have been characterized. Thus, a sizeable number of comparisons may be made between experiment and computation.

Structure prediction of the arm of AraC is challenging however, because neither the holo nor apo structures possesses regions of regular secondary structure. Normally, the structures of such regions are much more easily predicted than those of loop regions. In fact, only loop regions of thirteen residues or shorter are regularly correctly predicted.^{39,40} Furthermore, loops are structurally more constrained than the arm of AraC, because in typical unstructured loops, the location of both ends are specified, constraining the possible conformations, whereas for the arm of AraC, the position of only one end is specified.

Previous computational studies of mutant proteins have used stability calculations to predict the effect of a mutation on binding.⁴¹ We tried a similar technique in which the residue of interest was altered, and the energy of the protein was minimized and noted. This scheme gave poor results; only 2 of 17 mutants' regulatory behavior were correctly predicted. Most likely the poor performance of this method results from the fact that it is the energy differences between the two folded states and unfolded state of the AraC arm that determines its behavior, not the free energy difference between the wild type and mutant structures.

Before predicting the folding properties of mutant arms, it was necessary to show that a computational algorithm could correctly fold the arm of the wild type sequence to its correct structures. In the absence of arabinose, the arm should fold to the known apo structure, and in the presence of arabinose, the arm should fold to the known holo structure. Preliminary experiments showed these goals were achievable when the structures from which folding operations began were not too distant from the holo or apo structures and when the remainder of the dimerization domain was present. In contrast to the situation for the holo structure, reconstruction of the apo structure did not require the presence of arabinose. It is interesting to note that in the crystal lattices from which the apo structure was determined, the arm makes

one intra-subunit interaction. Computational reconstruction of the apo arm structure did not, however require the presence of this interaction.

Attempting to refold a partially unfolded arm did remarkably well in predicting the arabinose response of mutant sequences. A total of 43 different mutants were examined for their ability to fold to the apo structure in the absence of arabinose and the holo structure in the presence of arabinose. The *in vivo* regulatory behavior, that is, the level of expression of the *ara* genes induced by the mutant proteins in the absence and presence of arabinose, had also been measured. Because several different folding behaviors of the arm can produce the same constitutive *in vivo* regulatory response, the structure predictions of only 17 of the mutants could be directly compared to the *in vivo* regulatory behavior. The remaining 26 mutants were used as controls to ensure that the computational prediction algorithm was not biased towards always folding to the correct structure. Those for which the computational results can be directly compared to experimentally measured behavior are the mutants that show further induction upon the addition of arabinose. Such mutant proteins ought to fold to the holo structure in the presence of arabinose. Sixteen of the seventeen mutants displayed this behavior in the computations. If there were no connection between arabinose inducibility and the correct arm folding, the probability that by chance, 16 or 17 would have been correctly predicted is 0.005.

The close correspondence that we found between the computational predictions and the *in vivo* regulatory behavior of AraC demonstrates that in special situations structure prediction can now be used to answer questions beyond determining the specific fold of a particular sequence. In our case, the energy differences between the two folded states of the N-terminal arm of AraC are small. AraC controls gene activity over a dynamic range of about 300, that is, the induced level of gene activity is around 300 times the uninduced level. The free energy differences between the “off” and “on” states are then about $RT\ln(300) = 3.4$ kcal/mol, considerably less than the 10 kcal/mol that are typical for the stabilities of folded proteins.⁷ AraC is, in fact, dimeric, and depending on the details of DNA looping, the 3.4 kcal/mol may be shared equally between the protein’s subunits.

We would expect that most homologs of AraC that also control gene expression in response to arabinose would possess N-terminal arms that fold to close to the same structure as the *E. coli* N-terminal arm. This is so. On the other hand, random sequences of the same length as the arm ought not to possess any tendency to fold to the plus arabinose structure, and indeed, they do not. Finally, and also as expected if the plus arabinose arm structure has biological relevance and our algorithm is correctly predicting this structure, at a low probability, randomly substituting residues of the *E. coli* arm sequence with randomly chosen residues yields arms with considerably reduced tendency to fold.

Altogether, our findings show that the refoldability, when performed in the context of the remainder of the AraC protein, of a sequence to the known structure of the N-terminal arm of the *E. coli* AraC protein possesses a strikingly high correlation to the sequence’s biological activity. In short, we found that computational prediction of folded structure can be applied to problems in which the energy differences between structures is relatively small. Despite the success on predicting the behavior of AraC, considerable improvement in computational analysis of structure-function is still necessary. Only in rare cases is it likely that dramatic changes in structure occur and that understanding of the mechanism is sufficiently advanced to relate such structural changes to activity changes. In our case prediction failed one time out of seventeen, and prediction accuracy is likely to be lower in harder problems.

An interesting related question is that posed by the class of mutants in which our algorithm predicts an absence of apo arm structure. Superficially, such mutants ought to be unable to

repress in the absence of arabinose, and therefore will activate transcription in the absence of arabinose (be constitutive). Therefore, in addition to using the algorithm to fold all the mutant arms to the holo structure in the presence of arabinose, we also used the algorithm to fold the arms to the apo structure in the absence of arabinose. One mutant which we predicted not to form the apo structure was found experimentally however not to be constitutive. One attractive possibility is that when the DNA binding domain is present, as it is *in vivo* but not in the structure predictions of this mutant, its presence helps the arm to fold. Hence, in the absence of the DNA binding domain, a situation analogous to what was calculated, the arm ought to be unfolded. This prediction is experimentally testable.

The ability of computational methods to refold the two arm structures of AraC strongly suggests that they are biologically relevant. Because the arm is not an integral part of a globular domain, it could legitimately be feared that crystal lattice interactions might have forced the arm into positions or structures dissimilar to those normally adopted by the arm. The structure prediction results presented here diminish concerns of this nature that the crystal structures are misleading.

The high fraction of mutants successfully predicted in this study not only indicates that computational methods can be applied to the analysis of mutant behavior, but it also means that the mechanistic model of AraC which was used, the light switch mechanism, is likely to encompass the major part of the protein's regulatory behavior. Specifically this means that our protocol of holding the structure of residues 24-166 fixed and allowing only residues 7-23 to vary provides a close approximation to the behavior of the regulatory essentials of the protein. The high success also suggests that careful biochemical and computational analysis of the single failure out of 17 cases in which the computational prediction did not match behavior may yield deeper understanding of the biochemistry and/or the computations.

In summary, we have shown that computational structure prediction is highly successful in predicting *in vivo* regulatory behavior of AraC mutants and of related sequences. In two cases of a superficial discrepancy between predictions and experimental results, it is possible that interactions between the DNA binding domain and the arm of AraC stabilize the arm structure. This possibility can be directly tested with laboratory experiments.

Materials and Methods

The presence of a ligand during the calculations with the holo structure are modeled after the RosettaLigand protocol.⁴² In the case of N-terminal arm folding, the ligand is parameterized and input from a file and remains fixed for the duration of the simulation. The scoring function takes into account the atoms in the ligand for energy calculations.^{43,44} The root mean square deviation (RMSD) is calculated over all C α atoms in the protein as described⁴³.

Simulations for structure prediction of the N-terminal arm start from the crystal structure of either the holo or apo conformation, in which idealized bond lengths and angles have been substituted⁴³ and side chains are replaced with rotamer approximations.⁴⁵ Mutants were generated by appropriate alteration of the side chain. These preprocessed native crystal structures (pdb codes 2arc²² and 1xja,²³ holo and apo respectively) are input into Rosetta and the arm is unfolded by varying each of the 34 ϕ and ψ angles of the arm from their native values by Gaussian distributed values with a standard deviation of 20 degrees.

From the unfolded structure, refolding is achieved by altering the ϕ and ψ coordinates of only residues 7-23 of the protein during a standard Rosetta conformation search. Additionally, the rotamers of these N-terminal residues and those within 10 Å of them are also altered. As indicated in the figure, the three steps are: 1) make a move, 2) minimize energy by variation of relevant degrees of freedom, and 3) accept all move+minimization operations that lower the energy and accept moves that increase the energy with a probability P given by the Boltzman

distribution, $P = e^{-\Delta E/kt}$ where ΔE is the energy difference between the current and trial conformations, k is the Boltzman constant, and T is the temperature.

Refolding combines “small moves” (ϕ and ψ of residue i are perturbed), “shear moves” (ϕ of residue $i-1$ and ψ of residue i are perturbed in equal magnitudes and opposite directions), Davidon-Fletcher-Powell (DFP) conjugate gradient minimization,⁴⁶ and line minimization⁴⁶ in a procedure that yields diverse sampling. Fig. S1 shows the order and number of cycles used during the refolding procedure. The complete protocol is used 1000 times, resulting in 1000 candidate structures for examination. With current 3 GHz processors, generation of 1000 folded structures requires 5 CPU-days.

Foldability is defined as the number of the 10 lowest energy refolded structures that are within 0.5 Å RMSD of the original structure. A sequence is considered to refold correctly if its foldability is 7 or greater. Initially, two of the authors separately did blind examinations (no knowledge of the experimental activity for any of the mutants) of the plots to determine which mutants could be considered to correctly fold to the native structure. These subjective classifications showed striking agreement, and the objective measure of considering an arm to fold if its foldability is 7 or greater most closely paralleled the subjective classifications.

The protocol for structure prediction of the N-terminal arm of AraC has been incorporated into the Rosetta software suite and generalized for use with any protein of known structure for which n residues of the wild type or mutant sequence at the N-terminus are to be partially unfolded and then refolded a specified number of times. The source code for this work will be available through Rosetta Commons at no cost to academic researchers in the next release of Rosetta.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Brian Kuhlman, George Rose, and Marc Ostermeier for their comments and discussion on the manuscript. This work has been supported by a Ruth Kirschstein Graduate Research Fellowship to MB (GM081901) and NIH grants GM018277 (RS) and GM078221 (JJG). JJG is a Beckman Young Investigator.

References

1. Dahiyat BI, Sarisky CA, Mayo SL. De novo protein design: towards fully automated sequence selection. *J Mol Biol* 1997;273:789–96. [PubMed: 9367772]
2. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. *Science* 1998;282:1462–7. [PubMed: 9822371]
3. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retroaldol enzymes. *Science* 2008;319:1387–91. [PubMed: 18323453]
4. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science* 2003;302:1364–8. [PubMed: 14631033]
5. Rodgers ME, Holder ND, Dirla S, Schleif R. Functional modes of the regulatory arm of AraC. *Proteins* 2009;74:81–91. [PubMed: 18561170]
6. Street AG, Mayo SL. Computational protein design. *Structure* 1999;7:R105–9. [PubMed: 10378265]
7. Robertson AD, Murphy KP. Protein Structure and the Energetics of Protein Stability. *Chem Rev* 1997;97:1251–1268. [PubMed: 11851450]
8. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004;2004:pl2. [PubMed: 14872095]

9. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng* 1999;12:549–55. [PubMed: 10436080]
10. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 2002;320:369–87. [PubMed: 12079393]
11. Harmer T, Wu M, Schleif R. The role of rigidity in DNA looping-unlooping by AraC. *Proc Natl Acad Sci U S A* 2001;98:427–31. [PubMed: 11209047]
12. Lobell RB, Schleif RF. DNA looping and unlooping by AraC protein. *Science* 1990;250:528–32. [PubMed: 2237403]
13. Lobell RB, Schleif RF. AraC-DNA looping: orientation and distance-dependent loop breaking by the cyclic AMP receptor protein. *J Mol Biol* 1991;218:45–54. [PubMed: 1848302]
14. Saviola B, Seabold R, Schleif RF. Arm-domain interactions in AraC. *J Mol Biol* 1998;278:539–48. [PubMed: 9600837]
15. Seabold RR, Schleif RF. Apo-AraC actively seeks to loop. *J Mol Biol* 1998;278:529–38. [PubMed: 9600836]
16. Carra JH, Schleif RF. Variation of half-site organization and DNA looping by AraC protein. *EMBO J* 1993;12:35–44. [PubMed: 8428590]
17. Ross JJ, Gryczynski U, Schleif R. Mutational analysis of residue roles in AraC function. *J Mol Biol* 2003;328:85–93. [PubMed: 12683999]
18. 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. [PubMed: 6392569]
19. Wu M, Schleif R. Strengthened arm-dimerization domain interactions in AraC. *J Biol Chem* 2001;276:2562–4. [PubMed: 11069917]
20. 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 U S A* 1984;81:5017–20. [PubMed: 6089170]
21. 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 U S A* 1986;83:3654–8. [PubMed: 3520549]
22. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. Structural basis for ligand-regulated oligomerization of AraC. *Science* 1997;276:421–5. [PubMed: 9103202]
23. Weldon JE, Rodgers ME, Larkin C, Schleif RF. Structure and properties of a truly apo form of AraC dimerization domain. *Proteins* 2007;66:646–54. [PubMed: 17173282]
24. Baker D. A surprising simplicity to protein folding. *Nature* 2000;405:39–42. [PubMed: 10811210]
25. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 1997;48:545–600. [PubMed: 9348663]
26. Onuchic JN, Wolynes PG. Theory of protein folding. *Curr Opin Struct Biol* 2004;14:70–5. [PubMed: 15102452]
27. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci U S A* 1995;92:3626–30. [PubMed: 7724609]
28. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? *Curr Opin Struct Biol* 2007;17:342–6. [PubMed: 17572080]
29. Cozzetto D, Kryshchuk A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. *Proteins* 2007;69(Suppl 8):175–83. [PubMed: 17680695]
30. Kryshchuk A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. *Proteins* 2005;61(Suppl 7):225–36. [PubMed: 16187365]
31. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. *Proteins* 2001;(Suppl 5):163–70. [PubMed: 11835494]
32. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5:823–6. [PubMed: 3709526]
33. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. *PLoS Comput Biol* 2007;3:e164. [PubMed: 17722975]

34. Wooll JO, Wrabl JO, Hilser VJ. Ensemble modulation as an origin of denaturant-independent hydrogen exchange in proteins. *J Mol Biol* 2000;301:247–56. [PubMed: 10926507]
35. Bueno M, Camacho CJ, Sancho J. SIMPLE estimate of the free energy change due to aliphatic mutations: superior predictions based on first principles. *Proteins* 2007;68:850–62. [PubMed: 17523191]
36. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res* 2006;34:W239–42. [PubMed: 16845001]
37. Parthiban V, Gromiha MM, Abhinandan M, Schomburg D. Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. *BMC Struct Biol* 2007;7:54. [PubMed: 17705837]
38. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins* 2007;66:41–52. [PubMed: 17068801]
39. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of Protein Loop Conformations using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J Chem Theory Comput* 2008;4:855–868. [PubMed: 18787648]
40. Zhu K, Pincus DL, Zhao S, Friesner RA. Long loop prediction using the protein local optimization program. *Proteins* 2006;65:438–52. [PubMed: 16927380]
41. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. *Structure* 2007;15:1567–76. [PubMed: 18073107]
42. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* 2009;385:381–92. [PubMed: 19041878]
43. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. *Methods Enzymol* 2004;383:66–93. [PubMed: 15063647]
44. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem* 2008;77:363–82. [PubMed: 18410248]
45. Dunbrack RL Jr. Rotamer libraries in the 21st century. *Curr Opin Struct Biol* 2002;12:431–40. [PubMed: 12163064]
46. Press, WH. *Numerical recipes in C++ : the art of scientific computing*. 2nd edit. Cambridge University Press; New York: 2002.

Fig. 1.

Light-switch mechanism of action of AraC protein. In the absence of arabinose, the N-terminal arm is in one position and AraC binds to the I_1 and O_2 half-sites and forms a loop that represses expression of the genes. When arabinose is added, the arm is in an alternate position and the loop opens as AraC binds to I_1 - I_2 from which it activates transcription. AraC consists of nearly independent DNA binding domains and dimerization domains and an N-terminal arm that controls the orientational freedom of the DNA binding domains in response to arabinose.

Fig. 2. Arm plus dimerization domain (native, cyan) showing good (lilac) and poor (blue) arm refolding attempts. The arabinose molecule is depicted using spheres

Fig. 3. Two mechanisms for producing constitutivity, the arm does not fold on the dimerization domain as the wild type arm does, and the arm folds, but the folded arm does not bind the DNA binding domain.

Table 1

Mutants of the N-terminal arm, and their regulatory and foldability.

Mutation	–Ara level ^a	+Ara level ^a	Holo foldability ^b	Apo foldability ^b
Wild Type	<1	100	10	10
D7A	<5	20	8	10
P8A	<2	150	9	10
P8G	100	200	6	9
P8H	30	200	8	10
P8T	25, 100	100	10	10
P8V	<5	<5	10	10
L9A	90	90	6	10
L9C	200	350	6	10
L9D	150	150	3	8
L9G	250	150	6	9
L9H	300	300	8	10
L9P	100	200	9	1
L9S	200	200	5	9
L10A	100	150	9	10
L10C	10	150	8	9
L10D	200	200	9	9
L10I	10	100	9	5
L10N	200	200	10	9
L10R	150	200	10	9
P11D	50	300	8	10
P11R	200	200	10	10
P11Y	30	100	4	10
G12A	10	200	8	9
G12F	<2	200	8	10
G12T	50	400	8	10
Y13E	100	80	2	10
Y13K	120	120	2	10
Y13P	200	100	1	10
Y13S	100	100	6	10
Y13T	100	100	3	10
S14A	100	150	10	10
S14E	100	100	7	9
S14L	200	100	6	6
S14T	10	100	10	10
F15A	<5	<10	4	10
N16A	5	200	10	10
N16R	<2	25	9	8
N16T	100	200	9	7
A17S	1	100	9	10

Mutation	–Ara level ^a	+Ara level ^a	Holo foldability ^b	Apo foldability ^b
H18L	100	100	9	10
H18R	5	110	10	10
L19Q	5	100	7	10
V20G	100	100	5	8

^aExpressed as per cent of fully induced wild type L-arabinose isomerase level.

^bFoldability as defined in Methods.

Table 2

The sequences and foldability of N-terminal arms on presumptive arabinose-responding AraC homologs.

Bacterial strain	NCBI similarity score to <i>coli</i> AraC	Number residues identical to <i>E. coli</i> AraC	Sequence of N-terminal arm ^a	Holo foldability ^b
<i>Escherichia coli</i>	611	292/292	DPLLPGYSFNAHLVAGLT	10
<i>Klebsiella pneumoniae</i> 342	538	266	DPLLPGYSFNAHLVtGLT	6
<i>Pectobacterium carotovorum</i>	321	202	nPLLPGYSFNAyLVAGLT	10
<i>Dickeya dadantii</i> Ech703	318	198	nPLLPGYafNAyLVAGLT	10
<i>Yersinia pseudotuberculosis</i> IP 32953	318	162	nPLLPGYtFNAYLVAGLT	10
<i>Aeromonas salmonicida</i>	313	197	nPLLPGYafdvfLVsGmT	6
<i>Leptothrix cholodnii</i>	303	195	nPLLPGYSFNlyLVAGLT	10
<i>Vibrio furnissii</i>	258	177	DPLkPGYnFdAHLVAGLT	8
<i>Actinobacillus succinogenes</i>	255	124	nPLLPGYqFgsyLVAGcT	8
<i>Reinekea sp.</i> MED297	250	175	DPLkPGYpFNAHLVAGiT	10
<i>Pantoea sp.</i> At-9b	229	167	sPLmrrftFNAfmVsGfT	10

^a Uppercase residues are same as *E. coli* arm sequence, lower case residues are different.

^b Foldability as defined in Methods.

Table 3

The foldability of arms with sequences different from that of the N-terminal arm of *E. coli* AraC. The first six random sequences Ranaa, each residue was chosen at random from the 20 amino acids. The sequences of the final twelve, Sim, are the *E. coli* sequences in which each residue, with a probability of 0.225 was replaced with a randomly chosen residue from the 20 amino acids.

Identifier ^a	Sequence ^b	Holo foldability ^c
Ranaa1	ifaetfktngpimiyaLT	0
Ranaa2	iikpiatmneptaygfLT	0
Ranaa3	tfiaiafmftenikpypT	0
Ranaa4	DyettqhmvpkacswnLT	0
Ranaa5	nqfshevatpcykmwLT	0
Ranaa6	anktsqqvmweHLfyedT	0
Sim1	DPkLwGYSFNAcLsAiLT	2
Sim2	DPLnkGYSFNAmLVAGfT	7
Sim3	DdLLPGYkFNAHLVAGLT	8
Sim4	DwLePGYmeNAHLVmtrT	0
Sim5	DPLwPGYSFNAHLVAaLT	7
Sim6	DvLLIGYSFNAHLVAGLT	9
Sim7	nqLLPfYrFNAILVAGLT	2
Sim8	DPLwdGYgFNAHLssGLT	6
Sim9	DPLLPfYSFNAHLVAGLT	7
Sim10	cPpLPGYIFNAHLVAGLT	9
Sim11	DPLLPGISFNpHLVArLT	0
Sim12	DPyLPGISFNpHLVArLT	0

^a Sequences Ranaa1-3 generated using Protein Sequence Randomizer with a random input sequence (http://www.cellbiol.com/scripts/randomizer/sequence_randomizer.html). Ranaa 4-6 generated using RandSeq with equal composition for all amino acids (<http://www.expasy.ch/tools/randseq.html>).

^b Residues in uppercase are the same as in the *E. coli* N-terminal arm, those in lowercase are different.

^c Foldability is as defined in Methods