EcoSal Plus
Cellular and Molecular Biology of
E. coli, Salmonella, and the *Enterobacteriaceae*

DOMAIN 5 RESPONDING TO THE ENVIRONMENT

A Career's Work, the L-Arabinose Operon: How It Functions and How We Learned It

ROBERT SCHLEIF[a]

[a]Biology Department, Johns Hopkins University

ABSTRACT Very few labs have had the good fortune to have been able to focus for more than 50 years on a relatively narrow research topic and to be in a field in which both basic knowledge and the research technology and methods have progressed as rapidly as they have in molecular biology. My research group, first at Brandeis University and then at Johns Hopkins University, has had this opportunity. In this review, therefore, I will describe largely the work from my laboratory that has spanned this period and which was carried out by 40 plus graduate students, several postdoctoral associates, my technician, and me. In addition to presenting the scientific findings or results, I will place many of the topics in scientific context and, because we needed to develop a good many of the experimental methods behind our findings, I will also describe some of these methods and their importance. Also included will be occasional comments on how the research community or my research group functioned. Because a wide variety of approaches were used throughout our work, no ideal organization of this review is apparent. Therefore, I have chosen to use a hybrid structure in which there are six sections. Within each of the sections, experiments and findings will be described roughly in chronological order. Frequent cross references between parts and sections will be made because some findings and experimental approaches could logically have been described in more than one place.

KEYWORDS biophysics, function, gene regulation, genetics, mechanisms of action, molecular biology, structure

Received: 3 March 2021
Accepted: 20 July 2021
Published: 18 August 2021
Editor Susan T. Lovett, Brandeis University
Citation: Schleif R. 2021. A career's work, the L-arabinose operon: how it functions and how we learned it. EcoSal Plus 9:eESP-0012-2021. https://doi.org/10.1128/ecosalplus.ESP-0012-2021.
Address correspondence to Robert Schleif, rfschleif@jhu.edu.

ORIGIN OF STUDY ON *ARA*, GENETICS, OPERON, AND THE REGULATORY REGION STRUCTURE

To facilitate understanding of the material which follows, Fig. 1 shows a sketch of the *araCBAD* regulatory region that contains the promoters *pBAD* and *pC*. DNA looping, which is described later, occurs when the dimeric AraC protein binds to *araO2* and *araI1*.

Work on *ara* began as a genetic exercise in a summer course at Cold Spring Harbor and then was continued at Santa Barbara by Ellis Englesberg, who mapped the genes *araCBAD* (1). Figure 2 shows the genes that he mapped plus the additional *ara* gene clusters and their promoters that are now

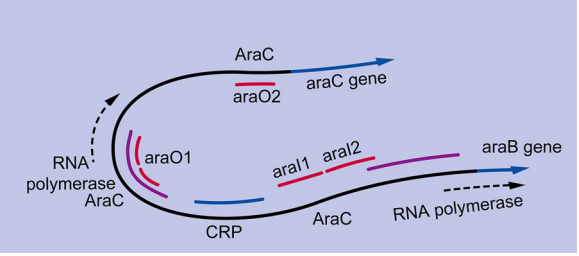

FIG 1 The regulatory region of *araCBAD* located between *araC* and *araBAD*. Shown are the binding sites for RNA polymerase (purple lines), AraC (red lines), and CRP (blue line), as well as the transcription start sites for *pBAD* and *pC*.

known. Englesberg determined the biochemical properties of the genes *B*, *A*, and *D* and, in addition, found that the activity of the *araC* gene was required for expression of the others (2). While this suggests that *araC* is a positive regulatory gene or element and is required for the expression of the other genes, this possibility was strongly resisted by the scientific community. At that time, only negatively acting repressor genes had been demonstrated, notably, the lambda phage *C1* gene and the *lacI* gene (3). Critics skeptical of the existence of positive regulatory genes argued that the AraC protein could be part of an active transport system and without active AraC, inducing levels of arabinose inside cells could not be achieved. Alternatively, they argued that AraC could merely be one part of a double-negative system, so that mutations in AraC left the other repressor fully expressed which then blocked expression of the other genes of the arabinose system. Later, while I was a graduate student, Englesberg published strong evidence that *araC* was a regulator that acted both positively and negatively (4). At about the same time, Gilbert published his detection of the Lac repressor protein, LacI, with an equilibrium dialysis assay (5). Up to this time, multiple efforts to detect the *lacI* gene product with a biochemical assay had been unsuccessful, and thus Gilbert's achievement was big news. I therefore decided to go to Gilbert's lab as a postdoctoral associate and isolate the AraC protein. There, I would determine whether or not AraC actually was a positive regulator as well as determine its mechanism of action. The project turned out to be somewhat harder than expected and, instead of taking 3 years, has required about 50 years.

A later section of this review contains a description of my initial and unsuccessful attempt at detecting AraC in Gilbert's lab and then the later successful assays.

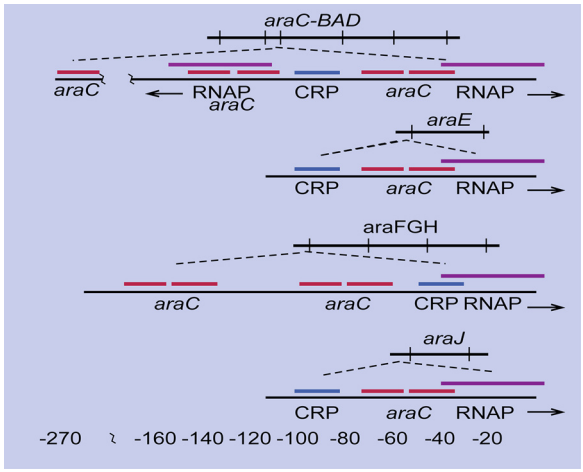

FIG 2 The AraC-responsive genes *araCBAD*, *araE*, *araFGH*, and *araJ* and the structures of their regulatory regions showing the binding regions for RNA polymerase, AraC, and CRP. The gene sizes are not to scale, but the protein binding sites and their locations are approximately to scale. The scale is in base pairs upstream from the start of transcription. Sequence analysis suggests that in the *araE* regulatory region there are two additional binding sites for AraC, about 150 bp upstream of the site shown, and in the *araJ* regulatory there are up to five AraC half-sites around 100 bp upstream of the site shown.

Here, I continue to describe the genetic aspects of the story. Soon after my experiments began, I needed mutants in the various *ara* genes for verification that an activity that I was detecting was AraC and I needed to replicate in *Escherichia coli* strain K-12 a good portion of the genetic work that had been done in strain B/r. In the end, this turned out to be beneficial, since B/r is not amenable to work with bacteriophage lambda, a phage that soon came to be an important tool in our research.

The hint that motivated our work and ultimately led to our discovery of DNA looping. One of the multiple lines of experiments that Englesberg developed in efforts to demonstrate that AraC was a positive regulator was to utilize deletions of AraC. Two that he isolated were of particular interest, Δ766 and Δ719 (6) (Fig. 3). Both seemed to eliminate *araC* activity and to end upstream of the *B*, *A*, and *D* genes, ending either in *araC* or between *C* and *B*. In these deletions and in the absence of AraC, these genes were expressed at near their uninduced, or basal, levels. However, when AraC was provided in *trans* from an episome, the chromosomal *B*, *A*, and *D* genes in both deletions were induced normally, by around 500-fold, by arabinose. The unexpected finding was that when AraC was provided but

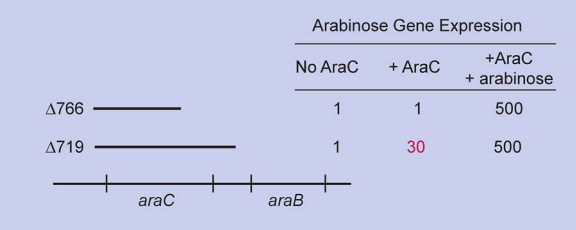

FIG 3 Schematic representation of the distances that the two Englesberg deletions, Δ719 and Δ766, enter the arabinose operon, along with a table indicating the levels of transcription of the *araBAD* genes in the absence of AraC protein provided in *trans* from an episome introduced into the cells, in the presence of AraC, and in the presence of AraC and arabinose.

arabinose was absent, in Δ719 only the genes *B*, *A*, and *D* were elevated about 30-fold above their basal level. A number of fairly trivial and uninteresting explanations for this elevated basal level were possible. A more interesting possibility, as suggested by Englesberg, was that both repressing and inducing conformations of AraC coexisted, even in the absence of arabinose, and that, in this absence, the repressing species normally prevented the inducing species from acting. Then, in the presence of arabinose, the near disappearance of the repressing species combined with increased levels of the activating species would lead to the observed 500-fold induction of *araBAD*. In order for this inducer-repressor idea to explain the data, at least a little bit of a site required for repression had to lie upstream from all the sites or bases required for full induction. This possibility seemed intriguing and potentially important to me, and in need of testing. This would require the isolation of multiple deletions of the nature that Englesberg had first isolated. This ultimate goal motivated the following: first, the construction of a fine structure deletion map as described in the next section, and then, from this, the determination of whether any or how many of the deletions possessed the behavior of the Englesberg deletion Δ719. This latter objective is described more fully in next section.

Another result suggestive of repression in the *ara* system was the response of the *araBAD* genes' expression in the simultaneous presence of two types of AraC protein. Constitutive mutations, C(con), retain AraC protein's capability of stimulating high expression levels of AraBAD products, but no longer require arabinose to do so. Cells containing both wild type (C^+) and C(con) genes express *ara* genes in the absence of arabinose at close to the basal level rather than at levels characteristic of the C(con) protein (4). That is, C^+ is dominant to C(con). While this appears to be consistent with a repressing activity of C^+ protein, it is not strong evidence. One weakness is that AraC could be oligomeric, and the activity of a mixed-subunit hetero-oligomer (if they existed) was unpredictable.

Fine-structure deletion map of *araCBA*. While I was a postdoc with Gilbert, and for about the next 10 years, the biology of bacteriophage lambda was under intensive study by the molecular biology community. The lambda phage community in the 70's and 80's was very open, cooperative, and supportive. This attitude, which arose from the leaders in the field, led to significant progress. As I have been involved in several different fields through my career, I have found that this type of cooperativity is not always the case and when the leaders in the field do not establish this tone, the field suffers.

Much of the lambda phage work was being done around me at Harvard in Cambridge, at the Harvard Medical School in Boston, and also at MIT. Both the lambda phage people and others interested in molecular biology met annually at the end of the summer at Cold Spring Harbor Labs to exchange results. This rich environment, plus the fact that Gilbert had increased the intracellular level of Lac repressor by infecting cells with a bacteriophage carrying the *lacI* gene, primed me to use phage lambda in my quest for AraC and an understanding of how it worked.

With the region of DNA involved in the regulation, at least in the promoter's induction, being located between the *araC* and *araB* genes, it was clear that a careful study of the biochemistry of gene regulation would require the isolation of the regulatory region DNA and subportions of the regulatory region lying between the two genes. As genetic engineering was then only beginning, two initial steps seemed sensible. The first was to determine whether Englesberg's hint was true, that repression of the *araBAD* genes could occur from upstream of the promoter and all the DNA that was required for its induction. The second was to generate a detailed, high-resolution genetic map of the *ara* genes. Lambda phage seemed like a useful tool in the pursuit of both of these goals.

At the time, it had just been discovered that infection of *Escherichia coli* that is deleted of the normal lambda phage chromosomal attachment site can give rise to very rare lysogens in which the phage has inserted into pseudo *att* sites elsewhere in the chromosome. One such insertion of the heat-inducible phage λCI857 was found to be in the leucine operon, which is close to the *ara* genes (7). Because the phage was not inserted into its normal site, excision from the pseudo site was inefficient. Among the cells surviving the phage induction could be those from which the phage had been deleted before the phage was induced, and those from which the phage excised from the genome without killing its host. I found that a small fraction of the surviving cells were *ara* negative and contained deletions from *leu* into or through *ara*. A Luria-Delbruck fluctuation test (8) showed that at least some of these resulted from deletions that existed in the culture before the heat induction step. The only instance that I knew of where the fluctuation test had been used was in the original Luria-Delbruck paper that had been published 30 years earlier. Therefore, I wrote to Luria that he might be interested to know that the fluctuation was still finding important use. He immediately wrote back something to the effect, "Of course it's still useful, it's a fundamental tool!" From about 4×10^{11} cells, a small number of survivors contained a chromosomal deletion extending from *leu* and ending in *ara*. Such rare cells could be identified by then selecting first for *ara*-negative cells by utilizing the arabinose sensitivity of Ara^+ cells to the presence of ribitol and then genetically screening candidates for the presence of the *araD* gene. Over the course of several months of continuous work, I isolated over 300 deletions from *leu* into *ara*. I also isolated over 250 *ara* point mutations, although this latter task was much easier and took only a week.

The deletions had been isolated in female cells and the point mutations were isolated on episomes. This facilitated performing genetic crosses between the deletion and point mutation strains to determine whether or not a deletion strain contained wild type sequence at the position of the point mutation. By this means, I constructed a detailed map of the point mutations and deletions (9). This map defined 25 segments within *araC* and 34 segments with *araB*. Additionally, a sizeable number of the deletions ended between *araC* and *araB*, but no point mutations mapped to this region. A very large number of genetic crosses were required to generate the map, and I developed methods for performing up to 1,500 crosses per day. This was before the existence of Eppendorf microcentrifuge tubes, microtiter trays, or multipipettors, and doing the crosses required the construction of some specialized but simple equipment in the departmental machine shop.

A statistical analysis of the distribution of the point mutations and deletions in the map indicated that at least the great majority of the deletion endpoints were randomly distributed. The sections below detail how we verified the existence of the upstream repression site in *ara* and used the phage to predict its position, and also how we utilized the map and additional lambda-*ara* phage to develop a means to isolate the *ara* regulatory region and subportions of it. This was done while genetic engineering was being developed and before that technology greatly simplified mapping and the isolation of specific pieces of DNA.

Genetic and physical mapping of the site required for repression. The fine structure genetic map that is described in the previous section was augmented with the isolation and mapping of an additional 150 point mutations and the insertion sites of more than 100 Mu phage (9, 10) (Phage Mu integrates into random sites throughout the chromosome.). Again, no point mutations were found in the *ara* regulatory region, suggesting that more than a single base change was required to reduce *ara* expression to the point that a mutant with such a mutation would be selected by the methods we used. Therefore, only deletion frequencies could be used to estimate distances in the region between *araC* and *araB*. The deletions of interest extended from *leu* and appeared to have lost the *ara* upstream repression site, but retained the induction site. The number of such deletions compared to the total suggested that the site was 200 to 400 bp upstream of the elements required for full expression in the presence of arabinose. Although the global distribution of deletion sites did seem close to random, such a mapping method is hardly rigorous. We needed a physical map of the regulatory region to definitively locate the upstream repression site or, more correctly, the upstream edge of this repression site.

To generate a physical map of the *ara* regulatory region, we again utilized phage lambda. This time we selected for the very rare insertions of lambda into *ara* genes (11). More than 100 large-scale attempts yielded one

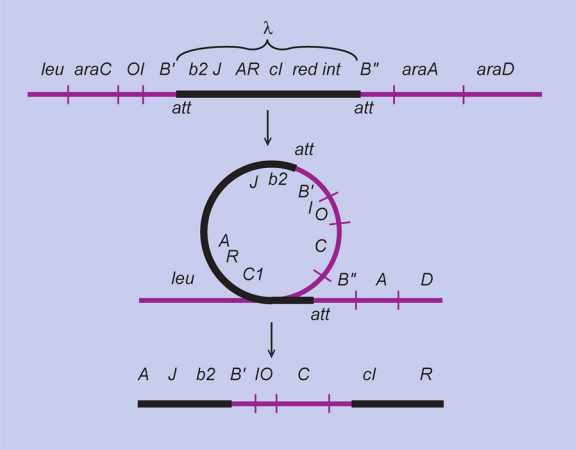

FIG 4 Creation of λ-*ara* transducing phage during improper excision from a pseudo attachment site within the *araB* gene. When the phage DNA enters a cell, the ends near the phage *A* and *R* genes join. The phage integrates into the chromosome with a crossover between its *att* site and a bacterial *att* site, thus essentially turning the phage inside out. Normal excision and an anomalous excision that produces a transducing phage reverse this process. As a result of an anomalous excision event, bacterial DNA replaces phage DNA to one side of the phage *att* site or the other.

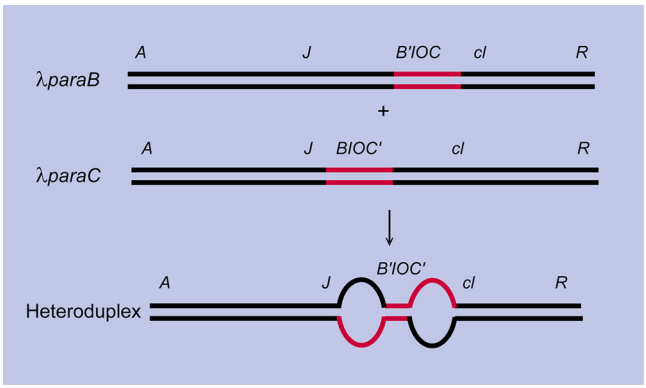

FIG 5 Double-bubble complex for the isolation of *ara* regulatory region DNA. Equal quantities of the two λ-*ara* transducing phage are mixed, denatured, and allowed to reanneal. Half the renatured DNA will be in heteroduplexes of the form shown, where the short *ara* region comprising a bit of *araC* and of *araB* flanking the regulatory region shown here as *IO* will be double stranded. It is flanked by the single-stranded regions. Digestion with the single-strand specific nuclease S1 leaves the two very long arms of λ and the short regulatory region.

insertion in *araC* and eight insertions, apparently all at the same point and in the same orientation, in *araB*. From these, we then isolated the very rare *ara* transducing phage that had excised improperly and replaced some phage genes with bacterial DNA sequences lying adjacent to the phage insertion sites (Fig. 4). As described below, we developed large-scale methods for the isolation of the *pBAD* regulatory region from DNA heteroduplexes formed from λ*paraB* and λ*paraC* phage (Fig. 5). At this time, the most efficient way to determine the locations and extents of such substitutions was by electron microscopy of DNA heteroduplexes between a reference phage and a phage with an insertion to be mapped. We therefore learned how to do the microscopy and then determined the sizes and locations of our substitutions. The microscopy was augmented by agarose gel electrophoresis of DNA fragments, which at the time was just being developed for size determination of DNA molecules. We were particularly interested in the sizes of fragments remaining after digestion of the heteroduplexes with the single-strand specific nuclease S1. To close in on the regulatory region and the location of the upstream edge of the repression site, we isolated deletions in the phage, mapped the deletion endpoints against the point mutations, and determined sizes of regions, again by electron microscopy and gel electrophoresis.

The isolation of the phage deletions was interesting. The phage DNA is rather tightly packed inside the phage head and the mutual repulsion between the charged phosphates of the DNA is neutralized by magnesium ions. If some are removed, the tightly packed DNA expands, bursts the phage head, and inactivates the phage. Thus, the selection of phage resistant to weak magnesium-chelating agents like pyrophosphate selects for phage containing deletions. Plating the candidate phage on *ara* minus cells on special indicating plates containing pyrophosphate yielded plaques in which deletions could be readily identified (11).

Isolating physical amounts of the regulatory region of *pBAD*, PEG size fractionation of DNA. To isolate physical quantities of *ara* regulatory region DNA, it was necessary to scale up the analytical methods that were described in the previous section (11). Phage was purified from heat-induced 16-liter cultures of lysogens of phage. Rather than letting the induced lambda phage lyse the cells, we used phage with a mutant in the S gene so that cells do not lyse until the cell membrane is dissolved with chloroform. This allowed the cells to be harvested by low-speed centrifugation and lysed when highly concentrated. The phage were further purified by centrifugation through block gradients of CsCl. After forming double-bubble DNA heteroduplexes from the two DNAs (Fig. 5), the single-stranded DNA was digested

with S1 nuclease that we had partially purified. After the digestion, the regulatory region or portions of it were on DNA duplexes of 1,000 or less base pairs and the two arms that were over 20,000 bp in length. We needed a simple large-scale preparative method for separating the two sizes of DNA fragments. We had read in a recent journal article that polyethylene glycol (PEG) can be used to precipitate viruses. John Lis in my lab noticed that the concentrations required to precipitate highly asymmetric viruses is lower than the concentrations required to precipitate symmetric viruses. We reasoned that the same was likely the case for precipitation of DNA and, indeed, we found that DNA can be size-fractionated by PEG precipitation (12). We applied this to our digested heteroduplex DNA samples and then completed the purification by gel electrophoresis. Our initial plasmid clones of *ara* DNA were made from DNA fragments purified from the double-bubble heteroduplexes.

High-resolution electron microscopy. With the DNA components of the *ara* regulation system in hand, one of the first questions to answer when we also had some of the protein components available was where the proteins bound in the *pBAD* regulatory region. Nowadays, we have a number of methods for determination of a protein's binding site on DNA, but at the time we were ready to proceed, there were none. One way to find the proteins' binding sites was merely to bind the proteins, look, and see. This meant performing high-resolution electron microscopy of protein-DNA complexes. A few papers on microscopy of proteins bound to DNA had been published, so it seemed feasible, but the techniques were hardly well worked out and seemed difficult even for expert microscopists. On the other hand, since not a lot of work had been done, we would not be at a great handicap if we plunged in and tried to use high-resolution microscopy to answer our questions. One of my students, Jay Hirsh, bravely undertook that difficult project that combined a requirement for high mechanical aptitude, great patience, and thoughtful scientific experiments. Although there was a good electron microscope available at Brandeis, it was not being used for high-resolution work. In fact, we were unaware of anyone in the Boston area who was doing high-resolution EM work of a nature that would be helpful in reaching our goals. Eventually, our efforts at developing techniques for the observation of regulatory proteins specifically bound to DNA succeeded. Our progress was aided in part by the mysterious and anonymous appearance of a 35 mm camera for the Philips EM 300 electron microscope.

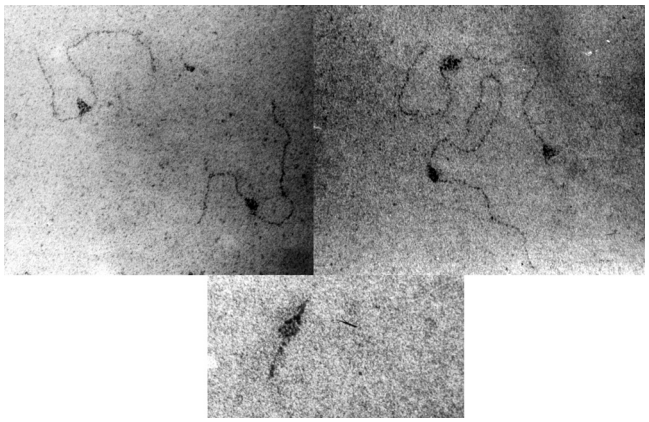

FIG 6 High resolution electron micrographs of AraC, CRP, and RNA polymerase incubated in the presence of D-fucose (an anti-inducer of the *ara* genes) bound to an 1,120-bp DNA fragment, the same proteins incubated in the presence of L-arabinose, and, bottom, Lac repressor bound to a 203-bp DNA fragment.

This allowed us to shoot 50 micrographs in a sitting, reduced costs, and greatly speeded our progress. We were able to visualize AraC, Lac repressor, lambda phage repressor, RNA polymerase, CRP-RNA, and AraC-CRP-RNA polymerase complexes, on fragments of DNA about 1,000 bp from *ara*, *lac*, or lambda phage (13–15) (Fig. 6).

The locations of the various protein binding sites observed in the EM were reproducible and consistent with the genetic and biochemical data. For example, the AraC-CRP-RNA-DNA complexes disappeared upon addition of ribonucleoside triphosphates and allowing transcription to proceed. Our measurements of the lengths of a 200-bp piece of DNA with and without bound *lac* repressor disproved a hypothesis put forward by Francis Crick on the disposition of the DNA under the repressor. We wrote to Crick and described our results and exchanged several letters. In the final letter, Crick very graciously admitted that he was wrong, greatly impressing me, a young faculty member, and Hirsh, my student.

Identification, mapping, and study of additional genes regulated by AraC protein. Study of additional promoters regulated by AraC seemed likely to identify and simplify unraveling the key elements in AraC's regulatory mechanisms. We reasoned that since the *araCBAD* gene cluster does not contain the genes for arabinose transport, they must lie elsewhere on the chromosome and should provide additional examples of promoters that are

regulated by AraC. Multiple genetic searches for arabinose negative mutants, however, had repeatedly failed to identify genes whose products transported arabinose into cells. This means that the cells contain at least two arabinose transport systems with overlapping functions and that it would require multiple simultaneous mutations to inactivate arabinose transport and render the cells unable to grow on arabinose.

We used two approaches (and two graduate students) to identify and begin study of additional AraC-responsive genes. In the first approach, we used the then-nascent cloning technology to seek DNA fragments that conferred arabinose dependence on expression of a promoterless tetracycline resistance gene on a plasmid (16). In the second approach, we used a variant Mu phage that had recently been constructed (17). As mentioned earlier, when phage Mu lysogenizes or even lytically infects cells, it inserts itself into random locations in the chromosome, inactivating any essential genetic element into which it has inserted. A special Mu-*lac* phage had been isolated which carries a promoterless *lacZ* gene. If the phage inserts in the correct orientation in an actively transcribed region, its *lacZ* gene is expressed, and colonies from such an insertion can readily be identified. With a simple colony screen, we identified and studied multiple Mu-*lac* insertions whose β-galactosidase expression was arabinose dependent (18). Once one transport mutant was identified, it was straightforward to select for Mu-*lac* insertions in the second transport system. We therefore had generated physical markers in both the low-affinity *araE* and the high-affinity *araFGH* gene systems. These markers greatly facilitated the identification of clones carrying the promoter regions from each locus. The studies also revealed that the promoter that had been isolated with the tetracycline-resistance approach controlled expression of a different gene, *araJ* (19, 20), whose cellular function is still uncertain. Other genes that are strongly induced by the presence of arabinose were not detected.

AraC$^+$ is dominant to AraC(con) at the *pE* and *pFG* promoters, as it is at *pBAD*, suggesting that both operons are subject to active repression as at *pBAD* (17). Both promoters are subject to catabolite repression and therefore the cyclic AMP receptor protein (CRP) is involved in their regulation in addition to its involvement in regulating the *pC* and *pBAD* (21–23). The promoter regions of all four of the *araC*-inducible promoters have been sequenced, the transcription start sites determined by S1 mapping, and the binding sites for AraC and CRP protein determined by DNase I footprinting (19–21, 24, 25) (Fig. 2). Thus far, no upstream binding sites for DNA looping have been biochemically identified at *pE* or *pFGH*. Sequence analysis of regions still further upstream, however, reveals potential binding sites that could be involved in DNA looping and repression of *pE* and *pFGH*. The locations of the CRP, AraC, and RNA polymerase binding sites at *pE* and *pJ* are like those at *pBAD*, but at *pFGH* two full AraC-binding sites are found that are located upstream from the CRP binding site, which lies adjacent to the RNA polymerase binding site. Little work has been done to determine the mechanism of repression (or induction) at these other promoters.

Homologs of the dimerization and of the DNA binding domains. Many years after the initial genetics studies mentioned at the beginning of this review, at a time when DNA sequencing was well developed and quite a few genes from *E. coli* and a few other bacteria had been determined, it seemed that analysis of homologs of *araC* might assist study of the structure, function, and mechanism of the protein. It turned out that not much could be gleaned from the few sequences available at this point because the homologs were not closely related to AraC. However, the rhamnose operon in *E. coli* appeared to be a source of a close AraC homolog, as rhamnose is a carbohydrate like arabinose and the rhamnose genes are under positive control like the arabinose operon (26). Therefore, we cloned and studied its positive regulator RhaR (27). This study and a subsequent study revealed a second regulatory gene in the system, RhaS (28). The sequences of both were similar to AraC, and the arrangement of the CRP, RhaS, and RNA polymerase binding sites in front of the *rhaBAD* genes was the same as that in front of *araBAD*. Further studies of the *rha* system have been greatly hindered by the extreme insolubility of the RhaR and RhaS proteins (29).

Now, after many thousands of bacterial genomes have been sequenced, it is tempting again to see what can be learned from homologs of AraC. First, it turns out that the DNA binding domain of AraC with its double helix-turn-helix DNA binding motif is very widely used for DNA binding proteins. The great majority of proteins with these sequences do not regulate arabinose catabolizing genes. However, the automated annotation

programs label such genes as homologs of AraC, so that searching for "araC" or for sequences homologous to the full-length AraC gene from *E. coli* yields tens of thousands of genes called AraC or labeled AraC homolog, when, in fact, they should not be. Searching with just the sequence of the dimerization domain of AraC yields many fewer supposed homologs, but closer examination shows that most of these also do not regulate arabinose catabolizing genes. Homologs coding for proteins that do regulate arabinose utilization genes can be identified by examining in sequenced genomes the genes immediately adjacent to an *araC* candidate. If the adjacent genes are homologs of the arabinose catabolizing genes from *E. coli*, then the gene in question can be concluded to be a regulator of an arabinose operon (30). As described later, authentic homologs provided the information necessary to identify an important contact made between the AraC DNA binding domain and dimerization domain when the protein is in its inducing state.

ARAC ACTION AND PHYSIOLOGY OF THE ARABINOSE SYSTEM

Once the possibility of positive regulation of *ara* by AraC protein had been raised, interest arose as to its mechanism. As explained in the previous section, I began study of the *ara* system partly for this reason. Initially, without the purified components of the regulatory system, the only experiments possible were those with growing cells. Most often these types of experiments suggest several possible mechanisms or models and usually exclude several possible mechanisms. Only rarely can such physiological experiments lead to definitive conclusions.

By 1973, study of a few genetic systems had not yet definitively demonstrated any mechanism for regulation of gene activity. On one hand, study of the *lac* operon suggested to some that a repressor controlled expression of the *lacZYA* genes by blocking access of RNA polymerase to the *lac* promoter. On the other hand, since the *lac* operator is positioned immediately adjacent to, but on the downstream side of the promoter, others postulated that *lacI* repressor blocked the further progress of an RNA polymerase that had completed at least some of the steps of initiation. In experiments described in the next section, we addressed the question of what step the *lac* repressor blocked. At around the same time, studies of the *trp* genes in *E. coli* indicated that, in addition to a repressor-operator, an additional level of regulation occurred downstream from the transcription initiation point. At the downstream site, depending on the availability of charged $tRNA^{trp}$, transcription was either terminated or allowed to continue (31). Similar antitermination systems have been found in a number of other bacterial genes, as well as in phage lambda (32). Thus, it appeared that Nature utilized a variety of regulation mechanisms in contrast to the initial expectations there would be just one universally utilized mechanism.

Our initial experiments at Brandeis were genetic, as described in the first section, and physiological, as described below. We also developed methods to purify AraC protein and *ara* regulatory region DNA, as described in third section. This was to enable the biochemical, biophysical, and mechanistic experiments described in final two sections.

Induction and basal levels, early time induction kinetics of *pBAD*, no stalled transcription complexes.

In light of the variety of potential gene regulation systems, the proposal of a transcription blockage mechanism in *lac* and the antitermination mechanisms in lambda and *trp*, it seemed possible that AraC acted in a positive manner to turn on *ara* gene expression by acting as an antiterminator. To look for these possibilities, we very carefully measured the time between the step of transcription initiation that is sensitive to rifampin and the time at which AraC acts to promote transcription (33). To make these measurements, we developed generally useful techniques to measure the time to the second that is required for inducers or rifampin that have been added to a culture to reach effective intracellular concentrations. The experiments showed that less than 10% of the *ara* operon copies in a cell possess an RNA polymerase molecule that has passed the rifampin-sensitive step and which can be triggered by arabinose to transcribe *araBAD*. Similarly, less than 5% of uninduced *lac* operon copies possess a rifampin-resistant RNA polymerase that can be released for transcription by the addition of a *lac* inducer. An offshoot of these experiments was the determination that, on average, transcription is initiated on the *ara* genes every 30 s, and on the *lac* genes every 50 s.

These were fun experiments to do. Because it was impossible to watch a timer and take the sample points at the very close time intervals that were required, we

used the audible beats from a metronome set to one per second for timing. In the resulting paper, we described the pace as *largo non rubato*. It was a minor disappointment that no one ever commented on this musical approach to increased precision in molecular biology. It was also disappointing that the work seems largely to have passed unnoticed. Perhaps this is a consequence of the fact that the work developed ideas and reasoning from enzymology to address a physiological problem, and few scientists are interested in both fields.

Catabolite repression and the roles of the cyclic AMP receptor protein CRP. The *in vitro* transcription-translation system that we used to detect AraC and definitively show that it was a positive regulator are discussed in the next section of this review (34). Those experiments, which actually preceded some of the experiments described in this section (no perfect order of topic presentation exists for this review), showed that expression of *pBAD* was cyclic AMP dependent and responsive to a phenomenon known as catabolite repression. Here, I describe some of our physiological experiments that were aimed at learning what we could about regulation of *ara* by CRP before we had learned how to purify components of the *ara* system for use in mechanistic experiments *in vitro*.

Generally, cells growing on glucose are poorly inducible for genes of other catabolic operons. We found this to be the case for genes of the *ara* system, and that the catabolic effects could be partially overcome by the addition of cAMP to the medium (35). At the time, it was surprising that the strengths of the catabolite repression effects were different between the *lac* and *ara* operons. This difference was yet another manifestation of the fact, now well known and accepted, that Nature utilizes multiple mechanisms of gene regulation and that a single regulatory protein may itself utilize multiple mechanisms in its regulation of genes. This is a far cry from the phenomena and objects studied in physics, where Newton's laws of motion apply everywhere in our common experience and all electrons have the same electric charge.

Several years after our early experiments on catabolite repression, in our determination of the locations of protein binding sites in the *ara* regulatory regions of *pBAD*, and then, a few years later, in our work on *pE*, *pFGH*, and *pJ*, we mapped the AraC and CRP binding sites as described in the previous section. Finally, we were able to study *in vitro ara* systems that mimicked their *in vivo* behavior, and also to investigate the effect of DNA looping on *pBAD* expression. We found with both *in vitro* and *in vivo* experiments that CRP both assists opening the DNA loop to relieve repression and also directly stimulates initiation of transcription, for a total of about a 30-fold effect (36–38).

Identification of specific residue-base interactions in the AraC binding site for induction. Our high-resolution electron microscopy and then the existing techniques for the study of protein-DNA interactions had allowed us to learn approximately where in the *araBAD* regulatory region AraC, CRP, and RNA bound (Fig. 1). Paradoxes and questions remained, however, and higher resolution methods were required to proceed. Protection experiments had indicated that, inexplicably, AraC bound at *O2*, *I* (*I1-I2*), and *O1* by contacting two, three, and four major groove regions, respectively. Later experiments showed that each subunit of AraC contacts two major groove regions, but at *I*, the fourth major groove was weakly contacted; only one subunit binds at *O2* whereas at *O1* both subunits bind to DNA. At that time, only indirect evidence indicated that the carboxyl third of AraC was a DNA-binding domain. This region of the protein contained two stretches of amino acids with significant similarity to the helix-turn-helix (H-T-H) motif that had been recently identified as present in only one copy in a small number of DNA binding domains (39). Thus, the questions were whether either or both regions were, in fact, H-T-H structures that directly contacted DNA. Finally, the binding regions located by DNA footprinting contained elements of inverted repeat sequence but also direct repeat sequence. Up to this point, only inverted repeat sequences had been observed in the DNA binding sites of regulatory proteins. One way to address all these questions on the binding by AraC was to identify specific residue-base interactions between residues in the two H-T-H regions in AraC and its DNA binding sites.

At the time, there existed one definitive, but very laborious method for determining specific residue-base interactions. In this, one needed to guess or infer that a particular residue was involved in direct contacts with

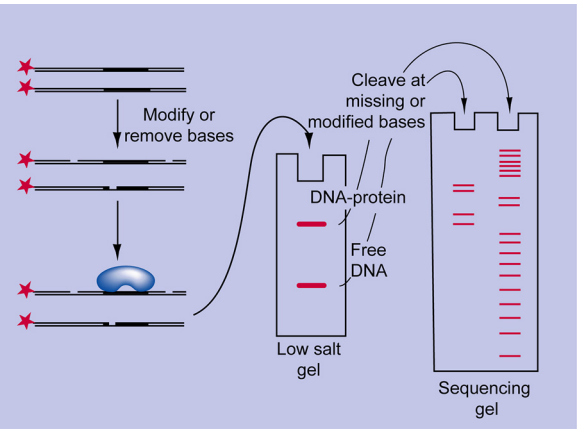

FIG 7 Missing contact method for identifying specific residue-base interactions. In this case, the protein makes contact with the DNA in two regions. DNA molecules with bases missing or modified in these two regions do not bind the protein well and run on the gel as free DNA. Subsequently, when these molecules are cleaved at the missing or modified bases, they give rise to specific bands on the sequencing gel. DNA molecules with bases missing or modified in regions not contacted by the protein give rise to bands on the sequencing gel in complementary positions.

a particular base of the DNA. Then one mutated or changed this residue to one with a smaller side chain, typically glycine or alanine. After that one then mutated the potential target base to show that while the binding affinity of wild-type protein was sensitive to the identity of the base in question, that of the mutant protein was not. Even if one can make a good guess as to a residue that contacts DNA, the difficulty with this method is correctly guessing which base is being contacted. The construction and confirmation process is quite lengthy and making multiple guesses would be painfully slow.

We thus developed the missing contact method to streamline the step of identifying the specific base that is contacted by a specific residue of an alanine-substituted DNA binding protein (40) (Fig. 7). First, a sample of a DNA fragment containing the binding site is lightly (approximately one per DNA duplex) depurinated or depyrimidated. Next, those DNA molecules that still bind the wild-type and alanine-substituted proteins with their normal affinities are separated from those which do not. Those that bind normally have not lost a base that is contacted by the protein, whereas those that bind less tightly have lost a contact. The wild-type protein will be sensitive to the loss of all the bases in its contact region, whereas the alanine-substituted protein will not be sensitive to the loss of the base that is normally contacted by the residue that was changed to alanine.

The next section of the review contains a description of the development of the gel electrophoretic binding and separation assay that was used here. After the separation, the DNA molecules are cleaved at the positions of the lost bases and run on sequencing gels for identification of the precise base at which the cleavage occurred. The DNA population that bound weakly because of a base missing from the contacted region would preferentially be cleaved at a contacted base, whereas the DNA population missing a base from a noncontacted region would preferentially be cleaved at positions not contacted (Fig. 7). Thus, the contacted bases could be identified.

With the missing contact method, we identified an interaction between a residue in the first H-T-H motif in AraC and two bases separated by 21 bases in the DNA of *araI*. This indicates that two subunits of AraC contact the DNA and that AraC most likely is a dimer, at least while bound to *araI*. The behavior of mutants with alterations in the second potential H-T-H region was consistent with it too contacting DNA, but these experiments were not conclusive. Later experiments resolved the question of the different apparent sizes of the binding sites. An intriguing outcome was that the binding sites for AraC contained a good direct repeat sequence component in addition to an inverted repeat component. The next section describes how we addressed the question of whether the AraC protein binding site at *araI* is a direct or inverted repeat.

AraC binding site is a direct repeat. The 21-base separation (two helical turns of DNA) between the contacts revealed in the missing contact experiments suggested that AraC recognized the *araI* site as a direct repeat, in contrast to the inverted repeat structure of the *lac* operator, but this was not a definitive proof (40). The AraC binding region can (as can any sequence) be broken down into direct and inverted repeat components, and it was not possible by this means to conclusively identify the AraC binding site through inspection of the sequences. By this time, we had identified the two AraC binding half-sites of *araI*, *I1*, *I2*, the left and right half-sites of the *O1* operator for repression of *pC*, the *O2* half-site, two half-sites ahead of *araE*, and four half-sites ahead of *araFG*, for a total of 11 half-sites. (The reason for the confusing nomenclature involving half-sites and full sites is that originally the *ara* DNA region utilized by AraC to induce *pBAD* was named *araI*. Since the two DNA-binding domains of the dimeric AraC bind to the full *I* site, the binding site for each monomer needs to be called a half-

site.) The *araI1* half-site was the closest to a consensus from all the half-sites. To test whether the full binding site was a direct or inverted repeat, we constructed direct and inverted repeats of the *I1* half-site sequence (41). The direct site bound AraC significantly more tightly than the inverted repeat. Together with the missing contact data, it was clear that AraC binds to a direct repeat of its half-site. This raised a new problem, however: what was the symmetry of AraC itself?

AraC protein appears not to possess a direct repeat structure. The finding that AraC binds to a direct repeat DNA sequence raised a problem. If the protein is a dimer with a direct repeat structure, there would be no reason for it to be limited to a dimer. Presumably, trimers and higher-order linear oligomers, particularly at high protein concentrations, should form. While such oligomers were not seen in solution, we also looked for them while bound to DNA containing three and four direct repeats of the *araI1* half site. No evidence was seen for binding by a trimeric or tetrameric form of AraC. This then suggested that AraC dimerizes in a closed form; that is, it possesses a 2-fold symmetry so that it cannot oligomerize beyond a dimer. In order that the DNA-binding domains bind to direct repeat sites, they cannot maintain the same 2-fold symmetry, and therefore have to be structurally separate from the part of the protein that specifies dimerization.

AraC stimulates both RNA polymerase binding and the transition to an open complex. One of the reasons for beginning a detailed study of the *ara* operon and AraC protein was to determine whether positive regulation of gene activity existed and, if it did, to determine the details of how it worked. We veered away from an intensive study of this question because many others in other systems were addressing the same type of question. Instead, we focused more and more on the question of what was happening within AraC itself. We did, however develop an *in vitro* assay for the initiation of transcription by RNA polymerase. The assay is useful in the study of systems like *ara*, in which proteins in addition to RNA polymerase are required to initiate transcription (42). We found that AraC played two almost equal roles in initiating transcription, the first in assisting the binding of RNA polymerase to the *pBAD* promoter and the second in stimulating the conversion of the closed RNA polymerase-DNA complex to an open complex.

Perhaps it is useful here to mention that the term promoter has drifted slightly over the years. Initially it was defined as the binding site for RNA polymerase. Then, unfortunately, a particular mutation ahead of the *lacZ* gene was considered the exemplar of a promoter mutation. It, however, turned out to lie in the CRP-binding site, not in the sequence initially bound by RNA polymerase. More recently, by promoter, some people mean just the sequence bound by RNA polymerase, but more commonly, both when referring to prokaryotic promoters and eukaryotic promoters, the term "promoter" now means not only those sequences but also sequences immediately adjacent that are required for full transcriptional activity.

The use of the *ara* system for controlled gene expression. Foreign gene expression on plasmid expression vectors utilizing the *lac* promoter is often found to be moderately high even without explicit induction. This occurs when the number of Lac operator copies in a cell exceeds the number of Lac repressor molecules available to repress the promoters. The resulting elevated expression, even in the absence of inducers, prevents cloning genes whose products are toxic. The *ara pBAD* promoter has sometimes been used instead of the *lac* promoter because it should not induce as a result of a shortage of AraC. The *ara* system has also been used when variable levels of a protein are sought in physiological experiments. It is often thought that inducing with intermediate levels of arabinose will give intermediate levels of expression. This is risky, since the arabinose system can show a maintenance phenomenon where active transport of even low extracellular levels of arabinose can give sufficient intracellular levels to fully induce some cells. These cells can then maintain their full *ara* induction. In this situation, a population may consist of two types of cells, fully induced and not induced at all. A safer approach for obtaining controlled amounts of a product whose synthesis is under *ara* control is merely to fully induce a culture, remove inducer, and allow dilution by cell growth to reduce the product's level to that desired.

DETECTION, HYPERPRODUCTION, AND PURIFICATION OF ARAC

While it is nice to collect random facts and correlate observations on a topic, the objective in our work has been to understand the system sufficiently well that we

could design and build other systems that utilize the same principles. This criterion requires more than describing what happens, it requires understanding the biochemistry and biophysics underlying the biological behavior. Such a requirement necessitates careful biochemical and biophysical study of the system's components and their interactions. Therefore, from the beginnings of our study of *ara*, one of the most important goals was first, the detection, and then the purification of sizeable quantities of the central player in the system, the AraC protein. Achieving this goal took, perhaps, 3 decades. This part of the review describes several of the major steps along the way.

The periplasmic binding protein detour. Since the economical utilization of resources is the main reason for regulating gene expression, it is logical that the intracellular levels of gene regulators be very low. This expectation was confirmed when, after much effort, Gilbert detected and roughly characterized the regulator of the *lac* operon, LacI protein repressor, in a biochemical assay (5, 43). Its detection relied upon its high affinity for a gratuitous inducer, IPTG (isopropyl-β-D-thiogalactopyranoside), which was made still higher by the isolation of a special high-affinity mutant. An equilibrium dialysis assay of concentrated cell lysate detected a slight excess of the concentration of radioactive IPTG in a dialysis sack compared to its concentration in the buffer surrounding the sack. Further study was aided by increasing the gene dosage of repressor by seeking repressor in cells that had been infected with a ϕ80-λ–*lac* phage that increased the number of *lacI* gene copies to about 50 (44).

Following the lead for the detection of LacI, in Gilbert's lab I tried to detect AraC by equilibrium dialysis with radioactive arabinose (45). I quickly discovered a very weak signal. Immediately, we checked that the binding activity was not an enzyme or transport protein which might also bind arabinose. At first it appeared that the activity was altering the arabinose, but soon we discovered that two thirds of our so-called ^{14}C-arabinose was not arabinose at all. Upon resolving that, the problem was to demonstrate that the gene responsible for the activity was AraC. Constitutive mutations, C(con), retain AraC protein activity but make that activity resistant to the inhibitory effects of the L-arabinose analog D-fucose (five-methy L-arabinose). Instead of altering the binding of arabinose or fucose to the activity being detected, the AraC constitutive mutations turned out to dramatically increase the levels of the binding activity. Further experiments then showed that the binding activity was not at all from AraC, but instead was from a periplasmic binding protein, one of a number of such proteins with rather high K_d $\sim 10^{-6}$ M for their various substrates. These are parts of high-affinity active transport systems.

Assays of AraC, DNA and arabinose binding. Measurement of induction of the arabinose operon as a function of the intracellular arabinose concentration could allow a reasonable estimation of the affinity of AraC for arabinose, and we tried to do this. Such measurements must be done with transport-negative mutants. Our supposed transport-negative mutant turned out to still possess some active transport activity, so our measurement of the likely K_d for arabinose binding was incorrect (46), but others quickly corrected this mistake (47). Both experiments showed, however, that the binding affinity of AraC for arabinose was much too weak for the equilibrium dialysis assay to work.

Simple DNA binding assays also failed to detect AraC, as did gel electrophoresis assays looking for an AraC-specific band on SDS gel electrophoresis. The one assay that had to work was assaying for the same *in vitro* activity that the protein has *in vivo*. That is, use an *in vitro* coupled transcription-translation system in which the synthesis of one of the *ara* enzymes depends on a DNA template containing that *ara* gene and also depends on the addition of AraC protein. One such system for the synthesis of β-galactosidase using added λ*dlacZ* DNA as a template had just been developed and described by Geoffrey Zubay (48). At this time, Beckwith's lab, which had developed difficult *in vivo* techniques for fusing genes to other gene's promoters, succeeded in placing the genes coding for AraCBAD on a defective ϕ80 phage (49).

In our successful assay of AraC, we constructed a strain deleted of *lacZ* and of *araCBAD* to use in making the coupled transcription-translation system. This allowed testing its activity by the synthesis of β-galactosidase and then using this delicate and cumbersome system, once it was working, to look for the synthesis of the product of the *araB* gene, ribulose kinase (34). As a source of AraC, we used partially purified lysates from

araBAD-deleted cells infected with a λ-ϕ80*ara* phage carrying a good *araC* gene and defective *araB gene* that we constructed from the original ϕ80d*ara* phage. To assay the products of the transcription-translation reactions, we then used a sensitive radiometric assay for the presence of AraB product, ribulose phosphate. To perform the assay, we added ^{14}C-arabinose and the product of the *araB* gene, arabinose isomerase, to convert some of the arabinose to ribulose, the substrate of the AraB enzyme, and ATP. After an incubation of up to 24 h, we determined the amount of ^{14}C-ribulose phosphate that had been synthesized. This required the specific precipitation of sugar phosphates away from the arabinose and ribulose by barium. Finally, the amount of radioactivity that was precipitated gave an indication of the amount of AraC that had been present in the initial transcription-translation reaction. Later, the barium precipitation step was replaced with separation by paper chromatography. This approach finally provided the definitive proof that AraC protein is a positive regulator of the *araBAD* genes. Nice as it was to have an assay, this was an extraordinarily cumbersome assay! Nonetheless, my student, David Steffen, was able to purify a tiny amount of the protein to a level where its band on an SDS gel could just be detected, giving its molecular weight as about 30,000. He also ran it through a Sephadex sizing column to show that the protein is a dimer (50).

At this time, only a few companies sold biochemicals and none sold molecular biology reagents like restriction enzymes, which were just coming into use. Soon after Steffen's success, I received a one page Xerox copy of a typewritten page from a new company offering to sell the restriction enzyme EcoRI and also DNA oligonucleotides containing the EcoRI restriction site. It was easy to construct a fake advertisement from the company offering, in addition, milligram quantities of purified AraC protein. I left this on Steffen's desk and waited... Six months later, Steffen, with the help of my other graduate students and several faculty members, perpetrated a far more devastating practical joke on me, in which they cleverly documented a huge (fake) advance in a competing lab that would have made further work by us on the *ara* system pointless. Over the decades, we have been the source of a number of notable practical jokes, one (harmless) which brought the wrath of the Hopkins art curator down on me for seemingly having replaced an ancient (and particularly ugly) oil painting of a turn-of the-century Biology Department professor with an inkjet print of me in exactly the same pose.

It has been pretty standard in the acknowledgments sections of molecular biology papers to thank associates for their generous provision of some of the biological materials used in the work. Being tired of this old song, in another paper we thanked Steffen, one of the authors on the initial purification paper for "grudgingly providing ..." An editor did catch this and so we wrote to the editor and said that, "Yes, he grudgingly provided...," and that it was fine with Steffen that we said so. Ultimately, however, "grudgingly" didn't make it into the literature. In another paper we acknowledged my dog Sigma, who came with me to the lab every day. A few years later, as genetic engineering was becoming more sophisticated, we constructed a lambda phage that overproduced AraC still more greatly (51). This facilitated purification, but the purification was greatly hampered by the protein's insolubility, instability, and sensitivity to precipitation by ammonium sulfate, a convenient and commonly used purification and concentration step.

A few years later, Crothers at Yale and Revzin in Minnesota found that under very nonphysiological conditions, *lac* repressor could bind to DNA and be retained on the DNA during gel electrophoresis, during which it retarded the migration of the DNA (52, 53). We tried it with AraC and found that this "gel retardation" assay worked. We spent a lot of effort showing that the assay could monitor AraC binding to its specific binding sequences in physiological buffers and that important biophysical quantities, like binding rates and dissociation rates, could be measured and yielded sensible values (54). This convenient assay of AraC activity greatly reduced the labor of assaying the protein.

After another few years, the very convenient and effective T7 pET vector hyperexpression system developed by Studier became available (55). With this system, AraC could be overproduced to such high levels that even initial steps in its purification could be assayed by SDS gel electrophoresis and protein staining, and with the wonderful Pharmacia FPLC system, purification steps became much simpler and no longer required fiddling with chromatography columns in the cold room. Somehow, the solubility and stability of AraC then improved to the point that a good many biochemical and biophysical studies became possible. Much of the

improved stability resulted from the effective removal of cellular proteases that, in the past, had digested the DNA-binding domain of AraC, which is very sensitive to proteases.

One amazing peculiarity that we later discovered was that AraC is specifically precipitated by IPTG! Why this happens is a mystery, although it should be noted that the D-galactose moiety of IPTG is a structural analog of L-arabinose. IPTG precipitation and resuspension will take a prep of 90% purity to greater than 99% purity.

Some of the experiments that then became possible required measuring the affinity of arabinose binding. We found that the binding of arabinose shifted the tryptophan fluorescence spectrum of AraC by about 2 nm. Surprisingly, this very small shift can be accurately measured by determining the average emission wavelength of the entire emission spectrum (56). The accuracy comes from the fact that the average depends on multiple emission-strength measurements.

DNA LOOPING: ITS DISCOVERY, DEMONSTRATION, AND HOW IT AIDS ARA REGULATION

On one hand, our discovery of the phenomenon of DNA looping as a feature in gene regulation from bacteria to eukaryotes was an unexpected and fortuitous finding. On the other hand, as described below, it was the culmination of more than a decade of very hard work directed at understanding an unexpected observation made by another scientist.

The Englesberg deletion Δ719 left *pBAD* fully inducible, and with a normal basal level in the absence of functional AraC (Fig. 3). In the presence of AraC, the basal level was increased 30-fold and, as explained earlier, this result suggests that in the cells carrying the deletion, the molecules of AraC that are in the repressing state can no longer block the inducing fraction from acting because the deletion removed the site required for repression. However, all of the sites required for the normal 200- to 500-fold induction likely remained intact.

AraC binds to the upstream repression site, and the helical twist experiment. The first section of this review describes our extensive genetic studies that verified that deletions in the *ara* regulatory region could damage an apparent repression activity without interfering with the ability to induce. These studies also developed the technology for isolation of different regions of the DNA involved in induction and repression. This was accomplished at a time before DNA isolation using genetic engineering methods became possible. When genetic engineering methods had developed to the point of being able to map protein binding sites on DNA, we finally used DNase protection experiments and controlled exonuclease digestion experiments to show that AraC could bind to DNA at a site well upstream of the binding sites involved in induction and in about the location predicted by our earlier genetic studies (57). We also ruefully set aside the years of work that we had invested in developing the technology for the isolation of the *araCBAD* regulatory region and subportions of it. Using the newly developed tools of genetic engineering was a more efficient way to proceed.

In a Monday group meeting, we discussed the new data on the autoradiographic films that had been developed that morning. They clearly showed the presence of an AraC binding site several hundred base pairs upstream from *pBAD*. We discussed how repression could be generated from so far upstream and I proposed that it was DNA looping and proposed the helical twist experiment (Fig. 8). In this experiment, at a nonessential region between *araO2* and *araI*, a small number of base pairs are inserted or deleted. An insertion or deletion of five base pairs rotates *araO2* half a turn with respect to *araI* and, if the binding surface of *araO2* previously faced *araI*, after the insertion or deletion it would face away. Looping such

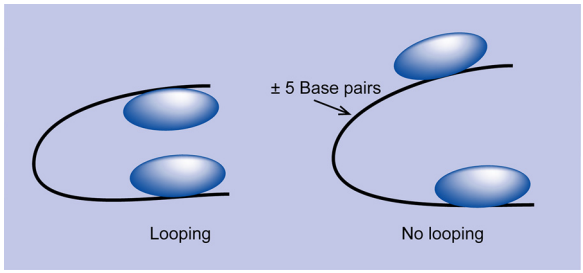

FIG 8 The helical twist experiment with DNA. Insertion or deletion of five base pairs of DNA between the two binding sites for the looping protein twists one with respect to the other half a turn of the DNA. If the free energy that is obtained from looping is substantially less than the energetic cost of twisting the DNA included between the binding sites by half a turn, looping will then be much reduced.

DNA would then additionally require over or under twisting the DNA by half a turn. If the energy required for this twisting is comparable to the free energy available for looping, the looping will be inhibited. Analogously, insertions or deletions of 11, 22, and 33 bp should restore DNA looping. (Later experiments showed that the helical pitch of DNA *in vivo* is 11 bp per turn.) All four of my graduate students at the time declined to try the experiment, which was far from easy given the state of genetic engineering at that time. Therefore, my technician and I began the experiment, which took about 6 months to complete and which, fortunately, gave a clear positive result (57, 58) (Fig. 8).

At the time of our proposal of DNA looping on the basis of the genetics, footprinting, and helical twist experiment, there were good reasons both to doubt its possibility or to believe it possible. On the basis of the known flexibility of DNA (59), forming a DNA loop of 200 bp seemed energetically unfeasible. On the other hand, those objecting on the basis of energetic considerations probably failed to realize that the DNA merely had to bend into a rather open U turn, with much of the bending to complete the circle of DNA-protein being done by AraC. They also did not consider that DNA *in vivo* is supercoiled, and as a result possesses fairly tight coils that might facilitate DNA loop formation.

Additional experiments supporting DNA looping. The genetics, footprinting, and helical twist experiments identified the *araO2* site as required for repression. The site is some 200 bp upstream from the *ara* sites required for induction. Although the experiments identified the site, they did not definitively prove that the repressive effect involved DNA looping. Therefore, we undertook a number of experiments with the goal of proving the presence of DNA looping or, lacking that, of accumulating sufficient evidence consistent with looping from such a wide variety of experiments that there could be no reasonable doubt as to its existence.

One obvious requirement of the DNA looping proposal was that AraC should occupy the *araO2* site *in vivo* in the absence of arabinose, and in the presence of arabinose either reduce its occupancy of the site or not occupy it at all. As it also seemed highly likely that the looping was mediated by AraC bound both at *araO2* and at the induction site *araI*, one consequence of this prediction was that AraC would be found at *araI in vivo*, even in the absence of arabinose. The problem was how to demonstrate that a protein was bound to a specific site *in vivo*. At the time, there were no methods available to do this.

We had already used dimethyl sulfate (DMS) footprinting in our *in vitro* experiments, locating the AraC, CRP, and RNA polymerase binding sites in *pBAD*. For an *in vivo* demonstration of binding, we hoped that DMS might enter cells and methylate DNA before its presence substantially changed intracellular conditions and rendered the findings meaningless. The experiment would also require adequate purification of the *ara*-containing plasmid DNA from the vast excess of chromosomal DNA. We tried the *in vivo* footprinting, and it worked (58). At the *araI* site, the pattern of methylations *in vivo* in the presence of arabinose in *araC*$^+$ cells matched the methylation pattern seen *in vitro*. Crucially for the looping hypothesis, *in vivo* the *araO2* site was occupied by AraC in the absence of arabinose and substantially less so in the presence of arabinose. Also, the *araI* site was occupied by AraC in the absence of arabinose. Thus, the *in vivo* footprinting that we developed provided important data supporting DNA looping.

Another type of experimental test of DNA looping was genetic. If DNA looping and repression occurs as a result of AraC binding at *araO2* and *araI*, then it should be possible to isolate point mutations in *araO2* and possibly also in *araI* that have lost the ability to repress, but can still induce. This proved to be the case (58), and we found mutations in both sites that were simultaneously repression-negative and induction-positive.

***In vitro* looping with small supercoiled DNA.** *In vitro* experiments failed to display DNA looping until we tried using small supercoiled DNA molecules containing *araI* and *araO2* instead of linear DNA (60). In the gel retardation assay, the circles with nothing bound, with a dimer of AraC bound at *araI*, and with a dimer looping the DNA with one subunit contacting *araO2* and one subunit contacting the left half of *araI* could all be distinguished. With the assay using radiolabeled DNA and radiolabeled AraC, it was also possible to show that, indeed, it was a dimer that formed the DNA loop, with one subunit binding DNA at either end of the loop. The loop was present in the absence of arabinose, and upon the addition of arabinose, the loop opened and the bound AraC molecule shifted from looping to binding to *araI1-I2*.

Experiments with and without the *araI2* half-site showed that forming the loop involved AraC binding only to *araI1* and not to *araI2*. DNA methylation experiments confirmed this finding. Another experiment addressed the question of whether the DNA sequence differences among the three sites involved (*araO2*, *araI1*, and *araI2*) were important factors in determining whether AraC looped between *araO2* and *araI1* or bound *cis* to *araI1-araI2*. As we found that the sites could be interchanged or altered without affecting the looping-unlooping reaction, it was clear that a property of the protein under the control of arabinose is the primary factor determining looping.

Why loop? Why should DNA loop in *ara*? First of all, it needs to be remembered that by far the most straightforward way that a transcription factor's activity can be directed to a specific gene is for the transcription factor to bind to a specific DNA sequence located at the gene to be regulated. On the contrary, for example, a transcription factor that bound to only RNA polymerase might affect the activity of hundreds of genes.

A number of possible reasons for looping have been raised in the past. Most notably, looping permits multiple transcription factors to be involved in regulating transcription from a promoter. Without looping, two or three transcription factors are about the limit of the number of proteins that could bind to DNA near a promoter and affect the activity of RNA polymerase.

Now that a thermodynamic mechanism that is consistent with virtually all the existing data has been developed, a reason for DNA looping is more apparent. Transcription in *ara* is subject to two regulatory activities of AraC, i.e., repression, which is dominant, and induction. By virtue of DNA looping, a shift in the state of the AraC molecules in a cell from mostly being in the repressing state to mostly being in the inducing state both reduces repression and increases induction. Thus, the transcriptional response can be proportional to the square of the amount of the shift to the inducing state instead of merely being linearly proportional to the amount of such a shift.

BIOPHYSICAL AND BIOCHEMICAL STUDIES OF ARAC AND DNA

Despite the power of genetic and physiological studies, these approaches usually yield data that is consistent with multiple and, frequently, very different physical explanations. Normally, biochemical and biophysical experiments are necessary for the definitive demonstration of the validity of a theory. Sometimes, however, none of the general approaches (genetic, physiologic, biochemical, and biophysical) provide conclusive proof of a hypothesis. As mentioned earlier, however, sometimes when a multitude of different experimental results consistent or inconsistent with one hypothesis have been found, then, for all practical purposes, the hypothesis can be considered proven or disproven. This was the case for "proving" DNA looping and also for disproving one of the most appealing of the mechanisms for how the binding of arabinose shifted AraC from looping to binding *cis* to *araI1-I2*. This was called the light switch mechanism and is discussed more fully in the next section.

This part of the review describes multiple experiments through which we learned the biochemical and biophysical properties of the *ara* system, and the next section describes work directed toward determining the more intimate details of AraC protein's response to its binding of arabinose, that is, its internal mechanism.

DNA digestion methods to locate domain boundaries and functional DNA binding positions of AraC. Before we had sufficient quantities of pure AraC for extensive biochemical studies, we sought to determine the domain structure of AraC and then, if it consisted of relatively autonomous domains, how they were connected. This was done with manipulations of DNA coding for AraC and then determining the activities of the various protein products *in vivo*. For example, we constructed genes for chimeras of AraC's DNA-binding domain and the dimerization domain of the coiled coil protein eukaryotic C/EBP transcription factor (61) (Fig. 9). Similarly, we fused the AraC dimerization domain to the LexA DNA-binding domain. Both types of chimeras functioned *in vivo* to bind to their DNA binding protein's recognition sequence, thus demonstrating the autonomous nature of the DNA-binding domains and the dimerization domains of AraC.

While the chimeras and sequence comparisons identified the approximate extents of the dimerization and DNA-binding domains, we sought also to determine the minimal functional domains. To do this, we digested DNA from either end of the region coding for each domain,

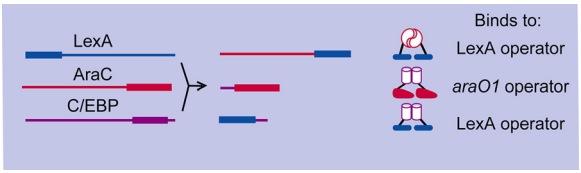

FIG 9 Chimeras used to demonstrate the autonomous nature of the dimerization and DNA-binding domains of AraC. The known autonomous dimerization domain from the protein C/EBP and the known autonomous DNA-binding domain from the LexA protein were used in the constructs. Binding was measured *in vivo* by the proteins' ability to repress the appropriate operators. On the left, the blue, red, and purple represent the LexA, AraC, and C/EBP proteins, respectively, with the thicker portions of the lines representing the proteins' DNA-binding domains. The middle portion represents the three chimeras used. On the right are representations of the three dimensional structures of the chimeras where the red yin-yang circles represent the dimerization domains of AraC and the purple cylinders represent the coiled-coil dimerization regions of C/EBP.

rejoined, and transformed the reconstructed DNA into cells whose growth or colony behavior depended on the domain's activity. To search explicitly for a linker region connecting the two domains of AraC, we modified the technique to generate random insertions throughout the gene (62). Presumably, a linker region would tolerate insertions, whereas insertions would not be tolerated within a normal globular structure. Interestingly, while the method did identify the AraC linker region, sometimes insertions elsewhere did not inactivate the protein. We also used these same techniques to generate a wide variety of positions of the *araI* site with respect to the RNA polymerase binding site in both the normal and inverted orientation. The only position and orientation with significant activity was the natural one (63). As AraC bound to DNA in this position could make direct contact with RNA polymerase, these results make it likely that it does so.

Measurement of the relative binding affinities of AraC to different DNA half-sites. As we devised various quantitative theories for the mechanism by which the binding of arabinose shifted AraC from DNA looping between *araO2* and ara*I1* to binding to *araI1-araI2*, it became necessary to know the relative binding affinities to these sites. Despite the convenience of the gel retardation assay, it is very difficult with this assay to reproduce direct measurements of the K_d of a DNA-binding protein to its target DNA sequence. In part this results from the low protein and DNA concentrations that must be used in such experiments, coupled with the sizeable changes in relative protein concentration that result from small losses of protein to glass or plastic surfaces. A workaround was to simultaneously measure the relative affinities between two DNA sites in the same reaction mix. In such a case, both DNA sites would be exposed to the same concentration of free AraC and, within the proper range of AraC protein, the relative AraC occupancy of the two DNA sites would be proportional to their relative affinities for AraC. By placing the two sites on DNA fragments of different sizes, the four resulting DNA species (two with AraC bound and two free of AraC) could be separated and the ratios of AraC-bound DNA to free DNA for each of the two DNAs could be measured. This provided the relative affinities for the sites (64).

Story of the crystallization of the dimerization domain. As the study of AraC progressed, it became clear that significant further understanding was dependent upon knowing the tertiary structure of the protein. At that time, X-ray crystallography was the only available technique for determining the structure of a protein. Therefore, my technician and I set up a production line where every 3 weeks we would grow 30 liters of cells and purify AraC. As this was before the advent of the pET expression vectors, our yields varied from 25 to 100 mg of ~95% pure AraC. We would concentrate the protein as much as possible, but usually taking significant losses, mainly through aggregation. Then we attempted crystallization using the hanging drop method. After several months of failures and hundreds of trays of hanging drops, we finally obtained some beautiful crystals. A quick test in an X-ray beam disappointingly showed that the crystals were unsuitable for structure determination. SDS gel electrophoresis of the protein from a crystal showed that the DNA-binding domain was absent, presumably having been digested away by contaminating proteases. We therefore turned to purifying AraC, intentionally digesting the DNA-binding domain away and attempting to crystallize the remaining dimerization domain in the presence of arabinose. Months of failures still followed. In desperation, in an examination of drops hung relatively early in the process, we found one crystal. Attempts at growing more crystals under the same conditions failed. Therefore, with considerable trepidation, we sacrificed

our one decent crystal by crushing it and used its fragments to seed new crystallization attempts. These were successful, but the folks who finally determined the structure for us complained that they were too small. We persevered and in a few months our collaborators, Soisson and Wolberger, were complaining that our crystals were too large. Eventually we also succeeded in growing crystals in the absence of arabinose and in the presence of D-fucose, an analog of L-arabinose that blocks arabinose binding and does not activate AraC. The structures of the three forms were all determined (65, 66). One of the most important findings from these structure studies was that the N-terminal 20 amino acids of the protein form an arm that folds over arabinose bound in a pocket in the rest of the dimerization domain. In the absence of arabinose, the arm adopts a different structure and position.

DNA binding properties of the DNA-binding domain. To advance our understanding of DNA binding by AraC, we engineered the hypersynthesis of its DNA-binding domain. Using SDS gel electrophoresis as an assay, it was relatively easy to purify the domain. Quite surprisingly, the DNA-binding domain was highly soluble (67). Since full-length AraC was not highly soluble, but its dimerization domain was soluble, it seemed that the solubility problems must lie in the DNA-binding domain itself. However, as each isolated domain turned out to be soluble, the explanation could have been that each was soluble under different conditions, and that there was no buffer condition at which both were very soluble. This turned out to be only part of the explanation, as later experiments showed that despite the structural autonomy of the domains, significant domain-domain interactions also occur.

Initially, it was surprising that the gel retardation assay indicated that the isolated DNA-binding domain of AraC bound to *araI1* rather tightly. Appropriate control experiments soon showed, however, that the DNA-binding domain was binding to DNA only after mixtures of DNA-binding domain and DNA were loaded on the gel and rapid diffusion of salt at this time lowered the salt concentration in the protein-DNA mixture to the point that the DNA-binding domain could then bind to the DNA and retard its migration through the gel. Perhaps it should be mentioned that DNA-binding proteins and cations compete with each other in "binding" to DNA and, therefore, at the very low salt concentrations in the gels used in the gel binding assay, the binding of DNA binding proteins is strongly enhanced (68).

Computational prediction of structure (Rosetta) and of function (CHARMM). Since proteins are just complicated polymeric organic molecules, it would seem that the structures of proteins could be predicted from their amino acid sequences. Similarly, since the motions of atoms in proteins and DNA can be well described by classical mechanics, it would also seem that applying Newton's equations of motion with the interatomic forces known from chemistry and physics should be sufficient to fully describe the dynamic behavior of any protein. The application of both approaches could thereby eliminate the need to do laboratory experiments. Structure prediction grounded in basic principles has gotten fairly good, particularly when augmented with information extracted from massive sequence databases (69). Even better, truly excellent structure prediction is just now becoming possible using artificial intelligence (70), but the burden of computation in dynamic simulations is much too great for this approach to fully explain the basis of a protein's activity now or at any foreseeable time in the future. Nonetheless, partly to learn how to use structure prediction and dynamics prediction, we applied both approaches to assist our understanding of AraC.

We used the protein structure prediction program Rosetta to address the question of whether the N-terminal arm is self-structured, or whether it is structured only when it interacts with other parts of AraC. These studies indicated the N-terminal arm likely can assume a quasi-stable structure without interactions with other parts of AraC. We also found that the N-terminal arms of a number of homologs of AraC also possessed quasi-stable structures. Random sequences or scrambled arm sequences did not possess such structures (71).

In wild-type AraC, the side chain of residue F15, which lies in the middle of the arm, makes direct contact with bound arabinose, suggesting that this interaction is critical in causing the repositioning of the N-terminal arm that follows the binding of arabinose and which leads to induction. With the molecular dynamics program CHARMM, we simulated the behavior of the wild-type arm and of the 19 single amino acid variants possible at position 15. The

results correctly predicted the laboratory behavior for all the variants and suggested that the formation of a hydrophobic cluster of residues in the arm is critical to correct regulation by AraC (72).

DNA-assisted measurement of weak domain-domain interactions. As described more fully in the next section, the properties of AraC in the absence and presence of arabinose predicted that the DNA-binding domain bound to the dimerization domain and/or the N-terminal arm in the absence of arabinose and that this interaction was weaker or absent in the presence of arabinose. Because the domains are held in close proximity to one another by the relatively short interdomain linker, the effective concentration of one domain in the presence of the other is on the order of millimolar and, hence, if binding versus no binding between the domains is to play a regulatory role, the dissociation constant for the binding has to be on the order of millimolar as well. This is a very weak binding affinity and normally its biochemical study would require excessively high concentrations of the two domains, also on the order of millimolar. To overcome this problem we developed the DNA binding assistance assay to measure very weak protein-protein interactions (73) (Fig. 10). In this assay, single-stranded DNA tails possessing a short region of complementarity are added to each of the proteins. Consequently, the binding energy between the DNA tails is added to the protein-protein binding energy (minus, of course, the entropic cost from the flexibility of the DNA tails). To simplify the assay of association between the dimerization-DNA and DNA-binding domain-DNA components, one of the DNAs was labeled with a fluorophore and the other with a fluorescence quencher. This allowed a simple and sensitive fluorescence assay to measure association of the two complexes. Using this method, we measured an interaction of about the expected strength between the dimerization and DNA-binding domains in the absence of arabinose. Its strength was reduced, but not eliminated, on arabinose addition. Reluctantly, however, we have come to believe that the results from these experiments are misleading, probably because the full surface area of each domain was available for interaction in this assay, but in the normal AraC protein the connection between domains by the linker greatly restricts the portions of the surface that are available for interaction between the domains.

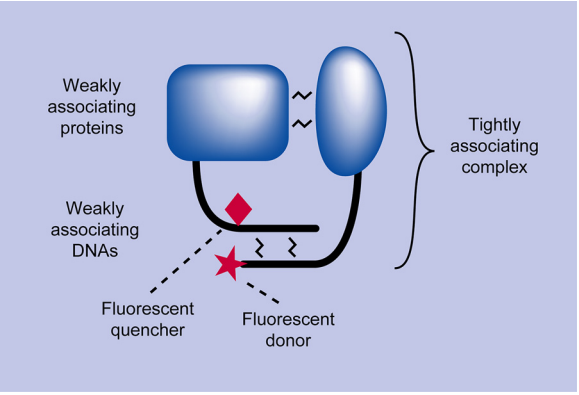

FIG 10 The DNA assistance method for measuring very weak protein-protein interactions.

Authentic domain-domain interactions in AraC. The previous section described detection of domain-domain interactions that we now think are not present in the wild-type AraC. Here is a more trustworthy demonstration of a functional interaction between the domains of AraC. We found that the dissociation rate of arabinose-bound full-length AraC from a single *I1* half-site, which involves only one DNA-binding domain, is 150 times faster than the dissociation rate of free DNA-binding domain from the same DNA (74). This result was supported by measurements of the arabinose binding affinity of full-length AraC with and without *araI1* bound, with chemical cross-linking experiments, and with NMR measurements.

The macromechanism of AraC. "The mechanism of AraC" can be viewed at three levels. At the macro level, the question is when or how does AraC action lead to transcription initiation at an *ara* promoter? At the next level, the question is how do interactions between AraC and RNA polymerase, and possibly also CRP, stimulate transcription? Finally, at the most atomistic, or micro, level, how does the binding of arabinose to AraC shift the protein from largely repressing transcription at *pBAD* to activating the transcription? Our laboratory has focused on the first and third aspects of the mechanism and has left transcription factor-RNA polymerase interactions to others. This section describes our understanding of the macro level question, and, in the last major section of the review, the micro level mechanism will be addressed.

Much evidence has already been described that leads to the following mechanism for regulating *pBAD*. AraC molecules in solution have two states that are in equilibrium with each other. In the absence of arabinose, their equilibrium favors

the repressing state where the DNA-binding domains are held in positions that favor DNA looping. When arabinose is bound, the DNA-binding domains are less tightly held and can now bind to the adjacent *araI1-I2* half-sites, where their binding stimulates the activity of *pBAD*. All that arabinose changes is the equilibrium between the repressing and inducing states. An AraC molecule in the repressing state can only bind to *araO2* and *araI1* and loop the DNA. In this looped DNA state, *pBAD* cannot be active both because *araI2* is not occupied by AraC and because looping interferes with the access of RNA polymerase to both *pBAD* and *pC* (23). An AraC molecule in the inducing state can only bind *to araI1-I2* and, when bound there, stimulates the *pBAD* promoter.

Quantitative estimations of the activity of *pBAD* can be made in terms of the equilibrium constants between the repressing and inducing forms of AraC in solution in the presence and absence of arabinose and the effective dissociation constants of the repressing form of AraC for looping DNA and the inducing form of AraC for open DNA. The situations for which the expression can be calculated in terms of the wild type basal level are: expression in the presence of arabinose, expression with *O2* absent both in the presence and the absence of arabinose, and expression in the presence of a large excess of AraC. When biochemically plausible values are chosen for the four free parameters, that is, the four dissociation constants, the predictions do remarkably well in predicting the experimentally observed data, typically within a factor of two. Perhaps the largest unknown in the calculations is whether or not AraC bound in the looped state can directly bind arabinose and transform to the inducing state without dissociating from the promoter region DNA.

MECHANISTIC STUDIES OF ARAC

The final section in this review describes a number of our attempts to learn the mechanism by which arabinose binding changes the tightness with which the DNA-binding domains of AraC are held in positions that favor DNA looping. For the most part, the experiments described in the previous part were designed to learn specific biochemical or biophysical properties of AraC, and some of the experimental methods and approaches described in this section were devised to test specific mechanistic hypotheses or to ask specific questions closely related to mechanism.

The first mechanistic problem: does the N-terminal arm interact with the DNA-binding domain? In the previous section, the agonizing attempts to determine the structure of AraC were described. The structures that were initially determined for the N-terminal arm plus dimerization domain in the presence and absence of arabinose initially suggested a mechanism (65). For that preliminary mechanism, the N-terminal arm, whose position is controlled by the binding of arabinose, determines which of two potential interfaces was used for dimerization and that this controlled whether the protein looped the DNA or not. Careful thinking and further experiments quickly discredited this model and the important question then became one of devising and testing a plausible model that was consistent with all the known experimental data. In the "corrected" mechanism (64, 75), known as the light switch mechanism, in the absence of arabinose, the two N-terminal arms from the two dimerization domains bind to the DNA-binding domains and help hold them in positions that favor DNA looping, and hence repression. Then, in the presence of arabinose, the arms bind over the arabinose, which binds in a pocket of the dimerization domain. This frees the DNA-binding domains to assume other positions and orientations and the lowest energy state then becomes one where the two DNA-binding domains of AraC bind in *cis* to the two adjacent half-sites at *araI* (*araI1* and *araI2*). Some of the earlier sections, and more of the sections below, describe the multiple experimental attempts to verify this attractive theory. None of which did. Finally, our attention turned to developing and testing other possible mechanisms.

AraC is rigid in the absence of arabinose and flexible in the presence of arabinose. In the investigation of whether AraC binds to a direct repeat or inverted repeat DNA sequence, we found that upon the addition of arabinose, the affinity of AraC for both relative orientations of the binding site increased (41). Either of two mechanisms can yield this behavior. The first is that, in the absence of arabinose, the DNA-binding domains are held in positions relative to each other such that only one or the other of the two DNA-binding domains can bind to one of the two adjacent half-sites. Presumably, the DNA-binding domains are held in positions suitable for DNA looping instead. Then, upon the addition of arabinose, the binding domains are less tightly held in these positions and it is significantly less energetically expensive than before for them both to

bind to the two binding sites, whether they are in direct or inverted repeat orientation. The second mechanism is that the binding of arabinose to AraC increases the intrinsic DNA binding affinity of each DNA-binding domain, and the sequence specificities are such that AraC then binds to *araI1-I2*.

Evidence consistent with the hypothesis that the DNA-binding domains are held in positions suitable for DNA looping also comes from two *in vivo* experiments (64). When the *O2* half-site is reversed but the protein-contacting major grooves are kept on the same face of the DNA, DNA looping is abolished. A related experiment is also consistent with the holding hypothesis. *In vivo*, AraC still loops the DNA when the *I2* half-site is replaced by the *O2* half-site, thus leaving AraC free to choose between looping between *I1* and the distal *O2* or binding *cis* to the *I1* and *O2* half-sites.

Further evidence for the conclusion that the DNA-binding domains are held in positions consistent with DNA looping in the absence of arabinose and less tightly held in the presence of arabinose comes from connecting the two half-sites at *araI* with 24 bases of single stranded DNA (76) (Fig. 11). The single-stranded DNA allows both half-sites to be positioned such that both DNA binding domains can bind a half-site, independent of where they are positioned in space. If arabinose addition then increases the flexibility of AraC, and the DNA-binding domains are more free to position themselves, AraC will continue to bind, and with little change in its overall affinity for the DNA. On the other hand, if the addition of arabinose increases the intrinsic affinity of each DNA-binding domain for binding to a half-site, then the addition of arabinose will substantially increase the affinity of AraC for the split half-site DNA substrate. Arabinose was found to have no appreciable effect on the affinity of AraC for the split half-site DNA, thus indicating that arabinose increases the apparent flexibility of AraC. The same split half-site approach was used to investigate the mechanism by which cAMP binding increases the affinity of CRP for its inverted repeat half-site structure. Our experiments showed that cAMP increases the intrinsic affinity of the DNA-binding domains of CRP rather than controlling their freedom to reposition. This mechanism was later confirmed by the 3D structure of cAMP-bound CRP (77).

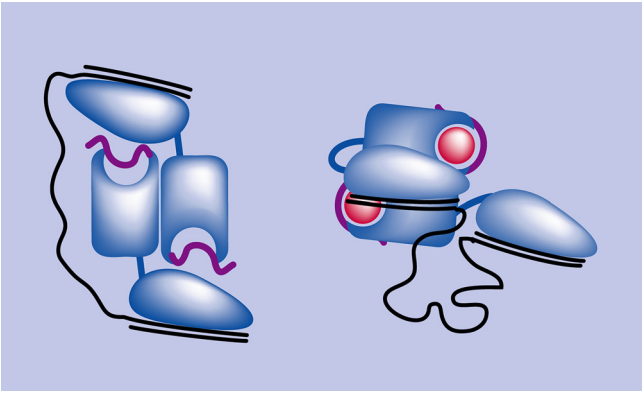

FIG 11 Determination of the mechanism of arabinose-induced change in DNA looping versus binding *cis*. If the DNA-binding domains reposition upon arabinose addition, binding remains unchanged if the two binding half-sites are connected by a flexible linker of sufficient length (single-stranded DNA). If the intrinsic affinity of the individual DNA-binding domains is changed by the binding of arabinose, the affinity of AraC for the flexibly connected DNA half-sites changes.

In plasmon resonance experiments, free N-terminal arm binds to the dimerization domain, but not the DNA-binding domain. The light switch mechanism predicts that the N-terminal arm of AraC would bind to the dimerization domain only in the presence of arabinose and bind to the DNA-binding domain in the presence and absence of arabinose. Plasmon resonance experiments are capable of real time observation of binding and dissociation between proteins from millimolar to nanomolar concentrations. We found that in the presence of arabinose, arm peptides immobilized on a glass chip bound to dimerization domain lacking an arm (78). This is as expected, as the X-ray structures of apo- and holo-AraC showed that arabinose binds in a pocket of the dimerization domain and that the arm then binds over the arabinose, and that, in the absence of arabinose, the arm makes fewer specific interactions with the dimerization domain. With respect to arm-DNA binding interactions, the plasmon resonance experiments failed to show any interaction: strike one against the light switch mechanism.

No constitutive mutations in the DNA-binding domains. Virtually all amino acid changes to the N-terminal arm result in constitutive behavior of AraC, as do deletions of the arm (79, 80). These AraC mutants still retain the ability to induce *pBAD* and can do so in the absence of arabinose, but they have lost the ability to repress. The arm must therefore play an active role in repression. As described above in the light switch mechanism, the most likely target for the arm was the DNA-binding domain. The

problem with this mechanism has been that, despite extensive searches for mutations conferring constitutivity, none have been found in the DNA-binding domain. The problem has therefore become acute. How does the arm control the looping-inducing capabilities of AraC without making specific interactions with any other part of AraC? Sections below describe experiments designed to answer or help answer this question.

The N-terminal arm does not function as an entropic bristle. Maybe AraC functions by a something other than the light switch mechanism. We know that the wild-type arm is required in order that AraC act as a repressor rather than an inducer. Thus, the arm actively does something. If not by binding to the DNA-binding domains, then perhaps the arms act as entropic bristles and push the DNA-binding domains away from the dimerization domains until the binding of arabinose causes the arms to retract and bind over the bound arabinose (Fig. 12). Suggestive of the bristle idea is that the dimerization domain aggregates into indefinite polymers in the absence of arabinose, and its tendency to do so markedly increases when the arms are deleted (56).

We devised an experimental system to directly test whether polypeptides of the size of the N-terminal arm of AraC, ~20 residues, can significantly act as entropic bristles (56). Ubiquitin is a 76-residue globular protein lacking tryptophan. It is cleaved by trypsin after Arg74, releasing the final two amino Gly-Gly residues. We replaced the residue immediately before Arg74 with tryptophan and added various lengths of random peptides, up to 24 residues, after the glycines. In this construct, cleavage at Arg74 changes the exposure of the tryptophan to water, thereby altering its fluorescence. If the added peptide tail were to act as an entropic bristle, it would reduce access of trypsin to the cleavage sequence. Cleavage at the sequence was monitored in real time by the fluorescence change of the tryptophan. As neither the ubiquitin core nor the added peptides contained tryptophan, the assay was particularly sensitive. Although an entropic effect could be detected and measured as a function of peptide length, it was much too small to play a significant role either in interfering with aggregation or in positioning the DNA-binding domains of AraC in the presence or absence of arabinose.

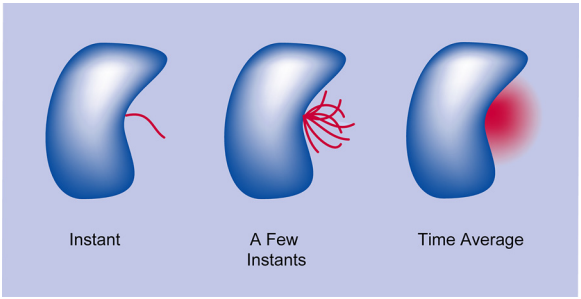

FIG 12 How an entropic bristle tends to exclude macromolecules from its vicinity.

Mutations in the interdomain linker and assay of linker flexibility. Since repression of *pBAD* requires a wild type N-terminal arm of AraC, the arm has to be doing something; thus we considered the eight-residue interdomain linker (81). We sought mutations in the arm that interfered with repression. Some constitutive mutations showed up, and almost all of them contained proline substitutions somewhere in the linker.

It seemed likely that the proline residues interfered with the formation of a rigid structure of the linker, and an alpha helix was an obvious candidate. When the linker was helical, it would be rigid and hold the DNA-binding domains in specific positions, and when the helix dissolved, the DNA-binding domains would not be rigidly held in position. We therefore tested whether arabinose changed the apparent flexibility of the linker (81). To do this, we conjugated the fluorophore IAEDANS to a cysteine residue that we substituted at the end of the linker of the dimerization domain. This permitted fluorescence anisotropy measurements to compare the apparent tumbling rate of the IAEDANS in the presence and absence of arabinose. Tumbling increased in the presence of arabinose, implying that the linker becomes more flexible in the presence of arabinose. The tumbling of IAEDANS conjugated elsewhere in the dimerization domain was unaltered by arabinose addition.

Arabinose binding to one subunit alters flexibility of the linker on the other subunit. In the dimeric AraC protein, the N-terminal arm of one subunit, which, we had learned, controls flexibility of the interdomain linker, is much closer to the linker on the other subunit than it is to the linker on the same subunit. Therefore, it seemed possible that the binding of arabinose to one subunit changed the flexibility of only the linker on the other subunit. Indeed, this proved to be the case (Fig. 13). This finding is an

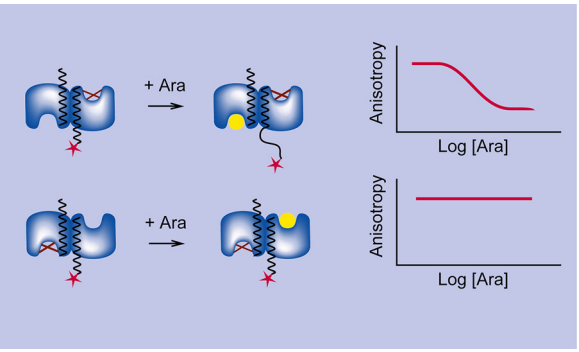

FIG 13 Heterodimers are made by adding a large excess of the unlabeled (star) dimers to fluorescently labeled dimers. After random mixing through subunit exchange, almost every labeled subunit has a nonlabeled partner. The X in the arabinose binding pocket indicates that that subunit type contains a mutation in the pocket that prevents arabinose from binding. The black coil represents the alpha helix that dimerizes the protein. Its last eight residues protrude from the dimerization domain and constitute the interdomain linker. The DNA-binding domain was not present in these experiments. Only when the fluorescent label is on the subunit opposite to the subunit that can bind arabinose does the addition of arabinose lead to a reduction in fluorescence anisotropy, indicating loss of helicity of the linker.

important component in the final model for the mechanism by which arabinose changes the flexibility of AraC.

Does the linker undergo an helix-coil transition?

AraC in the inducing state is more flexible than when it is in the repressing state. The shift from being more rigid to being more flexible could be a consequence of a helix-to-coil transition of the linker or just a weakening of a helical state of the linker that is somehow brought about by the binding of arabinose, either of which is also consistent with the behavior of the proline substitutions. We tested this idea by strengthening and weakening the tendency of the linker to form an alpha helix (82). To strengthen the helix, we substituted the helix-forming alanine residue for the first six residues of the linker and to weaken the helix, substituted the helix-destabilizing residues serine-glycine. The *in vivo* and *in vitro* activities of AraC with wild-type linker and with the two variants were consistent with the helix postulate.

A second type of test for helicity of the interdomain linker (82) is reminiscent of the helical twist experiment we devised to test for DNA looping (57). As discussed earlier, DNA looping and, therefore, repression depend on the two DNA-binding domains of AraC being fairly rigidly held in orientations favorable for looping. If the domains are misoriented

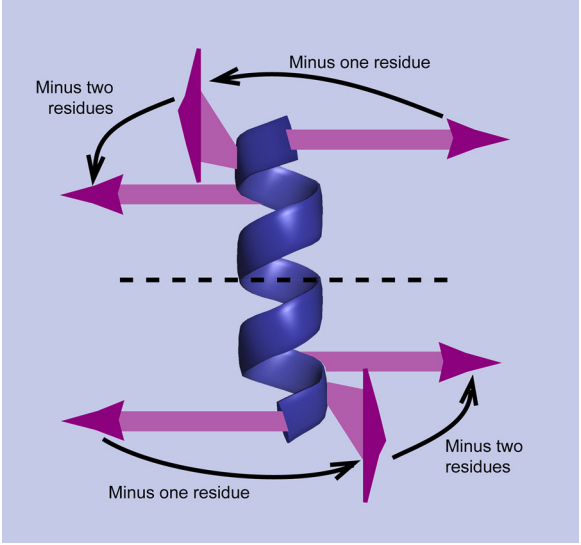

FIG 14 Protein helical twist. The dimerization helices and helical linkers point outward up and down from the protein. Here, the two linkers have retained their natural orientations but have been slid together, being separated by the dotted line. Initially the ends of the two helices point in opposite directions. Upon the deletion of a single residue from each, the ends now are parallel and point out of the paper. Upon the deletion of a second residue from each, the ends once again are roughly antiparallel.

with respect to the favorable orientation, DNA looping and repression will be inhibited. As an alpha helix contains 3.6 residues per turn and the two linkers in the dimeric AraC point in opposite directions, the insertion or deletion of a single residue in each linker will change the orientation of each DNA-binding domain by almost 90 degrees, such that the two DNA-binding domains change from pointing in opposite directions to pointing in nearly the same direction (Fig. 14). Then, the insertion or deletion of an additional residue points the two DNA-binding domains back rather close to their original antiparallel orientations. Thus, insertions or deletions of one residue ought to damage repression while insertions or deletions of two residues ought to largely restore repression. This is what we found experimentally. Unfortunately, though, hydrogen-deuterium exchange experiments, which could have provided an elegant confirmation of the helix-coil or helix-weakened helix idea, failed to provide definitive data (83).

Identifying a specific interdomain residue-residue interaction. The most useful information about the structure of AraC based on homologs has come from examination of covarying residues. In general, if a

mutation reduces the activity of a protein, most of the second site mutations that lead to a substantial restoration of activity should lie relatively close by in the tertiary structure of the protein. A search for such compensating pairs or second site suppressors could require testing huge numbers of candidate double mutations. Nature, however, through eons of evolution, has already isolated such functional double mutations. Among the sequence differences in a large set of homologs of similar function will be compensating double mutations. These will show up as residues that vary together, i.e., when one changes, the other is likely also to change. Such covarying pairs can therefore identify relationships between interacting residues, whether it is between neighboring residues in the structure, in different domains of a protein, or between different proteins. The web tool Gremlin, developed in David Baker's lab, can identify sequence homologs of an amino sequence and then identify pairs of residues with a high probability of covarying (84).

We used the Gremlin tool to seek covarying residues between the N-terminal arm of AraC and other parts of the protein and to seek covarying residues between the domains. In accord with the multiple other lines of evidence mentioned in this review, GREMLIN found covarying residues within the N-terminal arm and between the N-terminal arm and nearby residues in the dimerization domain. None were found between the arm and the DNA-binding domain—another strike against the light switch mechanism. Gremlin identified several interdomain covarying pairs between the dimerization domain and the DNA-binding domain. The presence of such covarying pairs is consistent with experiments mentioned earlier that indicated that functionally important domain-domain interactions are present in AraC. Therefore, we sought to verify that changes to one of the residues in one domain which alters the regulation properties of AraC can be corrected by a change in the covarying residue in the other domain.

We found that a mutation in residue N194 of the DNA-binding domain, which inactivated AraC, could be corrected by mutations H136 or N139 in the dimerization domain. These two residues lie on the concave face of the dimerization domains and alongside the dimerization interface. H136 on one subunit lies very close to N139 on the other subunit and both can simultaneously be contacted by N194 in one of the DNA-binding domains. Positioning a DNA-binding domain to make this interaction constrains it to lie along the dimerization domain interface. This structure (Fig. 15), and virtually no others, is compatible with AraC binding to the

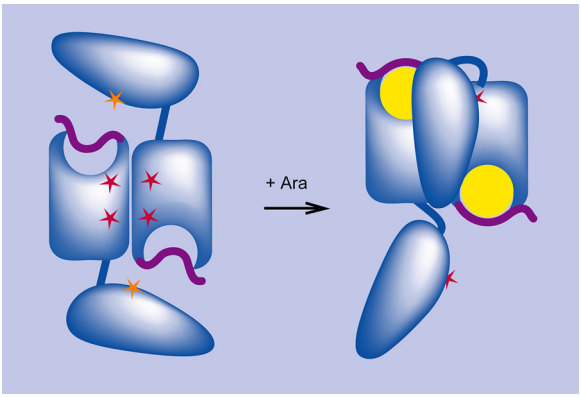

FIG 15 Approximate location of residue N194 in the dimerization domain of AraC (orange stars in the DNA-binding domains), which covaries with residues H126 and N139 (red stars in the dimerization domains). Mutations in N194 that inactivate AraC can be corrected by mutations in H126 and N139.

direct repeat *I1-I2* induction site. This mode of binding requires that when AraC binds to direct repeats of the half-sites, both DNA-binding domains of the AraC dimer are released from the positions in which they are held while the protein loops the DNA. Consistent with this requirement, we found that two molecules of arabinose must bind to AraC in order for it to bind normally to direct repeat half-sites like *I1-I2* (85).

Mechanism by which arabinose binding changes the flexibility of AraC. The unproven hypothesis, but one that is consistent with the multiple lines of evidence cited above, is that in the absence of arabinose, the N-terminal arms are structured and "push" against the β-sheets at the top of the jellyroll arabinose-binding pockets, moving each somewhat toward the dimerization interface (65, 86) (PDB entries 2ARC and 1XJA) (Fig. 16). These push against the helices that support the dimerizing α-helices, which in turn push against the dimerizing α-helices of the opposite subunit and stabilize the other subunit's interdomain linkers. Consequently, the DNA-binding domains are held in positions suitable for DNA looping.

When arabinose binds, the N-terminal arms reposition and bind over the arabinose, relieving the pressure and movements that stabilize the interdomain linker and, as a result, the helix unfolds or is at least weakened. As a result of this, the repositioning the DNA-binding domains so that they can bind to *araI1-I2* becomes energetically less costly. Consequently, the AraC protein dimer ceases looping and

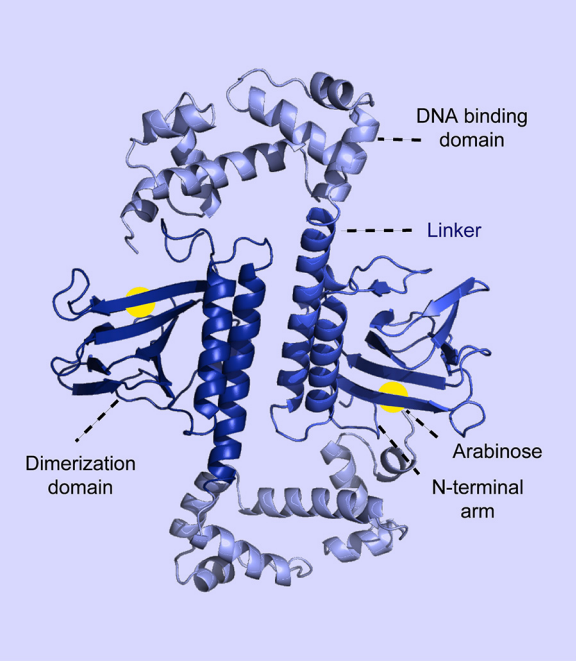

FIG 16 A cartoon representation of apo AraC. The dimerization domains are in dark blue and a guess as to the positions of the DNA-binding domains is in light blue.

binds *cis* to the *I1* and *I2* half-sites, where it stimulates RNA polymerase binding and isomerization to form an open complex. The repositioning of one of the DNA-binding domains is aided by its specific interaction with several amino acids of the dimerization domains as described above.

AraC stimulates the *pBAD* promoter by increasing the binding of RNA polymerase to the promoter and also by increasing the isomerization rate to form an open complex, both in about equal measures (42). The cyclic AMP receptor protein CRP also aids in the process of opening the loop and either increasing the binding of RNA polymerase or increasing its rate of forming an open complex.

The regulation of the *pC* promoter in the *ara pBAD* regulatory region is somewhat more complex. Experimentally we observe that the addition of arabinose increases the activity of the *pC* promoter by a factor of about five for 20 min (21–23), and then the activity of *pC* returns to its preinduction value. The initial increase in activity upon arabinose addition can be understood as resulting from the increased access of RNA polymerase to *pC*, resulting from opening the loop between *araO2* and ara*I1* (36). Turning the promoter down by the binding of AraC to *araO1*, the operator controlling *pC*, does not immediately occur however. This is because the binding of arabinose to AraC increases not only its binding affinity for its functional sites, it also increases the affinity of AraC for nonspecific binding to any DNA sequence or sequences resembling authentic sites. Consequently, the concentration of free AraC in the cell falls upon arabinose addition. This prolongs the time required for free AraC to bind to the *O1* site, but after about 20 min, probably aided by the increased levels of AraC, the protein binds to *O1* and turns down the activity of *pC*. Not surprisingly, artificially elevated levels of AraC appropriately decrease the derepression of *pC* upon arabinose addition (23).

My training and our research group. My training has been a bit unconventional. Except for a year of English, 3 years of Russian, and a year of economics, I took only mathematics, physics, and electrical engineering courses as an undergraduate. I began graduate school in physics, took the required courses, but then began a thesis research project in molecular biology in the laboratory of a physicist turned molecular biologist. Eventually, the physics department wised up and I was transferred to the biophysics department. After that, the major influences on my training are best shown in a drawing (Fig. 17).

For the first 18 years after my 3-year postdoc with Watson and Gilbert at Harvard, I was at Brandeis University. There, my research group grew and remained between three and six graduate students until my final few years at Hopkins, where it gradually shrank to just me. Although I've mentioned only a few by name, I am exceedingly grateful to the more than 40 Ph.D. students who did their research and

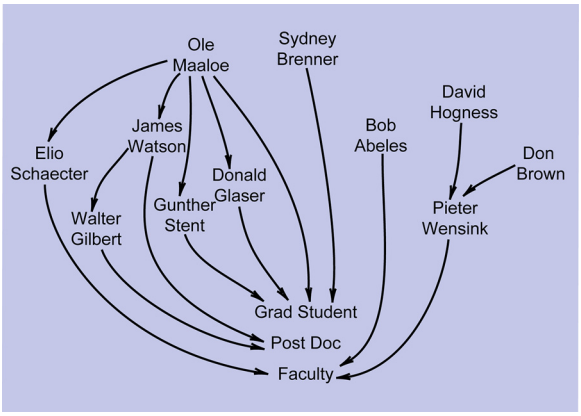

FIG 17 A genealogy tree of the major influencers on my development.

obtained their degrees in my lab. During those 5 decades, I also had eight postdoctoral research assistants and eight research technicians who have also importantly contributed to our work. Each research group is like a family and a little different from all others. Here are a few "features" of our family. In our weekly group meetings, everyone spoke every time, telling about their successes and failures over the past week, and we all tried to help. No slides were used and everyone had to learn how to convey their ideas with words and drawings made in real time on a blackboard. Because we all pursued diverse projects, this format allowed students to become deeply knowledgeable about multiple lines of research. Apparently this was successful, as the students have gone on to be highly successful in a wide variety of fields in the biological sciences. I felt it important that we be comfortable developing new techniques and using equipment in creative ways. Therefore, we did all our own equipment maintenance. (In those days, the electromechanical equipment failed much more often than contemporary equipment.) Once, the student whose responsibility was the liquid scintillation counter was repairing the elevator that lowered a vial into the counting chamber and accidently broke one of the two photomultiplier tubes. The manufacturer wanted $2,000 for a new one. We discovered however, that physicists needing matched pairs of photomultiplier tubes routinely ordered three of the tubes at $100 each from the tube manufacturer, picked the most closely matched pair, and returned the third. To keep the atmosphere in the lab as light and fun as possible, I executed, and was also the victim of, quite a few practical jokes. Finally, in order to test and develop new thesis projects, or to do experiments that no one else would touch, over almost the entire 50 years I worked at the lab bench about half of the day (and a good many evenings and weekends).

REFERENCES

1. **Gross J, Englesberg E.** 1959. Determination of the order of mutational sites governing L-arabinose utilization in *Escherichia coli* B/r bv transduction with phage Plbt. *Virology* **9:**314–331. https://doi.org/10.1016/0042-6822(59)90125-4.
2. **Englesberg E.** 1961. Enzymatic characterization of 17 L-arabinose negative mutant of *Escherichia coli*. *J Bacteriol* **81:**996–1006. https://doi.org/10.1128/jb.81.6.996-1006.1961.
3. **Jacob F, Perrin D, Sanchez C, Monod J.** 1960. Comptes Rendus Hebdomadaires Des Seances De l'Adademe. *Des Sciences* **250:**1727–1729.
4. **Englesberg E, Irr J, Power J, Lee N.** 1965. Positive control of enzyme synthesis by gene C in the L-arabinose system. *J Bacteriol* **90:**946–957. https://doi.org/10.1128/jb.90.4.946-957.1965.
5. **Gilbert W, Muller-Hill B.** 1966. Isolation of the *lac* Repressor. *Proc Natl Acad Sci U S A* **56:**1891–1898. https://doi.org/10.1073/pnas.56.6.1891.
6. **Englesberg E, Squires C, Meronk F, Jr.** 1969. The L-arabinose operon in *Escherichia coli* B/r: a genetic demonstration of two functional states of the product of a regulator gene. *Proc Natl Acad Sci U S A* **62:**1100–1107. https://doi.org/10.1073/pnas.62.4.1100.
7. **Shimada K, Weisberg RA, Gottesman ME.** 1972. Prophage lambda at unusual chromosomal locations: I. location of the secondary attachment sites and the properties of the lysogens. *J Mol Biol* **63:**483–503. https://doi.org/10.1016/0022-2836(72)90443-3.
8. **Luria S, Delbruck M.** 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28:**491–511. https://doi.org/10.1093/genetics/28.6.491.
9. **Schleif R.** 1972. Fine-structure deletion map of the *Escherichia coli* L-arabinose operon. *Proc Natl Acad Sci U S A* **69:**3479–3484. https://doi.org/10.1073/pnas.69.11.3479.
10. **Schleif R, Lis JT.** 1975. The regulatory region of the L-arabinose operon: a physical, genetic and physiological study. *J Mol Biol* **95:**417–431. https://doi.org/10.1016/0022-2836(75)90200-4.
11. **Lis JT, Schleif R.** 1975. The regulatory region of the L-arabinose operon: its isolation on a 1000 base-pair fragment from DNA heteroduplexes. *J Mol Biol* **95:**409–416. https://doi.org/10.1016/0022-2836(75)90199-0.
12. **Lis JT, Schleif R.** 1975. Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. *Nucleic Acids Res* **2:**383–389. https://doi.org/10.1093/nar/2.3.383.
13. **Hirsh J, Berg P.** 1976. Electron microscopy of gene regulation: the L-arabinose operon. *Proc Natl Acad Sci U S A* **73:**1518–1522. https://doi.org/10.1073/pnas.73.5.1518.
14. **Hirsh J, Schleif R.** 1976. High resolution electron microscopic studies of genetic regulation. *J Mol Biol* **108:**471–490. https://doi.org/10.1016/S0022-2836(76)80131-3.
15. **Hirsh J, Schleif R.** 1977. The araC promoter: transcription, mapping and interaction with the *araBAD* promoter. *Cell* **11:**545–550. https://doi.org/10.1016/0092-8674(77)90072-1.
16. **Kosiba BE, Schleif R.** 1982. Arabinose-inducible promoter from *Escherichia coli*. Its cloning from chromosomal DNA, identification as the *araFG* promoter and sequence. *J Mol Biol* **156:**53–66. https://doi.org/10.1016/0022-2836(82)90458-2.
17. **Kolodrubetz D, Schleif R.** 1981. Regulation of the L-arabinose transport operons in *Escherichia coli*. *J Mol Biol* **151:**215–227. https://doi.org/10.1016/0022-2836(81)90512-X.
18. **Kolodrubetz D, Schleif R.** 1981. L-arabinose transport systems in *Escherichia coli* K-12. *J Bacteriol* **148:**472–479. https://doi.org/10.1128/jb.148.2.472-479.1981.
19. **Reeder T, Schleif R.** 1991. Mapping, sequence, and apparent lack of function of *araJ*, a gene of the *Escherichia coli* arabinose regulon. *J Bacteriol* **173:**7765–7771. https://doi.org/10.1128/jb.173.24.7765-7771.1991.
20. **Hendrickson W, Stoner C, Schleif R.** 1990. Characterization of the *Escherichia coli araFGH* and *araJ* promoters. *J Mol Biol* **215:**497–510. https://doi.org/10.1016/S0022-2836(05)80163-9.
21. **Ogden S, Haggerty D, Stoner CM, Kolodrubetz D, Schleif R.** 1980. The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. *Proc Natl Acad Sci U S A* **77:**3346–3350. https://doi.org/10.1073/pnas.77.6.3346.
22. **Stoner CM, Schleif RF.** 1983. Transcription start site and induction kinetics of the araC regulatory gene in *Escherichia coli* K-12. *J Mol Biol* **170:**1049–1053. https://doi.org/10.1016/S0022-2836(83)80205-8.

23. Hahn S, Schleif R. 1983. In vivo regulation of the *Escherichia coli araC* promoter. *J Bacteriol* **155:**593–600. https://doi.org/10.1128/jb.155.2.593-600.1983.
24. Smith BR, Schleif R. 1978. Nucleotide sequence of the L-arabinose regulatory region of *Escherichia coli* K12. *J Biol Chem* **253:**6931–6933. https://doi.org/10.1016/S0021-9258(17)38009-2.
25. Stoner C, Schleif R. 1983. The araE low affinity L-arabinose transport promoter. Cloning, sequence, transcription start site and DNA binding sites of regulatory proteins. *J Mol Biol* **171:**369–381. https://doi.org/10.1016/0022-2836(83)90035-9.
26. Power J. 1967. The L-rhamnose genetic system in *Escherichia coli* K-12. *Genetics* **55:**557–568. https://doi.org/10.1093/genetics/55.3.557.
27. Tobin JF, Schleif RF. 1987. Positive regulation of the *Escherichia coli* L-rhamnose operon is mediated by the products of tandemly repeated regulatory genes. *J Mol Biol* **196:**789–799. https://doi.org/10.1016/0022-2836(87)90405-0.
28. Egan SM, Schleif RF. 1993. A regulatory cascade in the induction of *rhaBAD*. *J Mol Biol* **234:**87–98. https://doi.org/10.1006/jmbi.1993.1565.
29. Egan SM, Schleif RF. 1994. DNA-dependent renaturation of an insoluble DNA binding protein. Identification of the RhaS binding site at *rhaBAD*. *J Mol Biol* **243:**821–829. https://doi.org/10.1006/jmbi.1994.1684.
30. Schleif R. 2010. AraC protein, regulation of the L-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiol Rev* **34:**779–796. https://doi.org/10.1111/j.1574-6976.2010.00226.x.
31. Yanofsky C. 1981. Attenuation in the control of expression of bacterial operons. *Nature* **289:**751–758. https://doi.org/10.1038/289751a0.
32. Yanofsky C. 2000. Transcription attenuation: once viewed as a novel regulatory strategy. *J Bacteriol* **182:**1–8. https://doi.org/10.1128/JB.182.1.1-8.2000.
33. Hirsh J, Schleif R. 1973. *In vivo* experiments on the mechanism of action of L-arabinose C gene activator and lactose repressor. *J Mol Biol* **80:**433–444. https://doi.org/10.1016/0022-2836(73)90414-2.
34. Greenblatt J, Schleif R. 1971. Arabinose C protein: regulation of the arabinose operon *in vitro*. *Nat New Biol* **233:**166–170. https://doi.org/10.1038/newbio233166a0.
35. Lis JT, Schleif R. 1973. Different cyclic AMP requirements for induction of the arabinose and lactose operons of *Escherichia coli*. *J Mol Biol* **79:**149–162. https://doi.org/10.1016/0022-2836(73)90276-3.
36. Hahn S, Dunn T, Schleif R. 1984. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *J Mol Biol* **180:**61–72. https://doi.org/10.1016/0022-2836(84)90430-3.
37. Hahn S, Hendrickson W, Schleif R. 1986. Transcription of *Escherichia coli ara in vitro*. The cyclic AMP receptor protein requirement for PBAD induction that depends on the presence and orientation of the *araO2* site. *J Mol Biol* **188:**355–367. https://doi.org/10.1016/0022-2836(86)90160-9.
38. Lobell RB, Schleif RF. 1991. AraC-DNA looping: orientation and distance-dependent loop breaking by the cyclic AMP receptor protein. *J Mol Biol* **218:**45–54. https://doi.org/10.1016/0022-2836(91)90872-4.
39. Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO. 1982. Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature* **298:**447–451. https://doi.org/10.1038/298447a0.
40. Brunelle A, Schleif R. 1989. Determining residue-base interactions between AraC protein and *araI* DNA. *J Mol Biol* **209:**607–622. https://doi.org/10.1016/0022-2836(89)90598-6.
41. Carra JH, Schleif RF. 1993. Variation of half-site organization and DNA looping by AraC protein. *EMBO J* **12:**35–44. https://doi.org/10.1002/j.1460-2075.1993.tb05629.x.
42. Zhang X, Reeder T, Schleif R. 1996. Transcription activation parameters at *ara pBAD*. *J Mol Biol* **258:**14–24. https://doi.org/10.1006/jmbi.1996.0230.
43. Gilbert W, Muller-Hill B. 1967. The *lac* operator is DNA. *Proc Natl Acad Sci U S A* **58:**2415–2421. https://doi.org/10.1073/pnas.58.6.2415.
44. Muller-Hill B, Crapo L, Gilbert W. 1968. Mutants that make more *lac* repressor. *Proc Natl Acad Sci U S A* **59:**1259–1264. https://doi.org/10.1073/pnas.59.4.1259.
45. Schleif R. 1969. An L-arabinose binding protein and arabinose permeation in *Escherichia coli*. *J Mol Biol* **46:**185–196. https://doi.org/10.1016/0022-2836(69)90065-5.
46. Schleif R. 1969. Induction of the L-arabinose operon. *J Mol Biol* **46:**197–199. https://doi.org/10.1016/0022-2836(69)90066-7.
47. Doyle ME, Brown C, Hogg RW, Helling RB. 1972. Induction of the *ara* operon of *Escherichia coli* B/r. *J Bacteriol* **110:**56–65. https://doi.org/10.1128/jb.110.1.56-65.1972.
48. Zipser D, Beckwith J (ed). 1970. The *lac* operon. Cold Spring Harbor Laboratory of Quantitative Biology, Cold Spring Harbor, New York.
49. Gottesman S, Beckwith JR. 1969. Directed transposition of the arabinose operon: a technique for the isolation of specialized transducing bacteriophages for any *Escherichia coli* gene. *J Mol Biol* **44:**117–127. https://doi.org/10.1016/0022-2836(69)90408-2.
50. Steffen D, Schleif R. 1977. Overproducing *araC* protein with lambda-arabinose transducing phage. *Mol Gen Genet* **157:**333–339. https://doi.org/10.1007/BF00268671.
51. Schleif RF, Favreau MA. 1982. Hyperproduction of *araC* protein from *Escherichia coli*. *Biochemistry* **21:**778–782. https://doi.org/10.1021/bi00533a031.
52. Fried M, Crothers DM. 1981. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* **9:**6505–6525. https://doi.org/10.1093/nar/9.23.6505.
53. Garner MM, Revzin A. 1981. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res* **9:**3047–3060. https://doi.org/10.1093/nar/9.13.3047.
54. Hendrickson W, Schleif RF. 1984. Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay. *J Mol Biol* **178:**611–628. https://doi.org/10.1016/0022-2836(84)90241-9.
55. Studier FW, Moffatt BA. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* **189:**113–130. https://doi.org/10.1016/0022-2836(86)90385-2.
56. Weldon JE, Schleif RF. 2006. Specific interactions by the N-terminal arm inhibit self-association of the AraC dimerization domain. *Protein Sci* **15:**2828–2835. https://doi.org/10.1110/ps.062327506.
57. Dunn TM, Hahn S, Ogden S, Schleif RF. 1984. 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* **81:**5017–5020. https://doi.org/10.1073/pnas.81.16.5017.
58. Martin K, Huo L, Schleif RF. 1986. 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* **83:**3654–3658. https://doi.org/10.1073/pnas.83.11.3654.
59. Shore D, Langowski J, Baldwin R. 1981. DNA flexibility studied by covalent closure of short fragments into circles. *Proc Natl Acad Sci U S A* **78:**4833–4837. https://doi.org/10.1073/pnas.78.8.4833.
60. Lobell RB, Schleif RF. 1990. DNA looping and unlooping by AraC protein. *Science* **250:**528–532. https://doi.org/10.1126/science.2237403.
61. Bustos SA, Schleif RF. 1993. Functional domains of the AraC protein. *Proc Natl Acad Sci U S A* **90:**5638–5642. https://doi.org/10.1073/pnas.90.12.5638.
62. Eustance RJ, Bustos SA, Schleif RF. 1994. Reaching out. Locating and lengthening the interdomain linker in AraC protein. *J Mol Biol* **242:**330–338. https://doi.org/10.1006/jmbi.1994.1584.

63. Reeder T, Schleif R. 1993. AraC protein can activate transcription from only one position and when pointed in only one direction. *J Mol Biol* **231:**205–218. https://doi.org/10.1006/jmbi.1993.1276.
64. Seabold RR, Schleif RF. 1998. Apo-AraC actively seeks to loop. *J Mol Biol* **278:**529–538. https://doi.org/10.1006/jmbi.1998.1713.
65. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. 1997. Structural basis for ligand-regulated oligomerization of AraC. *Science* **276:**421–425. https://doi.org/10.1126/science.276.5311.421.
66. Soisson SM, MacDougall-Shackleton B, Schleif R, Wolberger C. 1997. The 1.6 Å crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *J Mol Biol* **273:**226–237. https://doi.org/10.1006/jmbi.1997.1314.
67. Timmes A, Rodgers M, Schleif R. 2004. Biochemical and physiological properties of the DNA binding domain of AraC protein. *J Mol Biol* **340:**731–738. https://doi.org/10.1016/j.jmb.2004.05.018.
68. Martin KJ, Schleif RF. 1987. Equilibrium DNA-binding of AraC protein. Compensation for displaced ions. *J Mol Biol* **195:**741–744. https://doi.org/10.1016/0022-2836(87)90193-8.
69. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* **355:**294–298. https://doi.org/10.1126/science.aah4043.
70. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577:**706–710. https://doi.org/10.1038/s41586-019-1923-7.
71. Berrondo M, Gray JJ, Schleif R. 2010. Computational predictions of the mutant behavior of AraC. *J Mol Biol* **398:**462–470. https://doi.org/10.1016/j.jmb.2010.03.021.
72. Damjanovic A, Miller BT, Schleif R. 2013. Understanding the basis of a class of paradoxical mutations in AraC through simulations. *Proteins* **81:**490–498. https://doi.org/10.1002/prot.24207.
73. Frato KE, Schleif RF. 2009. A DNA-assisted binding assay for weak protein-protein interactions. *J Mol Biol* **394:**805–814. https://doi.org/10.1016/j.jmb.2009.09.064.
74. Cole SD, Schleif R. 2012. A new and unexpected domain-domain interaction in the AraC protein. *Proteins* **80:**1465–1475. https://doi.org/10.1002/prot.24044.
75. Saviola B, Seabold R, Schleif RF. 1998. Arm-domain interactions in AraC. *J Mol Biol* **278:**539–548. https://doi.org/10.1006/jmbi.1998.1712.
76. Harmer T, Wu M, Schleif R. 2001. The role of rigidity in DNA looping-unlooping by AraC. *Proc Natl Acad Sci U S A* **98:**427–431. https://doi.org/10.1073/pnas.98.2.427.
77. Sharma H, Yu S, Kong J, Wang J, Steitz TA. 2009. Structure of apo-CAP reveals that large conformational changes are necessary for DNA binding. *Proc Natl Acad Sci U S A* **106:**16604–16609. https://doi.org/10.1073/pnas.0908380106.
78. Ghosh M, Schleif RF. 2001. Biophysical evidence of arm-domain interactions in AraC. *Anal Biochem* **295:**107–112. https://doi.org/10.1006/abio.2001.5213.
79. Ross JJ, Gryczynski U, Schleif R. 2003. Mutational analysis of residue roles in AraC function. *J Mol Biol* **328:**85–93. https://doi.org/10.1016/S0022-2836(03)00262-6.
80. Dirla S, Chien JY, Schleif R. 2009. Constitutive mutations in the *Escherichia coli* AraC protein. *J Bacteriol* **191:**2668–2674. https://doi.org/10.1128/JB.01529-08.
81. Malaga F, Mayberry O, Park DJ, Rodgers ME, Toptygin D, Schleif RF. 2016. A genetic and physical study of the interdomain linker of *E. coli* AraC protein—a trans-subunit communication pathway. *Proteins* **84:**448–460. https://doi.org/10.1002/prot.24990.
82. Brown MJ, Schleif RF. 2019. Helical behavior of the interdomain linker of the *Escherichia coli* AraC protein. *Biochemistry* **58:**2867–2874. https://doi.org/10.1021/acs.biochem.9b00234.
83. Tischer A, Brown MJ, Schleif RF, Auton M. 2019. Arabinose alters both local and distal H–D exchange rates in the *Escherichia coli* AraC transcriptional regulator. *Biochemistry* **58:**2875–2882. https://doi.org/10.1021/acs.biochem.9b00389.
84. Kamisetty H, Ovchinnikov S, Baker D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **110:**15674–15679. https://doi.org/10.1073/pnas.1314045110.
85. Rodgers ME, Schleif R. 2012. Heterodimers reveal that two arabinose molecules are required for the normal arabinose response of AraC. *Biochemistry* **51:**8085–8091. https://doi.org/10.1021/bi3005347.
86. Weldon JE, Rodgers ME, Larkin C, Schleif RF. 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins* **66:**646–654. https://doi.org/10.1002/prot.21267.