Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 2152–2156, April 1989
Biochemistry

Identifying determinants of folding and activity for a protein of unknown structure

(Arc repressor/DNA binding/protein folding/secondary structure prediction/mutagenesis)

JAMES U. BOWIE AND ROBERT T. SAUER

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by David Botstein, December 21, 1988

ABSTRACT **We have generated an extensive genetic map of functionally allowed and/or structurally allowed amino acid substitutions in Arc repressor, a DNA binding protein of unknown structure. Analysis of the allowed substitution patterns identifies residues that are likely to be involved in protein function and identifies side chains that play important structural roles, including residues likely to form the hydrophobic core. The identities of approximately one-third of the residues in Arc repressor are functionally important, about one-half are structurally important, and the remainder are unimportant for either structure or function. The patterns of obligatory hydrophobic positions permit strong predictions of secondary structure.**

With the advent of recombinant DNA techniques for modifying protein sequences, considerable work has been directed toward specifically altering or improving existing proteins. Usually, such studies have focused on proteins of known structure, where predictions can be more reasonably made and tested. Unfortunately, the structures of most proteins are not known. Studies of proteins of known structure have, however, shown that the types of substitutions accommodated at a residue position depend on the structural or functional importance of the side chain (1–3). Hence, it should be possible to use substitution patterns in a protein of unknown structure to probe the roles and importance of individual residues. Here, we identify substitutions that are structurally allowed and functionally allowed at each of the 53 residue positions of the phage P22 Arc repressor. Arc repressor is a sequence-specific DNA binding protein of unknown structure that is dimeric in solution and tetrameric when bound to DNA (refs. 4–6; B. Brown and J.U.B., unpublished data). Analysis of the allowed substitution patterns in Arc indicates the general role of each residue, suggests whether each side chain is solvent exposed or buried in the structure, and reveals regions of likely α-helix and β-structure. We propose that the methods and analysis applied here should be broadly applicable to problems of protein engineering and protein folding.

MATERIALS AND METHODS

Strains and Mutagenesis. *Escherichia coli* strains used in this work were UA2F and X90 (6). Plasmid pSA300 contains a synthetic *arc* gene (Fig. 1) under control of a *tac* promoter, a str^s gene under the control of an Arc repressible promoter, and an M13 origin of replication that allows the production of single-stranded plasmid DNA for sequencing (7). Details of the plasmid construction are available upon request. Six mutagenic "cassettes" spanning restriction sites in the synthetic gene were prepared by the method of Oliphant *et al.* (8). Regions within the cassettes were mutagenized by contaminating each base during the synthesis with 7.5% of each of the other three nucleotides.

Screens and Selections. To screen clones for high levels of inducible protein, UA2F/pSA300 cells were grown at 37°C in LB medium plus ampicillin (100 μg/ml) and Arc expression from the P_{tac} promoter was induced by the addition of isopropyl β-D-thiogalactopyranoside. After 3 hr, cells were harvested, lysed by heating in SDS, and portions were electrophoresed on 15% Laemmli gels (9). The selection for Arc activity takes advantage of the fact that a strain containing a str^s gene and a str^r gene is sensitive to streptomycin. Transcription of the str^s allele in the selection strain (UA2F) is repressed by Arc. Hence, cells containing active Arc are resistant to streptomycin (7). UA2F also contains a fusion of an Arc-repressible promoter to chloramphenicol acetyltransferase. Thus, cells containing active Arc are chloramphenicol sensitive (6). A mutant Arc protein appears to need at least 2–10% of the wild-type activity to pass the selection.

Protein Stability and Spectroscopy. Mutant proteins were purified by methods similar to those described (6, 7). Circular dichroism (CD) spectra of Arc proteins at a concentration of 8 μM in 10 mM potassium phosphate, pH 7.5/100 mM KCl were recorded at 20°C with an AVIV model 60DS spectropolarimeter. Thermal denaturations were monitored by CD using the buffer and protein concentrations described above. Fluorescence spectra were recorded and guanidine hydrochloride (Gdn·HCl) denaturations were performed at room temperature. For these experiments, the Arc concentration was 16 μM in 10 mM Tris·HCl, pH 7.5/50 mM KCl. The unfolding reaction for Arc is a two-state transition from folded dimer (A_2) to unfolded monomers (2U). Unfolding free energies in the absence of denaturant were determined by linear extrapolation of ΔG_u values calculated in the transition zone of the Gdn·HCl denaturation curves, where $\Delta G_u = -RT\ln([U]^2/[A_2])$.

Hydrophobic Moments. Sequence positions were categorized according to the hydrophobicity of the wild-type residue or the allowed substitutions using the hydrophobicity scale of Fauchere and Pliska (10). Positions where residues more hydrophilic than glycine were structurally accommodated were assigned a value of −1 and the remaining positions were assigned a value of +1. Since arginine and lysine can play ambiguous roles, these residues were not included unless they were the only residue accommodated at the position. Helical hydrophobic moments (11) were calculated by using an eight-residue window and vectors projecting radially every 100°. The β-strand moments used a six-residue window and vectors every 180°.

RESULTS

Experimental Design. The goal of these experiments was to identify the types of allowed sequence changes at each residue position in the sequence of Arc repressor. Random mutations were generated by a cassette method in a synthetic *arc* gene. For each cassette, 7–11 adjacent codons were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

```
               1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16
              Met Lys Gly Met Ser Lys Met Pro Gln Phe Asn Leu Arg Trp Pro Arg

CTCGAGGTGAAT  ATG AAA GGA ATG AGC AAA ATG CCG CAG TTC AAC CTG AGG TGG CCG CGG
GAGCTCCACTTA  TAC TTT CCT TAC TCG TTT TAC GGC GTC AAG TTG GAC TCC ACC GGC GCC
------                                                                  ------
Xho I                                                                   Sac II

17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
Glu Val Leu Asp Leu Val Arg Lys Val Ala Glu Glu Asn Gly Arg Ser Val Asn Ser Glu

GAA GTT CTA GAT TTG GTA CGC AAG GTA GCG GAA GAG AAT GGT AGA TCT GTG AAT TCT GAG
CTT CAA GAT CTA AAC CAT GCG TTC CAT CGC CTT CTC TTA CCA TCT AGA CAC TTA AGA CTC
        -------                                         ------  -------
          Xba I                                         Bgl II   EcoR I

37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
Ile Tyr Gln Arg Val Met Glu Ser Phe Lys Lys Glu Gly Arg Ile Gly Ala

ATT TAT CAA CGC GTA ATG GAA AGC TTT AAG AAG GAA GGG CGC ATT GGC GCG TAA TCGAT
TAA ATA GTT GCG CAT TAC CTT TCG AAA TTC TTC CTT CCC GCG TAA CCG CGC ATT AGCTA
            --------        --------                                  ------
              Mlu I          Hind III                                  Cla I
```

FIG. 1. Sequence of the synthetic *arc* gene in plasmid pSA300.

mutagenized so that each codon had a roughly 40% chance of encoding a mutant amino acid. The library of randomly mutagenized cassettes was ligated into an appropriate plasmid backbone to reconstruct the gene, and the resulting plasmids were introduced into a selection strain by transformation. Using separate selections and screens, candidates displaying an Arc$^+$ phenotype or encoding stable Arc protein were then identified, and the sequences of the corresponding *arc* genes were determined by DNA sequencing. This basic experiment was repeated for different cassette regions until information was available for the entire gene. Because the method of mutagenesis ensures heavy mutagenesis of each targeted codon, amino acid substitutions should be easy to find if they are allowed at a given sequence position. Conversely, when only the wild-type residue is recovered following selection, it is reasonable to infer that other mutations are deleterious and have been selected against.

Functionally Allowed Substitutions. Plasmid genes encoding functional Arc repressor were identified by applying a biological selection and active *arc* genes were isolated and

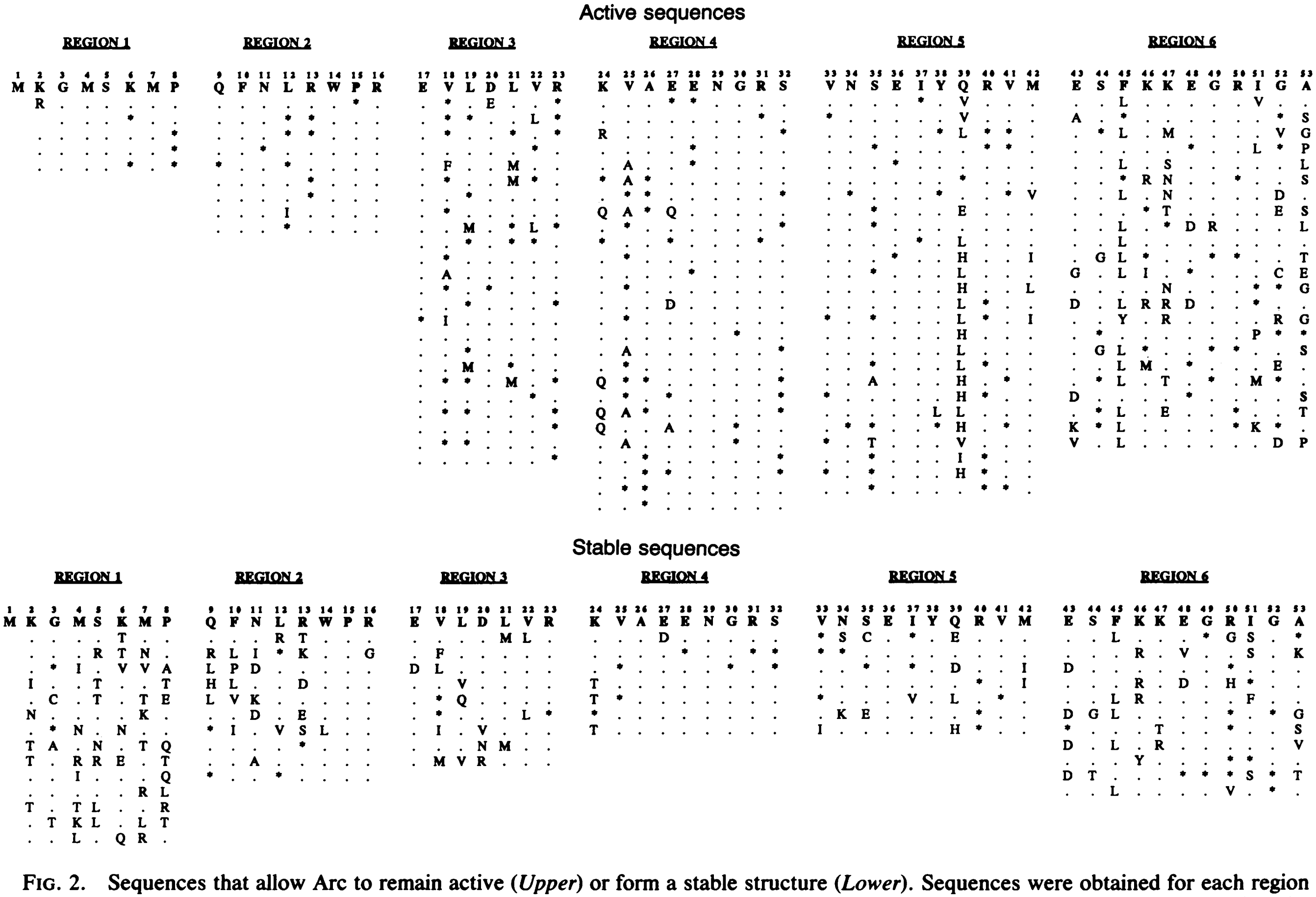

FIG. 2. Sequences that allow Arc to remain active (*Upper*) or form a stable structure (*Lower*). Sequences were obtained for each region in separate experiments. Each line represents a unique DNA sequence isolate. Asterisks represent silent base changes that do not alter the encoded amino acid. Dots represent the wild-type codon. For technical reasons, only the first position of codon 16 was mutagenized. Amino acids are designated by the single-letter code.

sequenced. Fig. 2 (*Upper*) shows the functional Arc sequences recovered in these experiments, grouped according to the regions mutagenized. In the C-terminal portion of Arc, many positions are able to tolerate significant sequence changes without loss of repressor function. We conclude that most of the side chains in this part of the protein do not play a significant role in either the structure or function of Arc. A different pattern emerges for the N-terminal and central regions of Arc. Here, only the wild-type residue was recovered at most of the positions following the functional selection. Moreover, at many of the remaining positions, the substitutions that are functionally tolerated are conservative with respect to the charge, size, or hydrophobicity of the side chain. Thus, the chemical identities of most residues in the N-terminal and central regions of Arc are apparently important in allowing Arc to function.

The experiments described so far have identified a set of different but related Arc protein sequences that can fold into the same structure and perform the same function. This set of sequences is similar to a set of phylogenetically related sequences, such as the cytochromes. In both cases, one finds that some residues are invariant, some are highly conserved, and others can vary freely. The variability of a side chain reflects the importance of the position with respect to activity. However, since activity demands both functional and structural integrity, it is not possible to use this kind of sequence information to identify the role of the important residues. To make this distinction, we removed the functional requirement by applying a second screen that requires only structural stability.

Structurally Allowed Substitutions. Arc mutants that are structurally unstable do not accumulate to high steady-state levels in the cell because they are rapidly proteolyzed (6, 7). Hence, Arc sequences that can fold into stable structures should be identifiable by screening transformants for moderate to high steady-state levels of Arc by SDS gel electrophoresis of crude cell lysates. Experiments of this type were performed with the same random sequence pools described above. Fig. 2 (*Lower*) shows the sequences, arranged by region, that allow Arc to fold into a protease-resistant structure.

Many sequence positions that were refractory to substitution when activity was required, became tolerant when the requirement for function was removed. To confirm that the amino acid substitutions present in the protease-resistant Arc molecules are indeed compatible with a folded protein structure, we purified 10 proteins that, as a set, contain sequence changes at 28 of the 53 residue positions of Arc. The set of purified proteins include nonconservative amino acid substitutions at all the sequence positions where only conservative substitutions were found in the active protein sequences.

The solution structures of the purified variant proteins were assayed at room temperature by monitoring their fluorescence and CD spectral properties. The fluorescence spectrum of Arc is dominated by the unique tryptophan residue at position 14, which is buried in a hydrophobic environment in the folded protein, and the CD spectrum is characteristic of a predominantly α-helical protein. Fig. 3 shows the CD spectra of the wild type and two of the most unstable (see below) mutant proteins. The spectra are extremely similar. In like fashion, the fluorescence and CD spectra of each of the variant proteins were found to be comparable to that of the wild-type protein. Clearly then, each of the variant proteins is able to assume a folded conformation similar to that of the wild-type protein.

The thermodynamic stabilities of the folded variant proteins were assayed by denaturation experiments. Table 1 shows that the variants have stabilities that range from 2 kcal/mol less stable to 3 kcal/mol more stable than wild type (1 cal = 4.184 J). While these values may seem large, it is important to consider that the contribution of each mutant side chain is counted twice, since the unfolding reaction is a concerted transition from the dimer to unfolded monomers. Thus, the stability changes per subunit are similar to those reported for other proteins with substitutions of surface residues (12). However, the important point is that the mutant proteins are still able to assume the basic three-dimensional structure of the wild-type protein. Clearly then, the Arc side chains that can be altered in the set of protease-resistant variants are not essential for folding, although some must play a role in determining the precise stability of the folded protein.

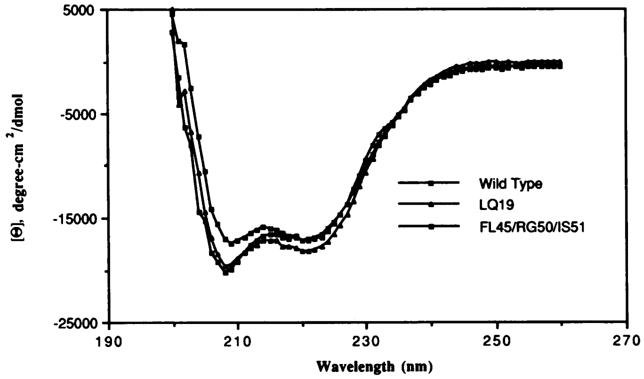

FIG. 3. CD spectra of wild-type Arc and two mutant proteins (sequence changes are designated by the single-letter code).

Fig. 4 lists the substitutions that we find to be compatible with a folded Arc protein. Note that because function requires structural integrity, the functionally allowed substitutions (boxed in Fig. 4) are a subset of the structurally allowed substitutions.

DISCUSSION

There are three important features of the methods that have been applied here. First, neutral mutants, not defective mutants, are studied. Second, levels of mutagenesis are used that are high enough to allow inferences from a negative result—i.e., the failure to recover anything but the wild-type residue at a given position. Third, separate lists are compiled of variant sequences that allow the protein to either function or to fold stably. In the discussion that follows, we describe how the information obtained can be used to identify residues likely to be key elements of the protein sequence and often

Table 1. Stability of Arc proteins to denaturation

Protein	t_m, °C	[Gdn·HCl]$_{1/2}$, M	$\Delta\Delta G_u$, kcal/mol
KT-2/MR-4/SR-5/KE-6/PT-8	67	2.0	2.8 ± 0.5
NS-34/SC-35/QE-39	63	—	—
KT-2/GA-3/SN-5/MT-7/PQ-8	60	1.5	2.9 ± 1.1
QR-9/FL-10/NI-11/RK-13/RG-16	54	1.3	0.4 ± 0.3
Wild type	54	1.2	0
VM-18/LV-19/DR-20	46	—	—
KT-24	44	1.0	−1.2 ± 0.3
KR-46/EV-48/IS-51/AK-53	37	0.9	−0.9 ± 0.4
FL-45/RG-50/IS-51	35	0.9	−1.2 ± 0.7
LR-12/RT-13	—	0.9	−1.8 ± 0.4
LQ-19	35	0.6	−1.8 ± 0.3

Sequence changes are designated by the single-letter code. Note that most mutants contain multiple substitutions. The Gdn·HCl concentrations and temperatures required to half denature the proteins are listed. ΔG_u values were determined from Gdn·HCl denaturation experiments. At a standard state of 1 M and 25°C, the folding of the wild-type Arc protein is favored over the unfolded state by ≈11 kcal/mol. $\Delta\Delta G_u = \Delta G_{\text{wild type}} - \Delta G_{\text{mutant}}$.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
MET	**LYS**	**GLY**	**MET**	**SER**	**LYS**	**MET**	**PRO**	**GLN**	**PHE**	**ASN**	**LEU**	**ARG**	**TRP**
	$\boxed{\text{ARG}}$	CYS	ILE	ARG	THR	ASN	ALA	ARG	LEU	ILE	$\boxed{\text{ILE}}$	LYS	LEU
	ILE	ALA	ASN	THR	VAL	VAL	THR	LEU	PRO	ASP	VAL	ASP	
	ASN	THR	ARG	ASN	ASN	THR	GLU	HIS	VAL	LYS	MET	GLU	
	THR		THR	LEU	GLU	LYS	GLN		ILE	ALA	ARG	SER	
			LYS		GLN	ARG	LEU					THR	
			LEU			LEU	ARG						

15	16	17	18	19	20	21	22	23	24	25	26	27	28
PRO	**ARG**	**GLU**	**VAL**	**LEU**	**ASP**	**LEU**	**VAL**	**ARG**	**LYS**	**VAL**	**ALA**	**GLU**	**GLU**
	GLY	ASP	$\boxed{\text{PHE}}$	$\boxed{\text{MET}}$	$\boxed{\text{GLU}}$	$\boxed{\text{MET}}$	$\boxed{\text{LEU}}$		$\boxed{\text{ARG}}$	$\boxed{\text{ALA}}$		$\boxed{\text{GLN}}$	
			$\boxed{\text{ALA}}$	VAL	VAL				$\boxed{\text{GLN}}$	ILE		$\boxed{\text{ALA}}$	
			$\boxed{\text{ILE}}$	GLN	ASN				THR			$\boxed{\text{ASP}}$	
			LEU		ARG								
			MET										

29	30	31	32	33	34	35	36	37	38	39	40	41	42
ASN	**GLY**	**ARG**	**SER**	**VAL**	**ASN**	**SER**	**GLU**	**ILE**	**TYR**	**GLN**	**ARG**	**VAL**	**MET**
				ILE	SER	$\boxed{\text{ALA}}$		VAL	$\boxed{\text{LEU}}$	$\boxed{\text{VAL}}$			$\boxed{\text{VAL}}$
					LYS	$\boxed{\text{THR}}$				$\boxed{\text{LEU}}$			$\boxed{\text{ILE}}$
						CYS				$\boxed{\text{GLU}}$			$\boxed{\text{LEU}}$
						GLU				$\boxed{\text{HIS}}$			
										$\boxed{\text{ILE}}$			
										ASP			

43	44	45	46	47	48	49	50	51	52	53
GLU	**SER**	**PHE**	**LYS**	**LYS**	**GLU**	**GLY**	**ARG**	**ILE**	**GLY**	**ALA**
$\boxed{\text{ALA}}$	$\boxed{\text{GLY}}$	$\boxed{\text{LEU}}$	$\boxed{\text{ARG}}$	$\boxed{\text{MET}}$	$\boxed{\text{ASP}}$	$\boxed{\text{ARG}}$	GLY	$\boxed{\text{VAL}}$	$\boxed{\text{VAL}}$	$\boxed{\text{SER}}$
$\boxed{\text{GLY}}$	THR	$\boxed{\text{TYR}}$	$\boxed{\text{ILE}}$	$\boxed{\text{SER}}$	VAL		HIS	$\boxed{\text{LEU}}$	$\boxed{\text{ASP}}$	$\boxed{\text{GLY}}$
$\boxed{\text{ASP}}$			$\boxed{\text{MET}}$	$\boxed{\text{ASN}}$			VAL	$\boxed{\text{PRO}}$	$\boxed{\text{GLU}}$	$\boxed{\text{PRO}}$
$\boxed{\text{LYS}}$			TYR	$\boxed{\text{THR}}$				$\boxed{\text{MET}}$	$\boxed{\text{CYS}}$	$\boxed{\text{LEU}}$
$\boxed{\text{VAL}}$				$\boxed{\text{ARG}}$				$\boxed{\text{LYS}}$	$\boxed{\text{ARG}}$	$\boxed{\text{THR}}$
				$\boxed{\text{GLU}}$				SER		$\boxed{\text{GLU}}$
								PHE		VAL

FIG. 4. Amino acid changes that allow the formation of a folded Arc structure. Those changes known to be functionally tolerated are boxed.

the likely role of a given residue in structure or function. We also describe how the analysis of sequence substitutions can be used to facilitate predictions of secondary structure.

Analysis of Structural Roles. At 27 positions in Arc, residue substitutions that dramatically alter the properties of the side chain can be structurally accommodated. These side chains are obviously not crucial elements of Arc structure or stability. At the remaining 25 positions, only the wild-type residue or conservative changes are structurally allowed. We infer that these residues are the primary determinants of Arc structure.

How do the structurally important residues determine or stabilize the folded form of Arc? Pro-15 and Gly-30 may be involved in turns or special backbone conformations, since proline and glycine residues often play these roles in proteins of known structure. Arg-23, Glu-28, Arg-31, Glu-36, and Arg-40 may participate in tertiary hydrogen bonds or ion pairs that stabilize the Arc structure. Phe-10, Leu-12, Trp-14, Val-18, Leu-21, Val-22, Val-25, Ala-26, Val-33, Ile-37, Tyr-38, Val-41, Met-42, Phe-45, and Lys-46 may be important for formation of the hydrophobic core of Arc. At these positions, only hydrophobic side chains or arginine or lysine (which contain long aliphatic regions and can substitute for hydrophobic residues under some circumstances) are found. It has long been established that buried positions strongly prefer nonpolar residues (1–3, 13, 14) and thus it seems likely that most of these 15 side chains of Arc are buried in the monomer or in the dimer interface. Conversely, the 27 positions that accommodate dramatic side-chain substitution are likely to be on the surface of the Arc protein since it is known that most surface side chains can be freely substituted (1–3, 15).

Prediction of Secondary Structure. Since hydrophilic side chains are commonly found on the surface of proteins but rarely in the protein interior, α-helices and β-strands sometimes show characteristic patterns of hydrophilic and hydrophobic residues that match the periodicity of the secondary structure (11, 16). For β-strands, this is a simple alternation of polar and nonpolar residues, while for α-helices, it is a more complicated pattern that matches the 3.6 residue per turn periodicity of the helix. To look for these patterns in Arc, we used the lists of structurally allowed substitutions to reveal positions that can accommodate hydrophilic residues. Fig. 5A shows hydrophobic moment plots for a composite Arc sequence. The magnitude of the hydrophobic moment in a region reflects the degree of amphiphilic character the sequence possesses in a given secondary structure. There is one short region of potential β-structure (residues 9–14) and two reasonably long regions of potential α-helix centered near residues 22 and 41. The most hydrophilic residues allowed at positions 16–28 and 35–47 are shown in helical wheel projections in Fig. 5B. The clustering of hydrophobic residues on one side of the wheel and hydrophilic residues on the other side is expected for an α-helix, with one side forming part of the hydrophobic core and the other side facing toward solvent. Recent NMR results have in fact confirmed that these regions are helical and that the N-terminal segment highlighted in the hydrophobic moment plot is in a β-conformation (M. Zagorski and D. Patel, personal communication).

Side Chains Involved in Function. The substitution patterns at a number of residue positions are consistent with the pattern expected for functionally important residues—i.e., they tolerate nonconservative substitutions when structure is demanded but tolerate only conservative substitution when both structure and repressor function is required. These residues may contact the DNA directly or may be involved

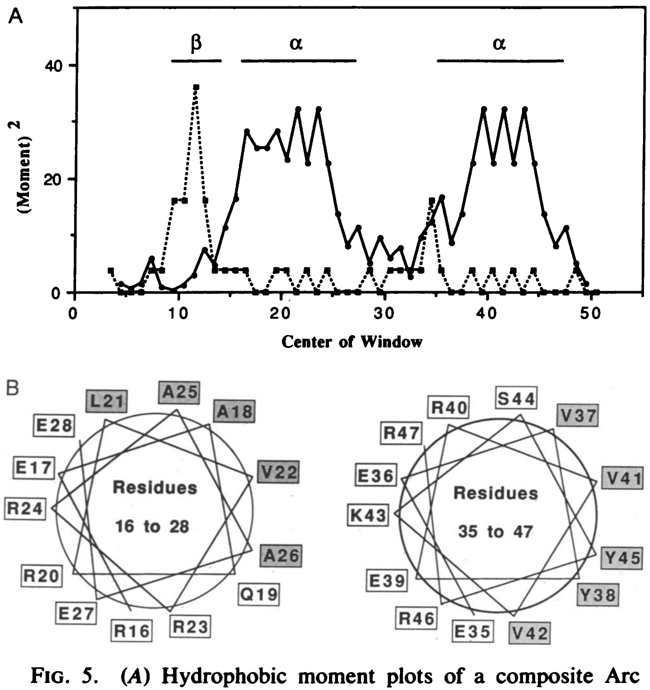

FIG. 5. (*A*) Hydrophobic moment plots of a composite Arc sequence assuming an α-helical conformation (solid line) or β-strand conformation (dashed line). The squares of the moments are plotted to highlight regions of greater amphiphilic character. (*B*) Helical wheel projections of two regions of Arc. The most hydrophilic residue observed at each position is plotted. Amino acids are designated by the single-letter code.

in stabilizing interactions between subunits in the bound tetramer. Ten of these positions are in the N-terminal third of Arc. In contrast, for the remainder of the protein, only Asp-20, Asn-34, Glu-48, and Arg-50 have the properties expected of functionally important residues. It would appear then that the N-terminal third of the protein plays a key functional role and may constitute a significant part of the DNA binding surface of Arc. This interpretation is supported by the fact that Arc mutants with substitutions at residues 2, 3, 4, 5, 8, and 10 have severely reduced operator affinity but are stably folded (6) and by the finding that a hybrid protein, containing the nine N-terminal residues of Arc fused to a related repressor, has the operator binding specificity of Arc (17).

Strengths and Weaknesses of the Method. Experiments of the type described here can be performed quite rapidly. For example, ≈3 months was required to complete these studies for Arc. However, to achieve this rapidity and still collect information about every sequence position, certain tradeoffs were necessary. For example, many residues are heavily mutagenized at the same time so that several substitutions may be seen in a particular sequence. While this allows more substitutions to be identified with fewer sequences, it raises the concern that some substitutions are only acceptable in combination with other substitutions. This may be true in some cases, but we doubt that it is a general problem because several different changes were recovered at most positions and the changes often occur in combination with a variety of other substitutions. Hence, we believe that most of the substitutions identified in multiply mutant backgrounds would also be allowed as single substitutions. A second problem concerns the spectrum of accessible substitutions at each position. Because many residues are mutagenized at once, it is not possible to fully randomize each codon and still recover a reasonable fraction of active sequences. Consequently, lower levels of mutagenesis must be used and, in these experiments, single base changes within a given codon are ≈4-fold more probable than double changes, and triple changes will be quite rare. Clearly, some substitutions will not be recovered simply because they are not sufficiently probable and some of the residues that are now grouped in the structurally or functionally important classes might be found to tolerate nonconservative substitution in a more exhaustive study. However, for most codons, the single and double changes do encode a reasonable spectrum of conservative and nonconservative residue changes so that if nonconservative changes were allowed, they could have been recovered. A final problem concerns the distinction between structurally and functionally important positions since these groups are not mutually exclusive sets. For example, local distortions of the structure could have significant functional consequences without causing global structural instability, and several candidates for DNA binding residues may be of this type.

Because of these concerns, there are gradations in the confidence levels of assignments at individual positions that are not easily quantified. However, we believe that the overall analysis of allowed residue substitutions provides an excellent overview of Arc structure–function relationships for several reasons: (*i*) Many of the residues identified by this analysis as likely to be important for DNA binding are known to be important from other studies (6, 17). (*ii*) The secondary structures predicted from the patterns of likely buried residues identified in this study are now known from NMR studies to be essentially correct (M. Zagorski and D. Patel, personal communication). (*iii*) Many of the 25 positions identified here as being structurally important are sites at which defective substitutions have been shown to result in proteolytic instability of Arc (6) and at which the same residue or conservative substitutions are found in Mnt, a homologous repressor.

The analysis of neutral mutations described here can be extremely useful for rapidly identifying residues likely to be key elements of the protein sequence so that more rigorous analysis can then be focused on important features of the sequence. Thus, our approach should be a useful first step in the characterization of structure–function relationships for other proteins. Obviously Arc is smaller than most proteins but, even in its current form, the experiments could be reasonably performed on larger proteins in short order. The primary requirement is a rapid screen for structure or activity. The analysis of neutral substitution data could also be used more sparingly to focus on a subset of residues, such as hydrophobic positions or residues in or near an enzyme active site, substantially reducing the number of residues that need to be mutagenized. Clearly, methods other than those used here could be used to generate neutral substitution data. For example, Miller (18) has shown that extensive maps of defective and neutral substitutions can be generated by the suppression of amber mutations.

We thank John Reidhaar-Olson, Neil Clarke, Peter Kim, Mike Mossing, Bronwen Brown, Robin Kelley, Dawn Parsell, Dawn Farrell, Kevin Shoemaker, and Jim Hu for ideas, help, and critical reading of the manuscript. This work was supported by National Institutes of Health Grant AI-15706.

1. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–768.
2. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196**, 199–216.
3. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241**, 53–57.
4. Susskind, M. M. (1980) *J. Mol. Biol.* **138**, 685–713.
5. Vershon, A. K., Youderian, P., Susskind, M. M. & Sauer, R. T. (1985) *J. Biol. Chem.* **260**, 12124–12129.
6. Vershon, A. K., Bowie, J. U., Karplus, T. M. & Sauer, R. T. (1986) *Proteins: Struct. Funct. Genet.* **1**, 302–311.
7. Bowie, J. U. & Sauer, R. T. (1989) *J. Biol. Chem.*, in press.
8. Oliphant, A. R., Nussbaum, A. L. & Struhl, K. (1986) *Gene* **44**, 177–183.
9. Laemmli, U. K. (1970) *Nature (London)* **227**, 680–685.
10. Fauchere, J.-L. & Pliska, V. (1983) *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369–375.
11. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982) *Nature (London)* **299**, 371–374.
12. Alber, T., Sun, D. P., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987) *Nature (London)* **330**, 41–46.
13. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171**, 479–488.
14. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136**, 225–270.
15. Hecht, M. H., Sturtevant, J. M. & Sauer, R. T. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5685–5689.
16. Schiffer, M. & Edmundson, A. B. (1967) *Biophys. J.* **7**, 121–135.
17. Knight, K. L. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 797–801.
18. Miller, J. H. (1979) *J. Mol. Biol.* **131**, 249–258.