Identifying Determinants of Folding and Activity for a Protein of Unknown Structure

by

James Ulrich Bowie

B. A., Chemistry, Carleton College
(1981)

Submitted to the Department of Biology in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

May, 1989

Signature of Author

_____

Certified by

_____

Robert T. Sauer
Thesis Supervisor

Accepted by

_____

Robert T. Sauer
Chairman
Department of Biology Graduate Committee

ABSTRACT

This thesis describes the isolation and analysis of sequence changes in the bacteriophage P22 Arc repressor protein, a small sequence specific DNA binding protein.

Chapter 1 discusses how proteins tolerate amino acid substitutions. A knowledge of allowed sequence variation can reveal the essential features of a protein sequence and leads to useful simplifications of the sequence.

Chapter 2 describes the generation of an extensive genetic map of functionally allowed and/or structurally allowed amino acid substitutions in Arc repressor. An analysis of the allowed substitution patterns identifies residues that are likely to be involved in Arc function and identifies side chains that play important structural roles, including residues likely to form the hydrophobic core. The patterns of obligatory hydrophobic positions permit strong predictions of secondary structure. Appendix I summarizes the amino acid substitutions identified in Arc and any relevant biochemical characterization that has been performed. The likely role of each residue for Arc repressor activity is discussed. Appendix II describes an attempt to identify proteins distantly related to Arc repressor by combining information from allowed amino acid substitutions into a homology search. In this manner, sequence similarity to TraY, another small DNA binding protein was identified.

Chapter 3 concerns the use of hydrophobic patterns in sets of related sequences to identify the tertiary fold adopted by the proteins in the set. Buried residues generally remain hydrophobic over the course of evolution whereas surface residues readily accept hydrophilic substitution. Consequently, the pattern of hydrodrophilicity allowed in a set of related sequences must be related to a pattern of solvent accessibility in the protein structure. If these patterns are characteristic of a particular fold, it should be possible to identify the structure a sequence family adopts, by finding a pattern of solvent accessibility in the database of known structures that best matches the observed pattern of hydrophobicity in the sequence set. An algorithm to test this possibility was created and it was found that in many cases, the similarity of these patterns is indeed sufficient to identify the tertiary fold adopted by a set of protein sequences. This project was a collaborative effort with Neil Clarke, a post doctoral fellow in Carl Pabo's laboratory at the Johns Hopkins University. I originated the basic concept and wrote the programs to search for the best alignments. Neil converted the database of known protein structures to solvent accessibility strings and secondary structures. We both contributed equally to methods development.

Chapter 4 describes the isolation and characterization of revertants of defective mutants in Arc repressor. Five of the six reverting mutations were frameshifts near the end of the coding sequence which resulted in proteins with C-terminal extensions. Each of the reverting mutations was found to prolong the half-lives *in vivo* of the proteins in which they reside, yet they do not alter the thermodynamic stability, structure, oligomeric form or DNA binding properties of these proteins. Fusion of one of these tails to the C-terminal end of a mutant form of the N-terminal domain of

lambda repressor also prevented proteolysis of this protein. These C-terminal sequences may prevent degradation by blocking the recognition of unstable proteins by the proteolytic machinery in the cell.

Chapter 5 discusses properties of the dissociation and unfolding reaction of the Arc repressor dimer. The stability of Arc is concentration dependent and the unfolding reaction is well described as a two state transition from folded dimer to unfolded monomer. The stability of the protein is decreased at low pH and increased by high salt concentration. The salt dependence suggest that two ions bind preferentially to the folded protein. In 10 mM potassium phosphate [pH 7.3] and 100 mM KCl, the unfolding free energy reaches a maximum near room temperature. The results suggest that at the low protein concentrations where operator DNA binding is normally measured, Arc is predominantly monomeric and unfolded.

ACKNOWLEDGEMENTS

I feel fortunate to have been able to work with such an outstanding group of people in the Sauer lab and I am indebted to virtually every member of the lab for help or new insights at some point during my time here. A few people in particular had a significant impact on my progress in graduate school. I would like to thank Lynne Regan who provided my initial lab training in molecular biology and Drew Vershon who took me under his wing when I entered the Sauer lab. I learned the nitty gritty of doing experiments largely from them. Neil Clarke was obviously a major player in the pattern matching project, but beyond that, I have discussed nearly every phase of my work with him at some point and benefitted greatly from his argumentative nature. Finally, I will always be thankful to Bob Sauer who was a fantastic advisor in every way.

TABLE OF CONTENTS

CHAPTER 1

The Structural Message in Amino Acid Sequences: Analysis of Tolerance to Sequence Variation

The genome is manifest largely in the set of proteins which it encodes. It is the ability of these proteins to fold into unique three dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is of fundamental importance to our understanding of biological processes. The problem is enormously complex, however, and has eluded solution even though the first protein structures were determined over 30 years ago (1; 2; 3).

The structural message in a sequence is complicated, in part, because the importance of individual amino acids is quite variable . Some sequence positions are critical and are unable to tolerate even conservative amino acid changes. Others can be changed to vastly different amino acids with little or no effect. Clearly, decoding the structural message in an amino acid sequence will require identifying those amino acids that are important, but this is not currently possible by inspection of the sequence alone. One way to recognize important side chains is to study mutations that prevent folding or function (see ref. (4) for review). It is also possible to ask the inverse question, i.e., what kinds of sequence changes can be tolerated without altering the ability of the protein to fold and function? The latter experiment is the focus of this chapter.

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first relies on experiments which have been performed by nature for billions of years. The process of evolution generates amino acid substitutions that are either accepted or rejected by natural selection. If a particular property of a side chain such as charge or

size is important at a given position, only side chains that possess the required property will be passed on to future generations. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be incorporated over the course of evolution. The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene, combined with genetic selections or screens to identify functional sequences. The end products of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding and function.

Comparisons of phylogenetically related sequences are limited to proteins that are members of large families, however. Moreover, natural proteins must have passed the rather stringent test of evolutionary fitness. As a consequence, sequence features of relatively minor importance may have been conserved throughout evolution. For example, Sherman and coworkers have found that non-conservative substitutions can actually be functionally tolerated at certain sequence positions that are invariant among the naturally occurring cytochrome sequences (5). Moreover, evolutionarily related proteins often differ at many sequence positions, making it difficult to determine if a substitution is allowed only because of a compensating change elsewhere in the protein. Some of these limitations are overcome by the genetic approach. For example, it is possible to explore the effects of all sequence changes at one or a few positions, and to set the level of required function. In addition, phenotypic differences that do occur can be readily assigned to particular sequence changes. However, genetic methods require the availability of cloned genes and reasonably efficient screens or selections for folding and activity. Moreover, tolerance to amino

acid substitutions at only a few positions can be tested in any given experiment. While the strengths and weaknesses of the phylogenetic and mutagenic approaches are different, the striking finding is that both types of studies suggest that considerable information is contained in a very simple representation of an amino acid sequence.

Tolerance to amino acid substitutions: Proteins can be surprisingly tolerant of amino acid substitutions. For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in the Lac repressor of *E.coli* , Miller and his coworkers found that about half of all substitutions were phenotypically silent (6; 7). At some positions, all substitutions were active. At others, no substitutions were allowed. Why can some residues be changed dramatically and not others? Certainly, residues that are directly involved in protein functions are among the most conserved. For example, many of the invariant residues in Lac repressor are directly involved in binding operator DNA or inducer. However, many of the important residues are essential for defining the scaffold that allows the functionally important side chains to be precisely placed in three dimensions. What sequence features are essential to forming this scaffold?

In their initial comparisons of the globin sequences, Perutz and coworkers found that residues at the protein surface were more tolerant to substitutions than their buried counterparts (8). Most buried residues were found to require non-polar side chains, whereas few features of surface side chains were found to be generally conserved. Similar results have now been seen for a number of protein families (8; 9; 10; 11; 12; 13; 23). A clear example of the difference in variability allowed at surface and

buried sites can be seen in recent cassette mutagenesis experiments with λ repressor (14). The substitutions allowed in a short region of this protein are shown in Figure 1. Below the list of allowed substitutions is a histogram of how buried the side-chain is in the crystal structure. At six positions, only the wild type residue or a single, relatively conservative substitutions are allowed. Five of these positions are buried in the protein. By contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. The significant role of buried residues in determining protein structure has been confirmed in studies of mutations that destabilize proteins (15; 16). For example, Alber et al.(16) have isolated temperature sensitive mutants at 20 positions in T4 lysozyme and have shown that these destabilizing mutations occur almost exclusively at positions of low side-chain solvent accessibility (Fig. 2).

What is important about the core? The tolerance of core residues to amino acid substitutions has been extensively explored in λ repressor (17). Over 100 acceptable core sequences were identified by mutagenizing packing units containing three or four residues at a time. The substitutions allowed at each of the core residues in λ repressor in isolation and in combination are summarized in Figure 3. As seen in other studies, only hydrophobic or moderately polar residues are found to be tolerated at these buried sites so that the hydrophobic nature of the core is maintained. The conservation of hydrophobic side chains in the core of soluble proteins is undoubtedly due to the large favorable contribution of the hydrophobic effect to protein stability (18; 19; 20). Hydrophobic side chains in the unfolded state have the entropically unfavorable effect of ordering water at their surface. Thus the

process of water release upon folding is entropically favorable and provides a large stabilizing free energy. In contrast, polar and charged side chains can form favorable hydrogen bonds to water in the unfolded state which need to be replaced if the side chain becomes buried. In fact, the hydrogen bonding potential of essentially all polar groups in protein structures are satisfied by interactions with other protein moieties or water (21)[1].

The cores of proteins are quite closely packed (22), but some volume changes are clearly acceptable. In λ repressor, all of the acceptable sequences fall into a range of 2 additional or 3 fewer methylene groups, which corresponds to a range of about 10% of the total core volume. This may be taken then as the approximate range allowed for the entire λ repressor core. Volume changes allowed at individual sites within the core can be even larger, however. For example, at positions 47 and 65 in λ repressor, residues as small as Ala and as large as Phe are acceptable in the proper sequence contexts. Similarly large volume changes at individual buried sites have been seen in phylogenetic studies as well and the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (11; 12; 23). These local

---

[1] It is certainly possible to imagine a protein structure that is stabilized by a core of residues forming hydrogen bonds and salt bridges. Indeed, sulfate is bound in a hydrogen bonding network in the core of sulfate binding protein (52). However, polar cores of this type are probably rare because hydrogen bonds require an exact geometry for each interacting group, and consequently there may be a limited repertoire of possible structures. In contrast, a hydrophobic side chain's contribution to stability is less dependent on the identity of its neighboring side chains. Lesk and Chothia (11) point out that a hydrophobic core would, for these reasons, be much less sensitive to mutation than one consisting of a hydrogen bonded network. Since protein structures must be resilient to change, the process of evolution will tend to favor the burial of hydrophobic residues and not polar residues.

volume changes can be accommodated not only by conformational changes in nearby side chains, but by a variety of backbone movements.

What are the constraints on residues allowed in the core? Clearly, with rare exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the the core is comprised of side chains that can assume only a limited number of conformations (24), it is also important that packing be maintained without steric clashes. How important are composition, volume, and the steric compatibility in determining whether a given sequence can form an acceptable core? Some estimates are possible for the λ repressor core experiment. In one experiment , three core positions were mutagenized simultaneously giving 8000 (20^3) possible residue combinations. Of these, 73% fall within the volume range + 2 to - 3 methylene groups compared to the wild-type core sequence. Thus, most random sequences have an appropriate volume and volume is not an informationally rich characteristic of an acceptable core sequence. In other words, identifying sets of residues that have the appropriate volume is unlikely to provide much information about possible core sequences, because most sets of residues will suffice. By contrast only 6% of the random sequences are composed of acceptable hydrophobic residues. Since the actual fraction of allowed core sequences is 1.4%, fully one fourth of all sequences that have the appropriate amino acids are sterically allowed in the core. Thus, residue hydrophobicity is an informationally significant feature of acceptable core sequences.

What is important about surface sites? Many conserved residues on the protein surface are essential for function. They may be required for ligand binding, catalysis, interactions with other macromolecules, etc. and in the absence of biochemical information, or crystal structures it is not possible to determine whether a conserved position is important structurally or functionally. As noted above, however, at the level of individual residues, very few surface sites seem to be essential to the formation of a stable protein structure. Many surface sites can tolerate a large number of chemically different residues, including hydrophobic residues, and surface positions are rarely sites of destabilizing mutations. Nevertheless, some surface residues do contribute to stability. Proline, is often found to be conserved at surface sites, particularly near the N-terminal end of helices, presumably because it uniquely restricts the allowed local backbone conformations (10; 25)[1]. Surface hydrogen bonds, salt bridges, and interactions with helical dipoles may also contribute modestly to overall stability (26; 27). Bashford, Lesk, and Chothia (10) in an extensive analysis of globin sequences have found a strong bias against large hydrophobic residues at many surface positions. At some level, it must be important that the overall protein surface be reasonably hydrophilic. Large patches of hydrophobic surface residues would presumably lead to aggregation and insolubility. On the other hand, these sites may play an important role in protein folding that is not yet understood.[2]

---

[1] Glycine is also unique because it is more flexible than any other amino acid. While glycine is also often found to be conserved, it appears to be rarely conserved for this property. Instead, it is usually found conserved at buried sites where side chains would not fit (see ref. 53 for example).

[2] One possibility is that residues at certain surface sites may be less solvent accessible in the denatured state relative to the native state so that hydrophobic residues at these positions

The structural message in a protein sequence: As shown by Anfinsen (28; 29), the sequence of amino acids contains all the information necessary to determine the structure of a protein. Thus, in theory, it should be possible to predict structure from sequence information. One problem, however, is that protein folding involves choices between enormously large numbers of different conformations whose free energies are determined by electrostatics, hydrogen bonding, solvent interactions, backbone and side chain dihedral torsional preferences, van der Waals interactions, etc. If all of these properties are essential aspects of the folding code, then one cannot be optimistic about prospects for a solution in the near future. It is possible, however, that many features of the sequence are subsidiary and that only a few properties of the sequence dominate the folding code. This is not necessarily meant in an energetic sense but in an informational sense. For example, one important feature of an amino acid sequence is the ability to form the pattern of backbone hydrogen bonds found in an α-helix. While these hydrogen bonds contribute in some way to the stability of the folded protein, this feature is unlikely to provide much information about the folded structure since all amino acids except proline can form these hydrogen bonds.

What do studies of neutral amino acid substitutions suggest are the most informationally rich features of an amino acid sequence? The principal finding is that most surface sites can tolerate diverse

---

would preferentially stabilize the unfolded protein. Mutants likely to act by this mechanism have indeed been isolated (54).

substitutions, and are probably important chiefly in maintaining a reasonably polar surface. This does not mean that surface positions contribute nothing to stability, but that the information is highly degenerate. The information contained in buried residues is also degenerate, with the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, a key message in an amino acid sequence may simply be its specific pattern of hydrophobic and hydrophilic residues. I wish to restate emphatically that I mean this in an informational sense. Clearly, the precise structure and stability of a protein cannot be defined without considering in detail every interaction in the structure. It is possible, however, that structural prediction at a primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence.

Implications for the prediction of structure: From aligned sets of sequences it is usually possible to infer which residues are buried and which are solvent exposed. Buried positions will tend to accept only hydrophobic or neutral side chains, whereas solvent exposed positions will almost always contain highly polar residues in at least some of the aligned sequences. As a result, residue positions can initially be classified with regard to the maximal level of hydrophilicity allowed at that position. This information can then be used to search for patterns that are characteristic of the periodicity of α-helices and β-sheets (30; 31; 32; 33; 34). For example, if a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected. As initially shown by Perutz et al. (8), amphiphilic patterns expected for simple secondary structures can be much clearer in a

set of related sequences. This principle is illustrated in Figure 4 which shows helical hydrophobic moment plots for the Anntenapedia homeodomain sequence alone and for a combined set of 30 homeodomain proteins . The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (32; 33; 34). The amphipathic character of the helical regions is clearly revealed by the combined set. In cases where aligned sequence data is not available from phylogenetically related proteins, it can be rapidly generated by cassette mutagenesis and analysis of neutral amino acid substitutions. In this manner, the secondary structure of Arc repressor, a small DNA binding protein from bacteriophage P22 was recently correctly predicted (35).

It is important to note that elements of secondary structure, depending upon their position in the protein structure, need not show simple amphipathic character. These hydrophobic patterns, along with expected steric restrictions were compiled by Lim as a set of rules for secondary structure prediction (30; 36). This method is comparable in accuracy to the propensity based methods (37) and it may be interesting to see what improvements are achievable by incorporating information from sets of sequences or neutral amino acid substitutions in a similar manner to that described above for the hydrophobic moment calculations[1].

The prediction of tertiary structure from secondary structure is far from trivial (38; 39). Indeed, the stabilization of secondary structure may be

[1] Only marginal improvements have been obtained by using combined predictions with propensity based methods (55).

largely a consequence of long range interactions. Consequently, it may be useful to attempt to bypass the secondary structure prediction problem entirely. Currently, the only convincing method for prediction of tertiary structure is by identifying homology to a protein that already has a known structure (40; 41; 42). This procedure becomes increasingly difficult as the level of homology decreases due both to structural divergence and difficulties in alignment. Moreover, it is often difficult to detect sequence similarity between distantly related proteins that adopt the same structure. Since many more sequences are known than three-dimensional structures, it would be advantageous to increase the reach of the structural information that is already in hand by improving methods for detecting distant sequence relationships and for subsequently aligning these sequences based on structural principles. If a protein of known structure is a member of a family of related sequences, information about allowed substitution patterns can be incorporated into sequence homology searches to help detect more distantly related proteins. In a normal homology search, the sequence database is scanned with a single test sequence. In this case, every residue in the sequence must be weighted equally. As discussed above, however, some residues are more important than others and should be weighted accordingly. Moreover, certain regions are more likely to contain gaps than others. This information is available from sequence sets and several methods have been used to combine the information into a more appropriately weighted sequence search and subsequent sequence alignment (see 43 for review).

How far can a simple consideration of sequence hydrophobicity take us toward the goal of tertiary structure prediction? There is certainly a

great deal of information in the basic pattern of hydrophobic residues and it is possible that a particular pattern of hydrophobic residues may be a necessary (although not sufficient) feature of a sequence that adopts a specific fold. If this is true, then the pattern of hydrophobic and hydrophilic residues should be a characteristic feature of that fold. Sweet and Eisenberg (44) have shown that given two protein sequences, a good criterion for their structural relatedness is the correlation of the pattern of hydrophobicity in their sequence. This suggests that a particular pattern is indeed specific to a individual fold. This is also indicated by work of Bashford, Lesk, and Chothia (10) with globin sequences and Pearl and Taylor (42) with aspartic protease sequences. Both groups identified conserved regions in aligned sequences and codified the allowed variability in a set of characteristic patterns referred to as templates. In these templates, most positions were classified by the preferred hydrophobicity, with only a few positions requiring a specific amino acid. For example, out of 77 positions in the conserved regions of the aspartate proteases, only four specified a particular amino acid. When these templates were used to search all protein sequences in the database, the globin template preferentially identified globin sequences, and the aspartic protease template identified only acid protease sequences. These studies show that very simple patterns are selective for sequences that adopt a given fold. It has recently been shown that patterns of hydrophobicity in sets of aligned sequences correlates sufficiently well with patterns of solvent accessibility in known protein structures that hydrophobic patterns can be used to selectively identify correct protein folds in the database of known protein structures (45). This method does not rely on sequence homology and

consequently it may be useful for identifying homologous protein folds in cases where traditional sequence homology is statistically insignificant.

While these studies indicate that the pattern of hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure *de novo*. It is important, first of all, that we understand how the hydrophobic pattern in sequence space can be related to conformation space. Dill and coworkers have begun to approach this problem (46). They have begun studying the properties of sequences composed entirely of hydrophobic groups and polar groups on two dimensional lattices. An example of such a representation is shown in Figure 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice and two residues cannot occupy the same space. Free energies of particular conformations are evaluated using a single term, an attraction of H groups. By considering chains of 10 residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence and/or conformation space could be explored. The significant results were: (i) Not all sequences can fold into a compact or "native" structure and only a few sequences form a unique native structure; (ii) The probability that a sequence will adopt a unique native structure increases with chain length; and (iii) The native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. While the gap between these two dimensional simulations and real three dimensional structures is large, it is encouraging that unique native state can apparently be attained by very simple rules and sequence representations.

Jernigan and Covell (47) have done full conformational searches for individual protein sequences on three dimensional cubic lattices that are limited in size and shape according to the known dimensions of the protein. The free energies of the thousands of possible conformations are evaluated using a scale that is similar in many respects to a simple hydrophobicity scale. They find that the correct structure is always among the 4% best conformations. While it is likely that the method will work less well without the size and shape constraint, it would appear that very simple representations can effectively limit the amount of conformational space that would need to be searched by more sophisticated methods.

Conclusion: There is clearly more information in sets of sequences than in a single sequence alone and a number of practical applications arise from an analysis of neutral amino acid substitutions. First of all, the information permits the evaluation of a residue's importance to the function and stability of a protein and the ability to identify the essential elements of a protein sequence, may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, incorporating a knowledge of allowed substitutions can improve our ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring. Third, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy surface positions. The characteristic patterns that emerge can be used to identify likely regions of secondary structure.

It is important to point out that as more and more sequences are being determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. Moreover, even if this is not the case, it is now often possible to obtain similar information using mutagenic techniques. Consequently, at least in the short term, it may not be necessary to solve the folding code for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by using an analysis of neutral amino acid substitutions to identify key residues and thereby simplifying sequence space, combined with simplifications of conformation space as in the lattice methods, something approaching true structure prediction could be accomplished.

ACKNOWLEDGEMENTS

I would like to thank Wendell Lim and John Reidhaar-Olson for contributing data and ideas to this chapter.

REFERENCES

1. Perutz, M. F., Rossman, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960) *Nature* **185,** 416-22.

2. Scouloudi, H. (1959) *Nature* **183,** 374-376.

3. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wycoff, H. & Phillips, D. C. (1958) *Nature* **181,** 662-6.

4. Shortle, D. (1989) *J. Biol. Chem.* **264,** 5315-5318.

5. Hampsey, M. D., Das, G. & Sherman, F. (1988) *FEBS Lett.* **231,** 275-283.

6. Kleina, L. G. & Miller, J. H. (1989) *J. Mol. Biol.* submitted.

7. Miller, J. H., Coulondre, C., Hofer, M., Schmeissner, U., Sommer, H. & Schmitz, A. (1979) *J. Mol. Biol.* **131,** 191-222.

8. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669-678.

9. Dickerson, R. E. (1980) *Sci. Amer.* **242,** 136-153.

10. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199-216.

11. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225-270.

12. Lesk, A. M. & Chothia, C. (1982) *J. Mol. Biol.* **160,** 325-342.

13. Taylor, W. R. (1986) *J. Mol. Biol.* **188,** 233-258.

14. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241,** 53-57.

15. Pakula, A. A., Young, V. B. & Sauer, R. T. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8829-33.

16. Alber, T., Sun, D., Nye, J. A., Muchmore, D. C. & Matthews, B. W. (1987) *Biochemistry* **26,** 3754-58.

17. Lim, W. A. & Sauer, R. T. (1989) *Nature* in press.

18. Privalov, P. L., Griko, Y. V., Venyaminov, S. Y. & Kutyshenko, V. P. (1986) *J. Mol. Biol.* **190,** 487-498.

19. Kauzmann, W. (1959) *Adv. Protein Chem.* **14,** 1-63.

20. Baldwin, R. L. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8069-8072.

21. Baker, E. N. & Hubbard, R. E. (1984) *Prog. Biophys. Mol. Biol.* **44,** 97-179.

22. Richards, F. M. (1974) *J. Mol. Biol.* **82,** 1-14.

23. Chothia, C. & Lesk, A. M. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 399-405.

24. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775-791.

25. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648-1652.

26. Perry, K. M., Onuffer, J. J., Touchette, N. A., Herndon, C. S., Gittelman, M. S., Matthews, C. R., Chen, J. T., Mayer, R. J., Taira, K. & Benkovic, S. J. (1987) *Biochemistry* **26,** 2674-2682.

27. Alber, T., Sun, D., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987) *Nature* **330,** 41-46.

28. Anfinsen, C. B. (1973) *Science* **181,** 223.

29. Epstein, C. J., Goldberger, R. F. & Anfinsen, C. B. (1963) *Cold Spring Harbor Symp. Quant. Biol.* **28,** 439.

30. Lim, V. I. (1974) *J. Mol. Biol.* **88,** 857-872.

31. Schiffer, M. & Edmundson, A. B. (1967) *Biophys J.* **7,** 121-135.

32. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982) *Nature* **229,** 371-374.

33. Eisenberg, D., Schwarz, D., Komaromy, M. & Wall, R. (1984) *J. Mol. Biol.* **179,** 125-142.

34. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 140-144.

35. Bowie, J. U. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 2152-2156.

36. Lim, V. I. (1974) *J. Mol. Biol.* **88,** 873-894.

37. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152-156.

38. Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979) *J. Mol. Biol.* **132,** 275.

39. Hurle, M. R., Matthews, C. R., Cohen, F. E., Kuntz, I. D., Toumadje, A. & Johnson, W. C. J. (1987) *Proteins: Structure Function and Genetics* **2,** .

40. Brown, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969) *J. Mol. Biol.* **42,** 65-86.

41. Greer, J. (1981) *J. Mol. Biol.* **153,** 1027-1042.

42. Pearl, L. H. & Taylor, W. R. (1987) *Nature* **329,** 351-354.

43. Taylor, W. R. (1988) *Prot. Eng.* **2,** 77-86.

44. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171,** 479-488.

45. Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1989) *Proteins: Structure, Function and Genetics* submitted.

46. Lau, K. F. & Dill, K. A. (1989) *Macromolecules* in press.

47. Jernigan, R. L. & Covell, D. G. (1989) in preparation.

48. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55,** 379-400.

49. Burglin, T. R. (1988) *Cell* **53,** 339-40.

50. Otting, G., Qian, Y., Muller, M., Affolter, M., Gehring, W. & Wuthrich, K. (1988) *EMBO Journal* **7,** 4305-4309.

51. Fauchere, J. -. & Pliska, V. (1983) *Eur. J. med. Chem.-Chim. ther.* **18,** 369-

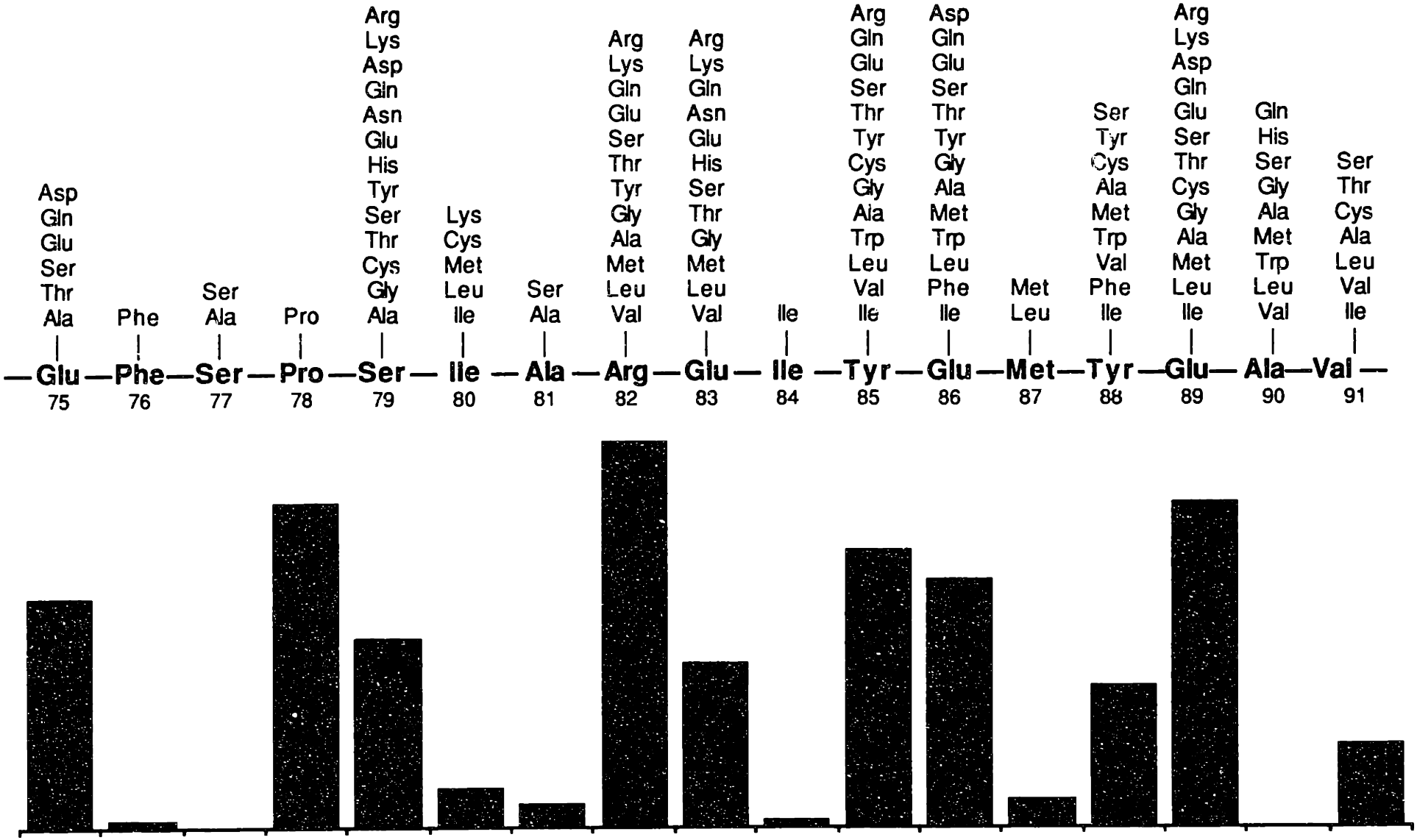
52. Quiocho, F. A., Sack, J. S. & Vyas, N. K. (1987) *Nature* **329,** 561-564.

53. Takano, R. & Dickerson, R. E. (1981) *J. Mol. Biol.* **153,** 79-94.

54. Pakula, A. A. & Sauer, R. T. (1989) submitted.

55. Zvelebil, M. J., Barton, G., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **195,** 957-961.

Figure 1. Amino acid substutions allowed in residues 75-91 of λ repressor (upper) and the fractional solvent accessibility of the wild type side chain in the crystal structure (lower). The wild type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutagenizing one to three codons at a time using a cassette method and applying a functional selection (14). The histogram represents the fractional solvent exposure of all the atoms in the wild type side chain relative to the same atoms in an Ala-X-Ala model tripeptide (48). The fractional accessibility of the most exposed residue, Arg82, is 0.86.


Asp Gln Glu Ser Thr Ala
Phe
Ser Ala
Pro
Arg Lys Asp Gln Asn Glu His Tyr Ser Thr Cys Gly Ala
Lys Cys Met Leu Ile
Ser Ala
Arg Lys Gln Glu Ser Thr Tyr Gly Ala Met Leu Val
Arg Lys Gln Asn Glu His Ser Thr Gly Met Leu Val
Ile
Arg Gln Glu Ser Thr Tyr Cys Gly Ala Trp Leu Val Ile
Asp Gln Glu Ser Thr Tyr Gly Ala Met Trp Leu Phe Ile
Met Leu
Ser Tyr Cys Ala Met Trp Val Phe Ile
Arg Lys Asp Gln Glu Ser Thr Cys Gly Ala Met Leu Ile
Gln His Ser Gly Ala Met Trp Leu Val
Ser Thr Cys Ala Leu Val Ile
Glu Phe Ser Pro Ser Ile Ala Arg Glu Ile Tyr Glu Met Tyr Glu Ala Val
75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91

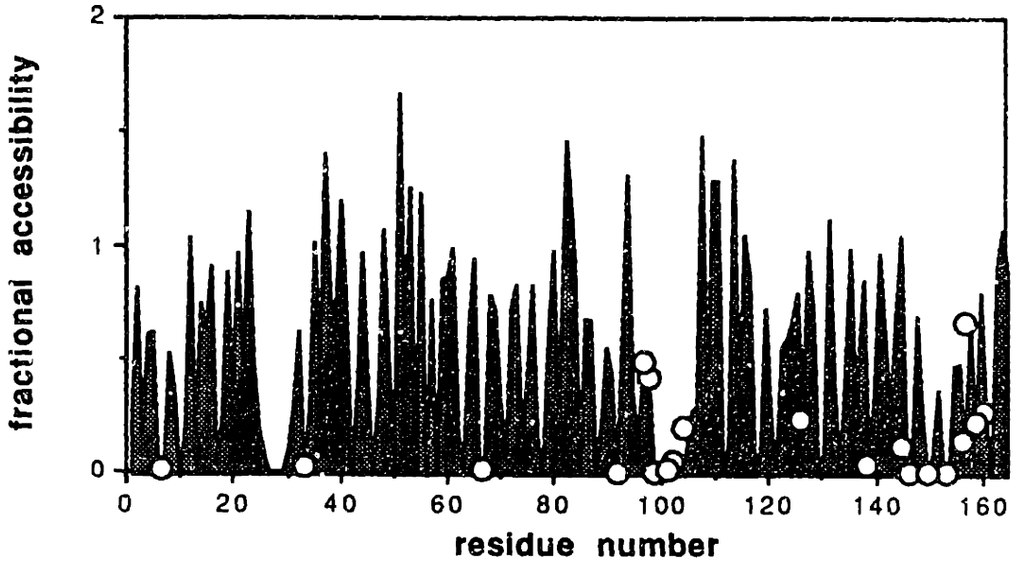
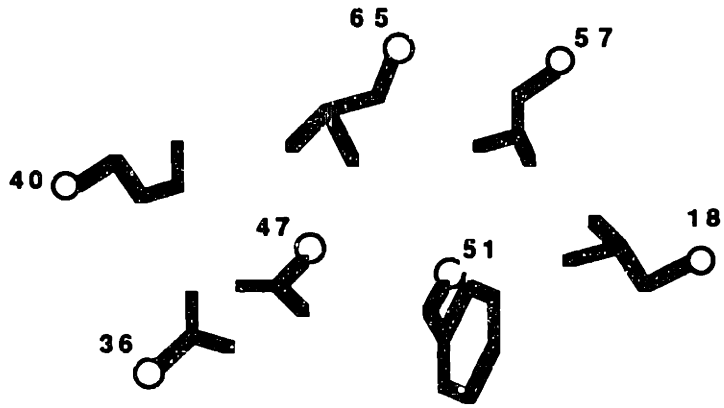


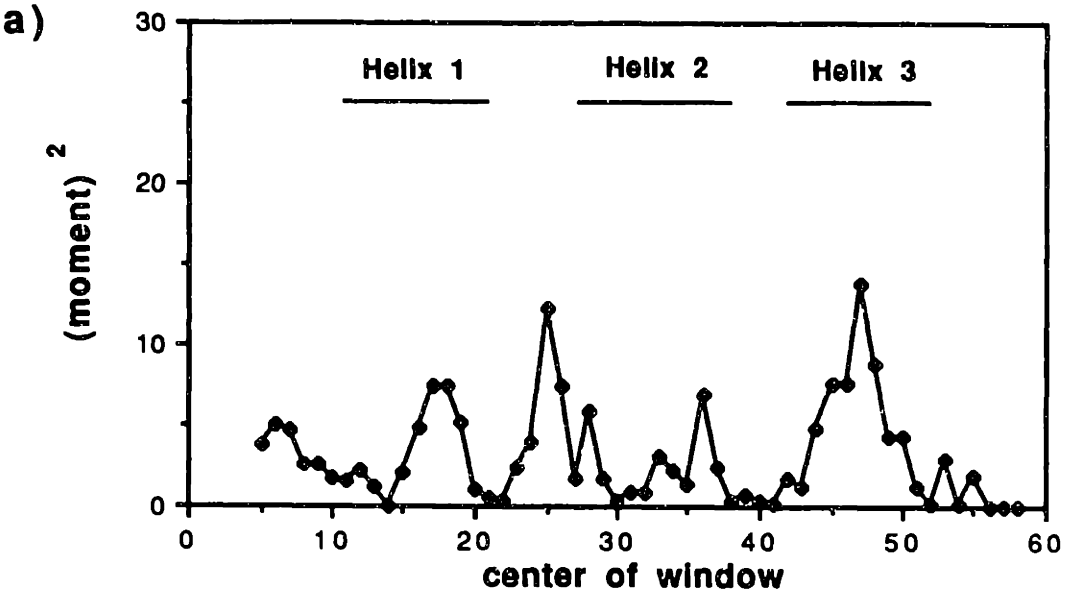
Figure 2. The solvent exposure of side chains in T4-lysozyme where destabilizing mutations were isolated (16). The sites of destabilizing mutations are indicated by the white dots. Solvent accessibility was determined using the Lee and Richards algorithm (48). Fractional solvent accessibility was calculated as the sum of the solvent accessibility of each side chain atom and the C_α backbone atom, divided by the same atoms in an Ala-X-Ala model tripeptide.

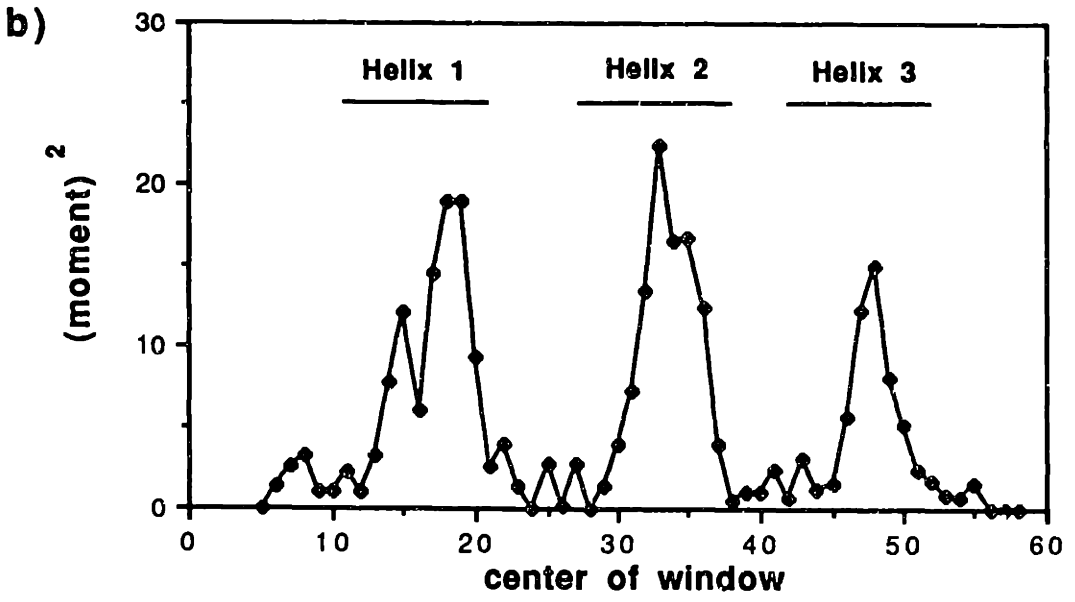


40	36	47	65	51	57	18
Ala	Ala	Thr	Ile	Val	Pro	Ala
Cys	Cys	Val	Leu	Ile	Val	Cys
Val	Thr	Ile	Phe	Leu	Ile	Val
Ile	Val	Leu		Met	Leu	Ile
Leu	Ile	Met		Phe	Met	Leu
Met	Leu	Phe			Phe	Met
	Met					

Figure 3. Amino acid substitutions allowed in the core of λ repressor. The wild type side chains are shown pictorially in the approximate orientation seen in the crystal structure. The lists of allowed substitutions at each position are shown below the wild type side chains. These substitutions were identified by randomly mutagenizing one to four residues at a time using a cassette method and applying a functional selection (17). Not all substitutions are functional in every sequence background.

Figure 4. Helical hydrophobic moments calculated using (a) the single Antennapedia homeodomain sequence or (b) a set of 39 aligned homeodomain sequences (49). The bars indicate the extent of the helical regions identified in NMR studies of the Antennepedia homeodomain (50). To determine hydrophobic moments, each position was classified according to its hydrophobicity and placed in one of three groups, H_1 (high hydrophobicity= Trp, Ile, Phe, Leu, Met, Val or Cys), H_2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser) and H_3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys or Arg). The vector magnitudes were assigned a value of 1, 0, or -1 for positions where the hydrophobicity group was H_1, H_2 or H_3 respectively. The hydrophobicity group for positions in the single sequence could be determined in a straightforward manner. For multiple sequences, the residues seen at each position were sorted by their hydrophobicity using the scale of Fauchere and Pliska (51). Arg and Lys were not counted unless no other residue was found at the position since they contain long aliphatic side chains and can thereby substitute for non-polar residues at buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was thrown out unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An 8 residue window was used and the vectors projected radially every 100°.


a)
Helix 1
Helix 2
Helix 3
(moment)^2
0
10
20
30
0
10
20
30
40
50
60
center of window


b)
Helix 1
Helix 2
Helix 3
(moment)^2
0
10
20
30
0
10
20
30
40
50
60
center of window

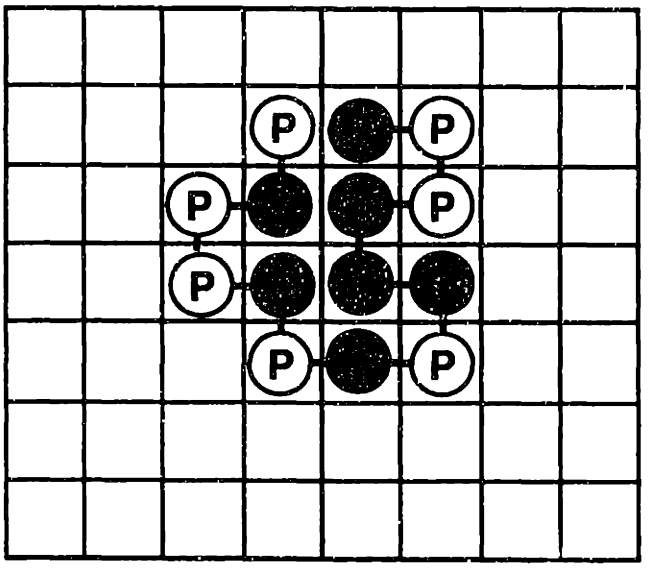


Figure 5. A representation of one compact conformation for a particular sequence of H and P residues on a two dimensional square lattice.

CHAPTER 2

Identifying Determinants of Folding and Activity for the Bacteriophage P22 Arc Repressor

INTRODUCTION

With the advent of recombinant DNA techniques for modifying protein sequences, considerable work has been directed towards 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 upon 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 (4-9). 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

Bacterial Strains: *Escherichia coli* strains used in this work were: UA2F, a derivative of strain US3 (*str*A, *thi*, *his*, *lac*Z, *sup*^{0}, *rec*A) containing an F' episome (*lac*I^Q, *lac*Z::Tn5[*kan*^{R}] *pro*$^+$) and a λAC201 prophage which bears a P_{ant} promoter fusion to *cat* and X90 (*ara*, Δ*lac-pro*, *nal*A, *arg*Eam, *rif* ^{R}, *thi*1; F' *lac*$^+$, *lac*I^{Q1}, *pro*$^+$) (7).

Buffers and Media: The following buffers were used: PCB: 50 mM Tris-HCl [pH 8.0], 0.1 mM EDTA, 5% glycerol, 1.4 mM 2-mercaptoethanol; TE: 10 mM Tris-HCl [pH 8.0], 1 mM EDTA; 5X Klenow buffer: 50 mM Tris-HCl [pH 8.0], 250 mM NaCl, 50 mM $MgCl_2$, 25 mM DTT, 1.25 mM each deoxynucleoside triphosphate.

General Molecular Biology: Plasmids were constructed by established procedures (10) and single stranded plasmid DNA for sequencing was prepared as described (7). DNA sequencing was performed by the dideoxy method (11) using Sequenase reagents and procedures from United States Biochemicals. Oligonucleotides were synthesized on a Systec Microsyn 1450A or an Applied Biosystems 380 B and purified as described previously (12).

Plasmids: New restriction sites were introduced throughout the Arc gene sequence by the construction of a synthetic gene. A synthetic DNA cassette

containing the entire coding sequence, the translation start site, a 5'-*Sal* I overhang and a 3'-*Sty* I overhang, was created by ligating eleven overlapping oligonucleotides. This cassette was ligated into the *Sal* I-*Sty* I backbone fragment of a pTA200 derivative (7), whose *Eco*R I site was destroyed (the site was destroyed by digesting with *Eco*R I and filling in with Klenow enzyme), and a defective mutant Arc gene, PL8. The resulting plasmid, bearing a synthetic *arc* gene downstream from the P_{tac} promoter, was digested with *Pst* I, a site within the *bla* gene, and *Cla* I, a site at the end of the synthetic *arc* gene that was introduced by the synthetic cassette. The *arc* containing fragment was then ligated into the *Pst* I-*Cla* I backbone fragment of plasmid pFG650 (F. Gimble, personal communication) to produce pSA100. For Arc activity selections, a plasmid pSA200 was constructed by excising the synthetic gene from pSA100 on a *Pvu* I-*Eco*R V fragment and ligating this fragment into the *Pvu* I-*Eco*R V backbone of pUS406 which contains an str^s gene under the control of an Arc repressible promoter and an m13 origin of replication from plasmid pZ150 (14). The construction of pUS406 is described elsewhere (13). A derivative of pSA200, called pSA300, was also constructed in which the *Sma* I site in the synthetic *arc* coding sequence was replaced with a *Sac* II site by inserting an appropriate synthetic DNA cassette. The synthetic *arc* genes are active *in vivo* and are expressed at levels comparable to the wild-type gene. The sequence of the synthetic Arc gene in plasmid pSA300 is shown in figure 1.

Mutagenic Cassettes: Mutagenic cassettes spanning restriction sites in the synthetic gene were prepared by the method of Oliphant *et al.* (15). Self-priming, synthetic oligonucleotides were prepared in the usual manner except that certain regions were mutagenized by contaminating each base

during the synthesis with 7.5% of each of the other three nucleotides. To prepare the second strand of the cassette, 10 μg of the purified oligonucleotides in 40 μl TE were heated to 65 °C for 1 min and cooled to room temperature to allow annealing. 10 ul of 5x Klenow Buffer was then added, followed by 5 units of Klenow enzyme. Another 5 units of Klenow enzyme was added after 1 hr. After 2 hr, the the enzyme was removed by phenol extraction and the DNA was recovered by ethanol precipitation. Ligatable cassettes were then released by digesting 1 μg of the double stranded DNA with the appropriate restriction enzymes and the cassettes were purified by electrophoresis on 12% polyacrylamide gels. Bands were visualized by ethidium bromide staining, and DNA from the appropriate excised band was then eluted from the crushed gel slice by diffusion into 3 ml TE overnight. The DNA was finally purified by the Elutip procedure (Schleicher and Schuell) and ethanol precipitated.

Cassettes were constructed to mutagenize six regions. All mutagenesis experiments except region 3 were performed in the pSA200 background. Region 1 (codons 2 through 8) and region 2 (codons 9 through 16) were mutagenized using cassettes spanning *Xho* I to *Sma* I. Since *Sma* I cuts to leave a blunt end and the fill-in step also leaves a blunt end, it was possible to mutagenize through the *Sma* I site. However, because of the location of the *Sma* I site, only the first position of codon 16 could be mutagenized in this manner. As the *Sma* I site is not unique in the pSA200 background, technical difficulties resulted in a significant background from wild-type plasmid in these experiments. Consequently, to mutagenize region 3 (codons 17 through 23), we constructed pSA300, in which the *Sma* I site in pSA200 was replaced with a *Sac* II site. Region 3 was then

mutagenized using a *Sac* II-*Bgl* II cassette in the pSA300 background. Region 4 (codons 24 through 32), region 5 (codons 33 through 42), and region 6 (codons 43 through 53) were mutagenized using *Xba* I-*Eco*R I, *Bgl* II-*Hin*d III and *Mlu* I-*Cla* I cassettes, respectively. The proportion of mutagenized genes which produced active protein was about 8% for region 3 and 1-2% for the other regions. The proportion of all genes which produced proteolytically resistant protein was about 30%, 9%, 13%, 1%, 2%, and 6% for regions one through six respectively.

Arc Protein Level Screen and Activity Selection: To screen clones for high levels of inducible protein, cells were grown in 96 well microtiter plates. LB medium (200 ul, containing 100 ug/ml ampicillin) was placed in each well and inoculated with 5 µl of an overnight culture. Cultures were incubated at 37 °C for 3 hrs and Arc expression from the P_{tac} promoter was induced by the addition of 5 µl of 100 mg/ml IPTG. Following an additional 3 hr incubation at 37 °C, cells were harvested by centrifugation in the microtiter plates, resuspended in 50 µl Laemmli sample buffer, lysed by heating at 90 °C for 2 min, and 25 µl portions were electrophoresed on 15% Laemmli gels (16). We do not know precisely how stable a protein must be to be protected from intracellular proteolysis, but the VG18 mutant protein, which is 50% denatured at about 30 °C under the conditions described below, is degraded rapidly in the cell and consequently is not present at a sufficiently high steady-state level to pass the screen. As discussed in Results, the LQ19 mutant protein, which is half denatured at 35 °C, is sufficiently stable to pass the screen. Thus, the stability of the LQ19 protein is probably close to the lower limit required for resistance to intracellular proteolysis.

The selection for Arc activity takes advantage of the fact that the streptomycin sensitive gene, *str*^{s} is dominant to the resistant gene, *str*^{r}. Consequently, in the presence of both alleles, a strain is sensitive to streptomycin. We have brought the *str*^{s} allele under the control of an Arc repressible promoter so that in the presence of active Arc, the sensitive gene is repressed and the cells are resistant to streptomycin. The selection strain, UA2F, also contains a P_{ant} promoter fusion to *cat* so that Arc activity can also be detected by screening for chloramphenicol sensitivity (7). Details of the selection procedure will be presented elsewhere (13). A mutant Arc protein that does not pass the selection, MI4, is stably folded and possesses between 2 and 10% of the wild type operator binding activity (7). Hence, mutants that pass the selection are presumed to have activities greater than this level.

Protein Purification: Mutant Arc proteins expressed from derivatives of pSA200 or pSA300 were purified from *E. coli* strains X90 or UA2F. Extracts from 1 liter of induced culture were prepared and carried through to the ammonium sulfate precipitation step as described (7,13). Precipitated protein was redissolved in 4 ml PCB and dialyzed for 4 hrs against the same buffer. This material was filtered through a 0.45 μm Millex HA filter and applied to a Accell CM Sep-pak cartridge (Waters) by means of a syringe. The cartridge was then washed with 8 ml of PCB plus 50 mM KCl and bound protein eluted step-wise with 4 ml portions of PCB plus 100, 150, 200, 250, 300, and 400 mM KCl. Mutant Arc proteins were found to elute at various salt concentrations depending on their net charge. Fractions containing Arc were concentrated to 0.5 ml by means of Centricon 10

concentrators (Schleicher and Schuell) and applied to a 1 x 25 cm column of Sephadex G-75 equilibrated in PCB plus 50 mM KCl. Arc containing fractions were pooled and stored at -20 °C.

Protein Stability and Spectroscopy: Denaturation experiments were performed as described previously (7, 13). Thermal denaturations of wild type or mutant Arc proteins were performed at protein concentrations of 8 μM in 10 mM potassium phosphate [pH 7.5] and 100 mM KCl. Guanidine hydrochloride denaturation was performed at a protein concentration of 16 μM in 10 mM Tris [pH 7.5] and 50 mM KCl. The unfolding reaction can be modeled as a two-state transition from folded dimer (A_2) to unfolded monomers (2U). As a consequence, the stability of Arc depends on protein concentration. Unfolding free energies in the absence of denaturant were determined by linear extrapolation of ΔG_u values calculated in the transition zone of the GuHCl denaturation curves. ΔG_u values for the unfolding reaction were calculated as $\Delta G_u = -RT \ln ([U]^2/[A_2])$.

Circular dichroism spectra of wild type or mutant Arc proteins at a concentration of 8 μM in 10 mM potassium phosphate [pH 7.5] and 100 mM KCl were recorded at 20 °C using an AVIV model 60DS spectropolarimeter.

Hydrophobic Moments: To incorporate data from allowed sequence substitutions into the hydrophobic moment calculation (17-19), vector magnitudes were defined as follows. Each position was categorized according to the hydrophobicity of the wild-type residue or the allowed substitutions using the hydrophobicity scale of Fauchere and Pliska (20).

Positions where residues more hydrophilic than Gly were structurally accommodated were assigned a value of -1 and the remaining positions were assigned a value of +1. Since Arg and Lys can play ambiguous roles, these residues were not included unless they were the only residue accommodated at the position. For the helical hydrophobic moments, an eight residue window was used and vectors were assumed to project radially every 100°. A six residue window was used for the β-strand moments and the vectors were assumed to project radially 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, seven to 11 adjacent codons were 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 sequenced. The top portion of Figure 2 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. For example, at position 47, the wild type residue is Lys, but Met, Ser, Asn, Thr, Arg, and Glu were also recovered among the active Arc sequences. Because many of the side chains in the C-terminal region of Arc can be changed dramatically without affecting activity, 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 which 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 (7,13). 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 using the same random sequence pools described above. The bottom portion of Figure 2 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 non-conservative 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 spectral properties. The fluorescence spectra of native and denatured Arc indicates that the unique tryptophan residue at position 14 is buried in a hydrophobic environment in the folded protein and becomes exposed upon unfolding. The circular dichroism spectrum of the folded Arc protein is characteristic of a predominantly α-helical protein. These spectroscopic methods are therefore useful probes of the Arc structure. Fig. 3 shows the circular dichroism 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 circular dichroism 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 are able to assume a folded conformation similar to that of the wild type protein.

To assay the thermodynamic stabilities of the structures of the variant proteins, we monitored their spectral properties as a function of temperature or GuHCl concentration. Table I shows that the variants have a range of stabilities. The least stable mutant is about 2 kcal/mole less stable than wild type, while the most stable variant is about 3 kcal/mole

more stable. 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 decreases per subunit are similar to those reported for T4 lysozyme mutants in which a single surface hydrogen bond is perturbed (21), while the increases are similar to those reported for mutant Cro proteins in which surface side chains are altered (22). However, the important point is that in all of these cases, the mutant proteins are still able to assume the basic three-dimensional structure of the wild type protein. Clearly then, the Arc side chains which 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.

Figure 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 the likely role of the residues 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 the Arc sequence, residue substitutions that dramatically alter the chemical 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 in the Arc sequence, 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? Two of the invariant residues, Pro15 and Gly30, may be involved in turns or special backbone conformations, since prolines and glycines often play these roles in proteins of known structure (23). Five of the invariant residues are charged (Arg23, Glu28, Arg31, Glu36, and Arg40). These side chains might participate in tertiary hydrogen bonds or ion pairs that stabilize the Arc structure. At 15 positions that are invariant or highly conserved (Phe10, Leu12, Trp14, Val18, Leu21, Val22, Val25, Ala26, Val33, Ile37, Tyr38, Val41, Met42, Phe45 and Lys46), hydrophobicity appears to be the important factor. At most of these positions, only hydrophobic side chains or Arg or Lys are found. Although Arg and Lys are not hydrophobic residues *per se*, they contain long aliphatic regions and

thus can substitute for hydrophobic residues under some circumstances. It has long been established that buried positions strongly prefer non-polar residues and tend to show a strong conservation of hydrophobicity among related sequences (1-3, 24-27). These 15 side chains are therefore likely to be buried in the Arc monomer or the dimer interface and form the hydrophobic core of Arc repressor. This is a reasonable number of buried side chains for a protein the size of the Arc dimer (26). 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, 28, 29).

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 (17-19, 30). For β-strands, this is a simple alternation of polar and non-polar residues, while for α-helices, it is a more complicated pattern that matches the 3.6 residue/turn periodicity of the helix. However, these characteristic patterns can be masked in natural sequences by hydrophobic residues that are solvent exposed. To look for these patterns in Arc repressor, we used the lists of structurally allowed substitutions to reveal positions that can accommodate hydrophilic residues. Figure 5 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 through 14) and two reasonably long regions of potential α-helix centered near residues 22 and 41. To further explore the potential α-helical regions, we chose the most hydrophilic residue that was allowed at each position and plotted these residues in helical wheel projection (30). Figure 6 shows that for residues 16-28 and 35-47, there is a striking clustering of hydrophobic residues on one side of the helical wheel and hydrophilic residues on the other side of the helical wheel. This amphipathic character is exactly that expected for an α-helix, with one side forming part of the hydrophobic core and the other side facing towards 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 (Zagorsky and 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 non-conservative substitutions when structure is demanded, but tolerate only conservative substitution when both structure and function is required. These residues may contact the DNA directly or may be involved in stabilizing interactions between subunits in the bound tetramer. At least 10 of these positions are in the N-terminal third of Arc repressor. In contrast, for the remainder of the protein, only Asp20, Asn34, Glu48 and Arg50 have the properties expected for 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 repressor. This interpretation is supported by an earlier biochemical analysis of Arc-defective mutants which showed that Arc proteins with sequence

substitutions at residues 2, 3, 4, 5, 8, and 10 had severely reduced operator affinity, but were stably folded (7). Moreover, a hybrid protein containing the nine N-terminal residues of Arc fused to the C-terminal region of Mnt, a related repressor, has the operator binding specificity of Arc (31).

Strengths and Weaknesses of the Method: Experiments of the type described here can be performed quite rapidly. For example, about three months were 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 employed and in these experiments, single base changes within a given codon are about four-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 non-conservative substitution in a more exhaustive study. However, for most codons, the single and double changes do encode a reasonable spectrum of conservative and non-conservative residue changes so that if non-conservative 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. Confidence in the overall analysis, however, is bolstered by several lines of confirmatory evidence: (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 (7, 31). (ii) The secondary structures predicted from the patterns of likely buried residues identified in the study are now known from NMR studies to be essentially correct (Zagorsky and Patel, personal communication). (iii) We have suggested that 25 residues of Arc repressor are structurally important because these positions accommodate only the wild-type side chain or conservative substitutions in the set of protease resistant variants. This inference is strongly supported by the finding that defective mutations, which result in proteolytic instability, have been

identified at 21 of these positions (7). Moreover, 19 of the 25 stucturally important positions in Arc are occupied by the same residue or conservative substitutions in the homologous repressor, Mnt . Thus, we believe that the overall analysis of allowed residue substitutions provides an excellent overview of Arc structure-function relationships.

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 and his colleagues have shown that extensive maps of defective and neutral substitutions can be generated by the suppression of amber mutations (32).

REFERENCES

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. Youderian, P., Chadwick, S. J. and Susskind, M. M. (1982) *J. Mol. Biol.*, **154**, 449-464

6. Vershon, A. K., Youderian, P., Susskind, M. M., & Sauer, R.T. (1985) *J. Biol. Chem.*, **260**, 12124-12129

7. Vershon, A. K., Bowie, J. U., Karplus, T. M., & Sauer, R. T. (1986) *Proteins: Structure Function and Genetics*, **1**, 302-311

8. Vershon, A. K., Liao, S., McClure, W. R., & Sauer, R.T. (1987) *J. Mol. Biol.*, **195**, 323-331

9. Brown, B., Bowie, J. U., & Sauer, R.T., in preparation

10. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

11. Sanger, F., Nicklen, S., and Coulson, A. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467

12. Knight, K. L., & Sauer, R. T. (1988), *Biochemistry*, **27**, 2088-2097

13. Bowie, J. U. & Sauer, R. T. (1989) *J. Biol. Chem.*, **264**, 7596-7602

14. Zagursky, R., and Berman, M. (1984) *Gene*, **27**, 183-191

15. Oliphant, A. R., Nussbaum, A. L. & Struhl, K. (1986) *Gene*, **44**, 177-183

16. Laemmli, U.K. (1970) *Nature*, **227**, 680-85

17. Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1982) *Nature*, **299**, 371-374

18. Eisenberg, D., Schwarz, E. Komaromy, M. & Wall, R., (1984) *J. Mol. Biol.*, **179**, 125-142

19. Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 140-144

20. Fauchere, J.-L., & Pliska, V. (1983) *Eur. J. med. Chem.-Chim. ther.*, **18**, 369-375

21. Alber, T., Sun, D.P., Wilson, K., Wozniak, J.A., Cook, S.P., & Matthews, B.W. (1987), *Nature*, **330**, 41-46

22. Pakula, A. & Sauer, R.T. (1987) *Proteins: Structure Function and Genetics*, in press

23. Richardson, J.S. (1981) *Advances in Protein Chemistry*, **34**, 167-337

24. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.*, **171**, 479-488

25. Rose G. D., Geselowitz, A. R., Lesser, G. J., Lee R. H., Zehfus, M. H. (1985) *Science*, **229**, 834-838

26. Miller, S., Janin, J. Lesk A. M., Chothia, C. (1987) *J. Mol. Biol.*, **196**, 641-656

27. Lesk, A. M., & Chothia, C. (1980) *J. Mol. Biol.*, **136**, 225-270

28. Alber, T., Sun, D. P. , Nye, J. A., Muchmore, D. C., & Matthews, B. W. (1987) *Biochemistry*, **26**, 3754-3758

29. Hecht, M. H., Sturtevant, J. M., & Sauer, R. T. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 5685-5689

30. Schiffer, M. & Edmundson, A., B. (1967) *Biophys. J.*, **7**, 121-135

31. Knight, K. L., & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 797-801

32. Miller, J.H. (1979), *J. Mol. Biol.*, **131**, 249-258

PROTEIN	T_m (°C)	$[GuHCl]_{1/2}$ (M)	$\Delta\Delta G_u$ (kcal/mole)
KT2/MR4/SR5/KE6/PT8	67	2.0	2.8 ± 0.5
NS34/SC35/QE39	63	*nd*	*nd*
KT2/GA3/SN5/MT7/PQ8	60	1.5	2.9 ± 1.1
QR9/FL10/NI11/RK13/RG16	54	1.3	0.4 ± 0.3
Wild Type	54	1.2	0
VM18/LV19/DR20	46	*nd*	*nd*
KT24	44	1.0	-1.2 ± 0.3
KR46/EV48/IS51/AK53	37	0.9	-0.9 ± 0.4
FL45/RG50/IS51	35	0.9	-1.2 ± 0.7
LR12/RT13	*nd*	0.9	-1.8 ± 0.4
LQ19	35	0.6	-1.8 ± 0.3

Table I. Stabilities of wild type Arc repressor and proteolytically resistant Arc variant proteins to thermal or guanidine hydrochloride denaturation. Most of the proteins contain multiple mutations. The sequence changes are designated using the single letter amino acid code with the first letter referring to the wild-type residue and the second indicating the mutant amino acid. The number indicates the position in the sequence. Slashes separate different mutations within the same protein sequence. $[GuHCl]_{1/2}$ is the concentration of guanidine hydrochloride and T_m is the temperature at which the protein is half denatured. ΔG_u values were determined from the guanidine denaturation profiles as described in Materials and Methods. 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 about 11 kcal/mole. $\Delta\Delta G_u = \Delta G_{wild\ type} - \Delta G_{mutant}$

nd; not determined

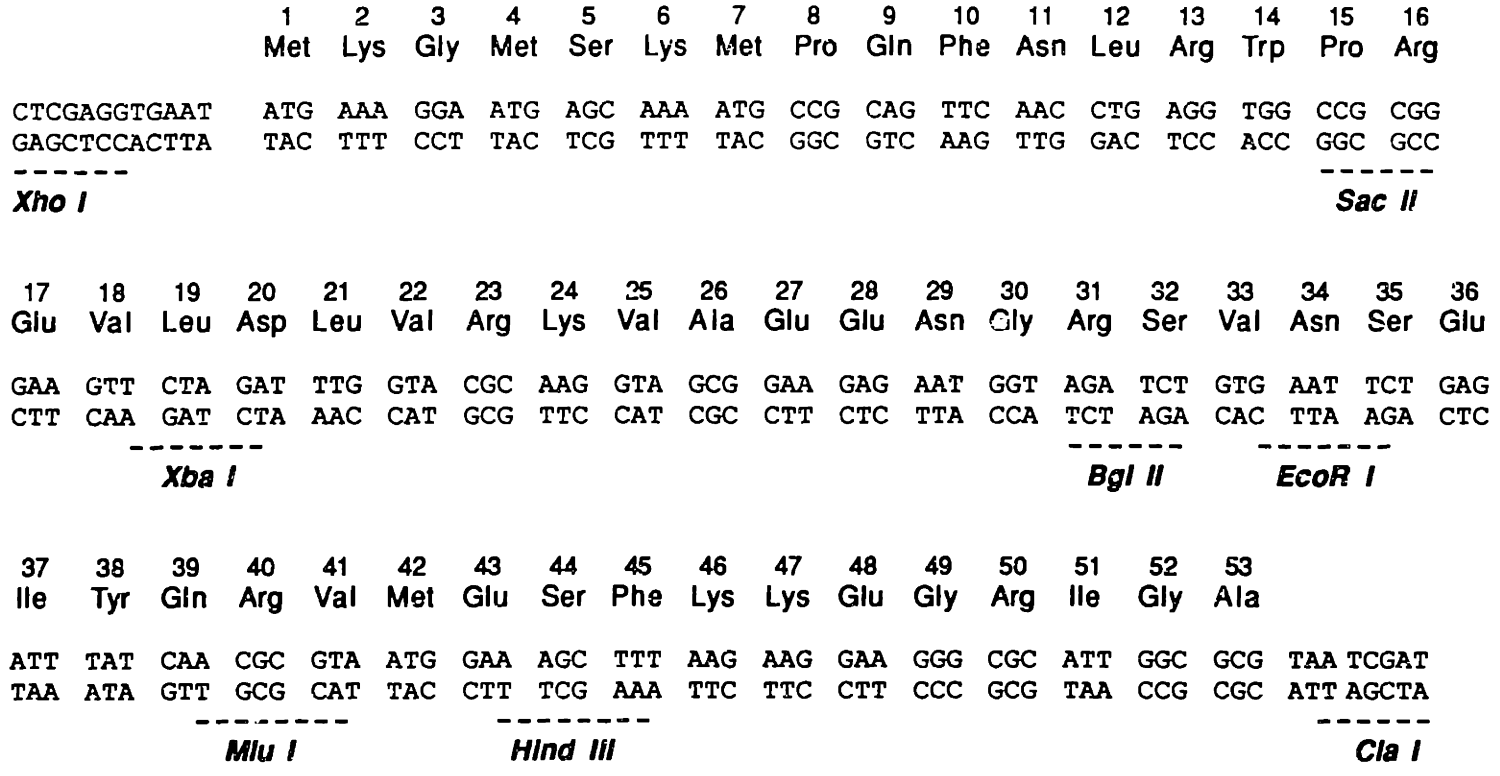
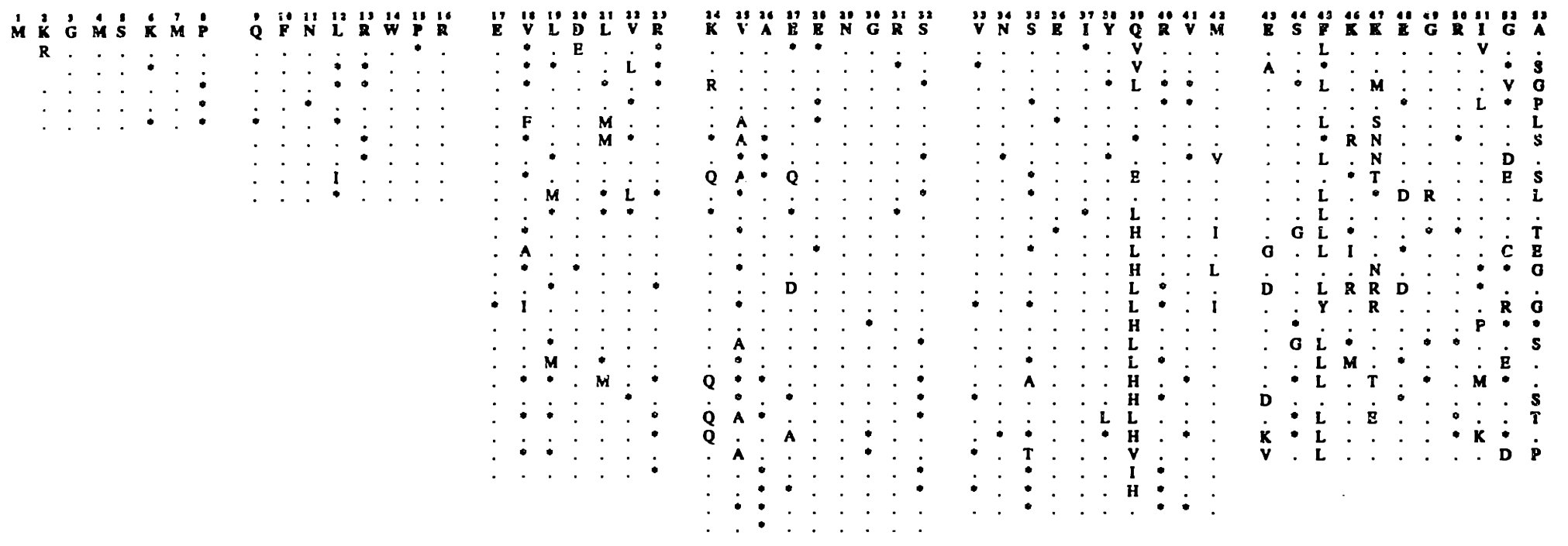


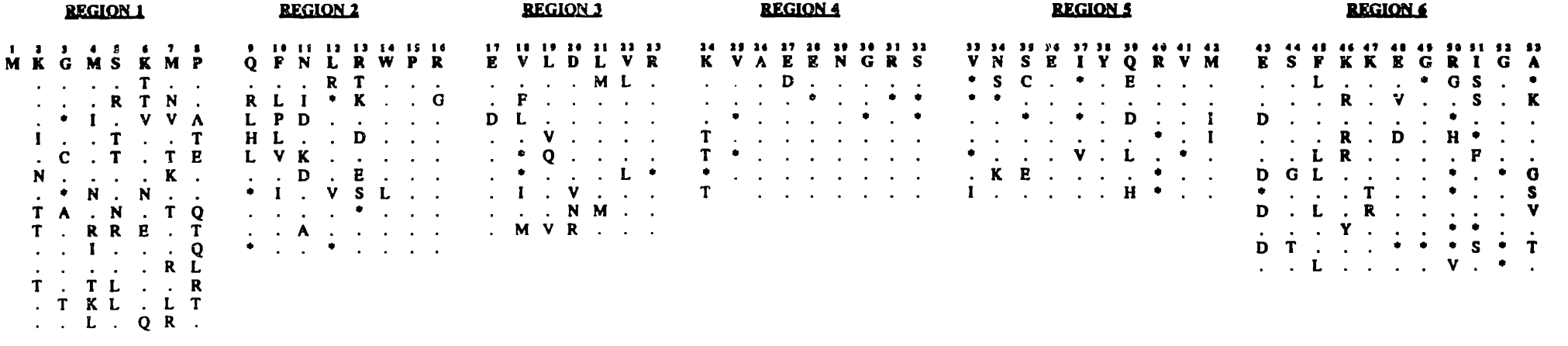
Figure 1. Sequence of the synthetic *arc* gene in plasmid pSA300. The gene sequence for plasmid pSA200 is identical to the one shown except that the *Sac* II site is replaced with a *Sma* I site.

Figure 2. Sequences that allow the Arc protein to remain active (top) or form a stable structure (bottom). 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.

ACTIVE SEQUENCES



STABLE SEQUENCES



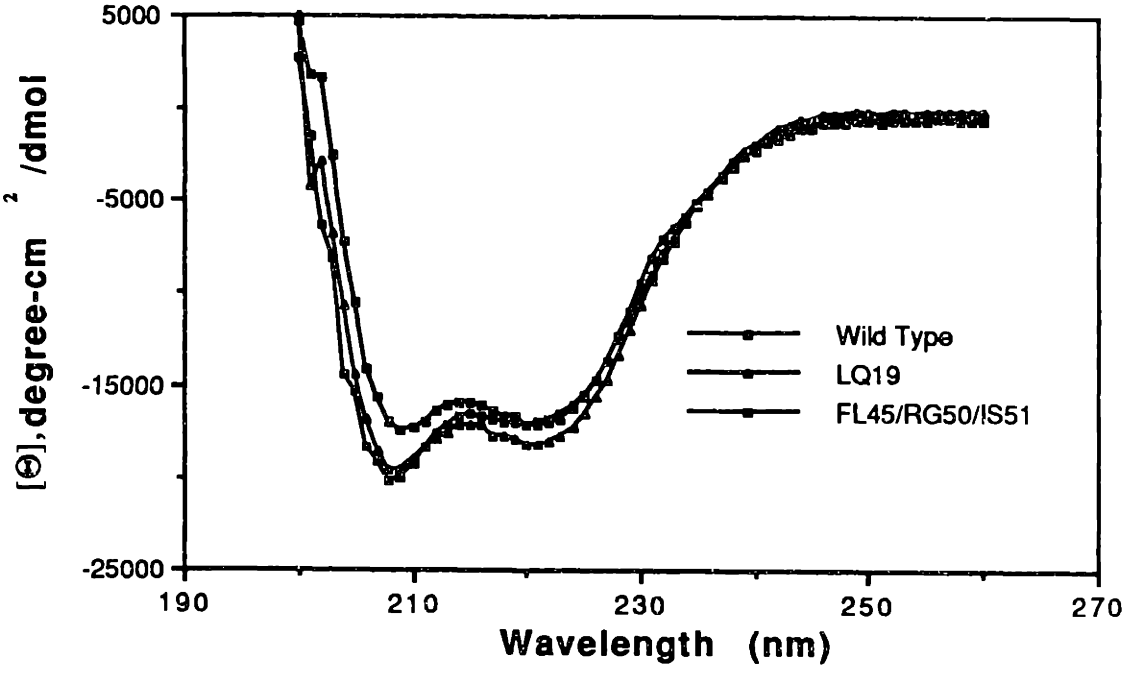
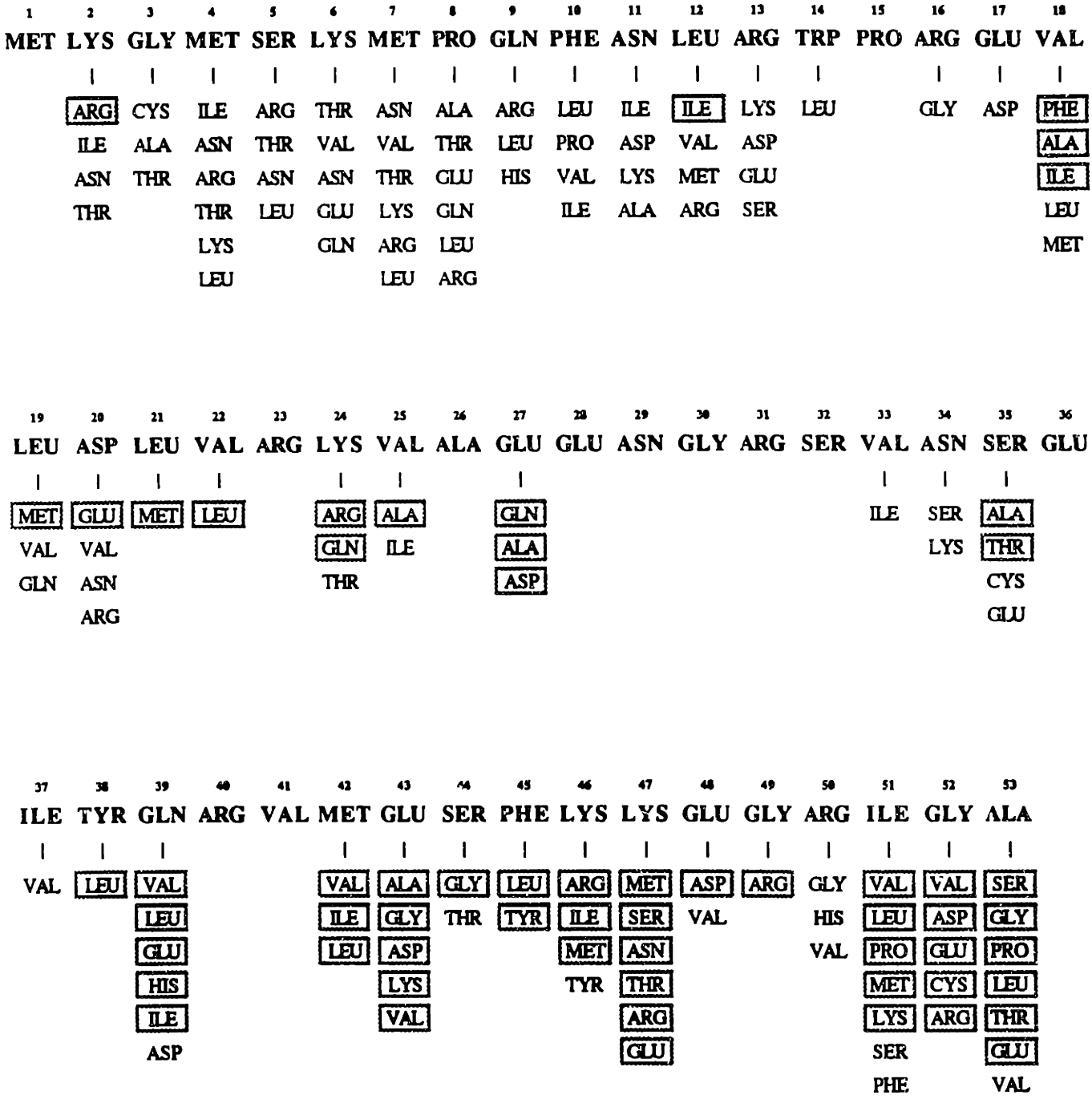


Figure 3. Circular dichroism spectra of wild type Arc protein (■), LQ19 protein (▲), and FL45/RG50/IS51 protein (□).

Figure 4. Summary of amino acid changes tolerated in the sequence of Arc repressor. All the sequence changes shown allow the formation of a folded structure. The subset of these changes that are known to be functionally tolerated are boxed.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
MET LYS GLY MET SER LYS MET PRO GLN PHE ASN LEU ARG TRP PRO ARG GLU VAL
ARG CYS ILE ARG THR ASN ALA ARG LEU ILE ILE LYS LEU GLY ASP PHE
ILE ALA ASN THR VAL VAL THR LEU PRO ASP VAL ASP ALA
ASN THR ARG ASN ASN THR GLU HIS VAL LYS MET GLU ILE
THR THR LEU GLU LYS GLN ILE ALA ARG SER LEU
LYS GLN ARG LEU MET
LEU LEU ARG
19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
LEU ASP LEU VAL ARG LYS VAL ALA GLU GLU ASN GLY ARG SER VAL ASN SER GLU
MET GLU MET LEU ARG ALA GLN ILE SER ALA
VAL VAL GLN ILE ALA LYS THR
GLN ASN THR ASP CYS
ARG GLU
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
VAL LEU VAL VAL ALA GLY ARG MET ASP ARG GLY VAL VAL SER
LEU ILE GLY THR TYR ILE SER VAL HIS LEU ASP GLY
GLU LEU ASP MET ASN VAL PRO GLU PRO
HIS LYS TYR THR MET CYS LEU
ILE VAL ARG LYS ARG THR
ASP GLU SER GLU
PHE VAL

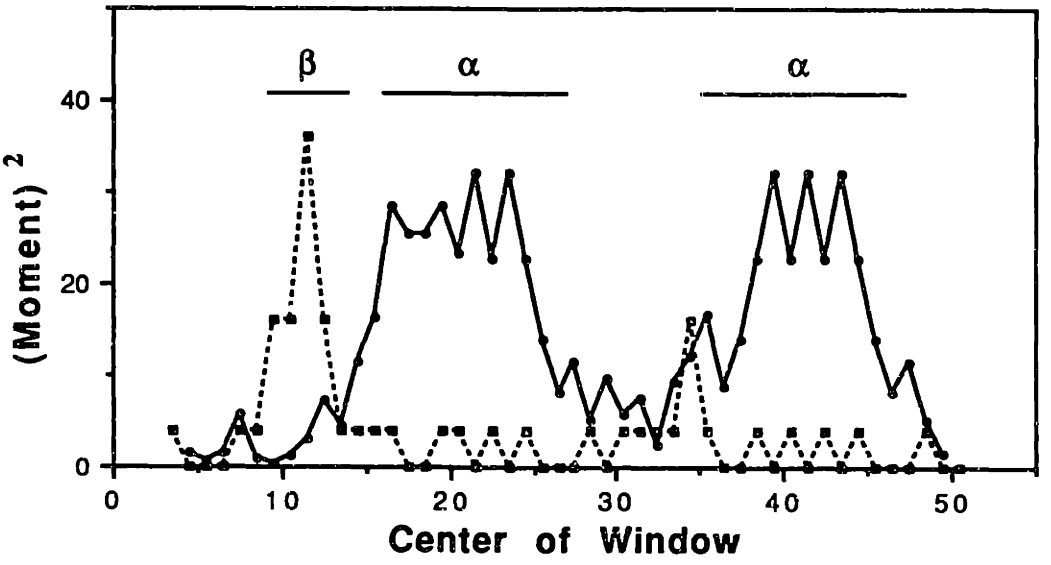


Figure 5. Hydrophobic moment plots of a composite Arc sequence assuming an α-helical conformation (solid line) or β-strand conformation (dashed line). The squares of the moments are plotted to highlight regions of greater amphiphilic character.

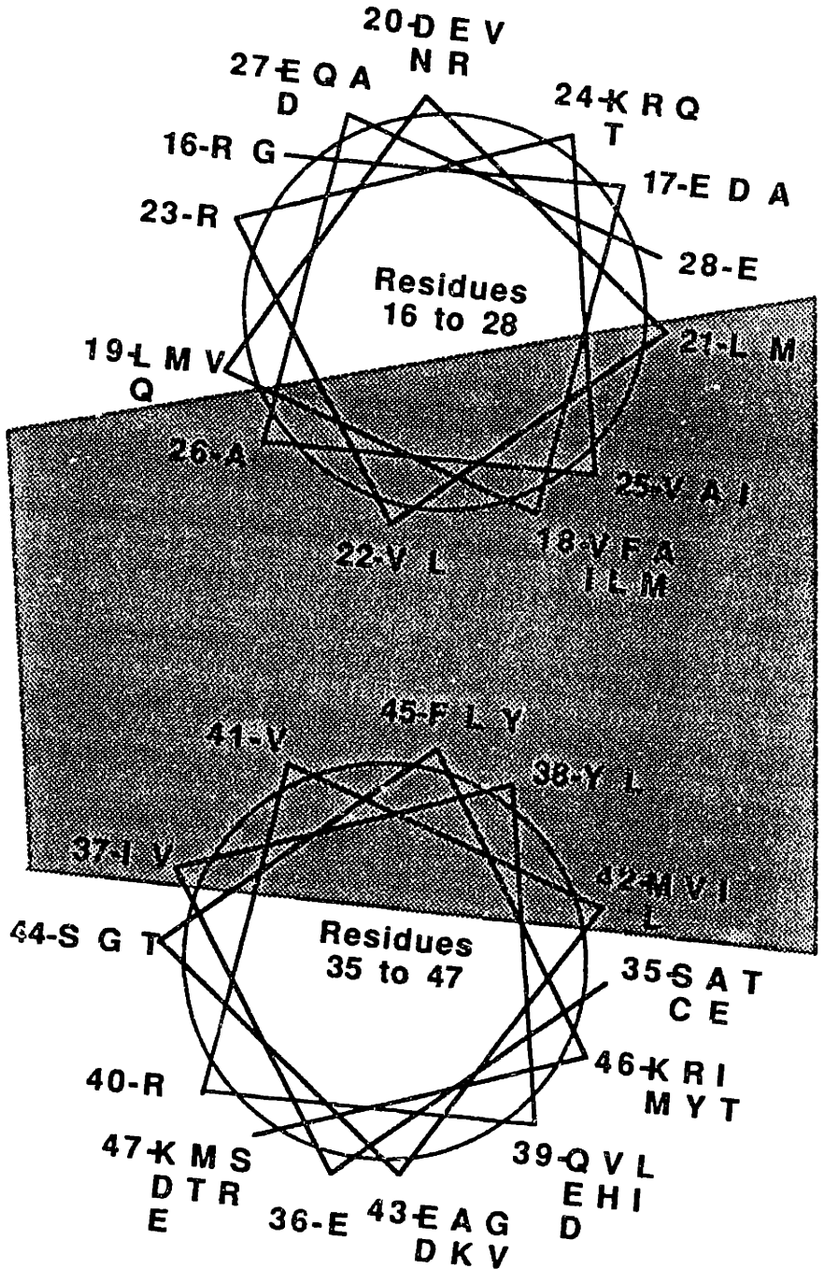


Figure 6. Helical wheel projections of of positions 16-28 and 35-47. The wild type residue and the structurally allowed substitutions are shown for the two regions of the Arc sequence. The hydrophobic faces of the helices lie within the shaded area.

APPENDIX I

Summary of Amino Acid Substitutions Identified in Arc Repressor

Res	WT Seq	Mnt Seq §	Functionally Allowed †	Structurally Allowed †	Functionally Disallowed¥	Comments
2	K	-	R*	R*, I, N, T	T, Q	Functionally important. Purified KT2 has wild-type stability, but operator binding down 250X (1).
3	G	-	-	G, C, A, T	R, K	Functionally important. Purified GR3 has wild-type stability, but operator binding down 200X (1).
4	M	A	-	I, N, R, T, K, L	L, R, I	Functionally important. Purified MI4 has wild-type stability, but operator binding down 27-60X (1; 2).
5	S	R	-	R, T, N, L	R, C	Functionally important. Purified SC5 has wild-type stability, but operator binding down 500X (1)
6	K	D	-	T*, V, N, E, Q	T¶	Functionally important. KT6, isolated as single mutant producing stable protein in randomization experiment, was inactive *in vivo* (J. Bowie, unpublished).
7	M	D	-	N,V,T,E,Q,L,R	-	Functionally important
8	P	P	-	A, T, E, Q, L, R	L	Functionally important. Purified PL8 is much more stable than wild type (Tm (pH4) = 56°C, $[GuHCl]_{1/2}$ = 1.9 M), DNA binding down 200X (1). Pro may be destabilizing in the wild-type protein relative to any other amino acid. Two multiple mutants that change position 8, KT2/MR4/SR5/KE6/PT8 and KT2/GA3/SN5/MT7/PQ8, are much more stable than wild-type(3).
9	Q	H	-	R, L, H*	H¶, L	Functionally important. QH9 was isolated as a single mutant producing stable protein in the randomization experiment and was purified. Stability unknown but was purified easily and behaved normally. QH9 was inactive in operator binding (R.D. Kelley, unpublished).

10	F	F	-	L, P, V, I	V, C, S	Functionally and structurally important/Partially buried. FV10 purified, wild-type stability, operator binding down120 X. May not make DNA contact but mutations could affect structure in the region that is very near the DNA.
11	N	N	-	I, D, K, A*	K	Functionally important.
12	L	F	I*	I*, V, M, R	W, F	Structurally important/Partially buried. Appears to require a hydrophobic residue although the aliphatic region of Arg can substitute and still allow the guanidino group to escape to solvent. May be important for function since substitutions could affect structure near the DNA.
13	R	R	-	K, D, E, S, T	W, Q	Functionally important.
14	W	M	-	L	R	Structurally important/Buried. Only one acceptable substitution found and was obtained in combination with 3 other substitutions. The fluorescence spectrum of Arc is dominated by this unique Trp and folding is associated with a blue shift in l_{max} and an intensity increase. This is consistent with the burial of Trp in a hydrophobic environment (4).
15	P	P	-	-	S, L	Structurally important. May be required for a specialized turn or to initiate helix1.
16	R	M	-	G	W, Q	Functionally important. Assignment is extremely tentative. In the randomization experiment, the only structurally allowed substitution, Gly, came through in combination with 4 other substutions. In this experiment, however, only the first position of codon 16 was mutagenized because of an overlapping restriction site. Consequently, few substitutions were expected.

17	E	E	-	D	A	Not important *in vitro*. From the substitution pattern alone, this would appear to be a structurally important position requiring a negative charge. EA17 was purified, however, and found to have wild-type stability and operator binding activity (1). On the other hand, it was produced at very low levels in the cell. It is possible that the negative charge is in some way important *in vivo*.
18	V	V	F^*, A^*, I^*	F^*, A^*, I^*, L, M	E, G	Structurally important/Partially buried. Appears to require a hydrophobic side-chain, although accepts considerable size and shape changes. VG18 was purified in the *st*1 and *lt*1 backgrounds. Was considerably destabilized ($[GuHCl]_{1/2}$ = 0.5-0.6 M) and DNA binding affinity reduced 27 X most likely due to the decreased stability (2).
19	L	R	M^*	M^*, V^*, Q^*	-	Structurally important/partially buried. Gln tolerated but the purified LQ19 protein has significantly reduced stability (Tm (pH7.5) = 35 °C, $[GuHCl]_{1/2}$ = 0.6 M) (3). This position may be partially buried so that the amide group can still escape to solvent.
20	D	E	E^*	E^*, V, N, R	N, Y, A	Not very important. The negative charge is probably favored but not crucial. DN20 was purified. Somewhat reduced stability, (Tm(pH4) = 26, $[GuHCl]_{1/2}$ = 1.0 M) and DNA binding was reduced only 4 fold (1).
21	L	K	M^*	M^*	F	Structurally Important/Buried. Tolerates a single residue that is conservative in size and hydrophobicity.
22	V	L	L^*	L^*	-	Structurally Important/Buried. Tolerates a single residue that is conservative in size and hydrophobicity.
23	R	K	-	-	C, H	Structurally Important RH23 was purified and has reduced stability ($[GuHCl]_{1/2}$ = 0.8 M). This protein is stable enough to pass the Arc level screen, however. Perhaps His can provide enough of a positive charge to substitute, albeit poorly.

24	K	F	R*, Q*	R*, Q*, T*	T¶	Not very important. KT24 was purified, has somewhat reduced stability (Tm(pH7.5) = 44 C, $[GuHCl]_{1/2}$ = 1.0 M).
25	V	R	A*	A*	I, G	Structurally important/Buried. VI25 was purified. Stability very close to that of wild type (Tm(pH4) = 31, $[GuHCl]_{1/2}$ = 1.1 M) and operator binding is only diminished 4 fold.
26	A	A	-	-	T, V	Structurally important/Buried. Apparently cannot even tolerate slight size variation.
27	E	E	Q, A, D*	Q, A, D*	-	Not very important.
28	E	A	-	-	K, A	Structurally important.
29	N	N	-	-	Y, T, K	Structurally important.
30	G	G	-	-	-	Structurally important.
31	R	R	-	-	W, L	Structurally important.
32	S	S	-	-	A, F	Structurally important.
33	V	M	-	I	G	Structurally important/Buried.
34	N	N	-	S, K	H, K	Functionally important.
35	S	S	A, T	A, T, C, E	-	Not very important. All the functionally allowed substitutions have small side chains, however. This feature may be important in allowing Arc to function.
36	E	E	-	-	K, A, G	Structurally important.
37	I	L	-	V	M	Structurally important/Buried.

38	Y	L	L	L	D, S, C	Structurally important/Buried.
39	Q	Q	V, L, E, H, I[#]	V, L, E, H, I, D[#]	P	Not very important. Pro mutation would force a kink in helix2.
40	R	I	-	-	Q	Structurally important.
41	V	V	-	-	A	Structurally important/Buried.
42	M	Q	V*, I*, L	V*, I*, L	-	Structurally important/Buried
43	E	D	A, G, D*, K, V	A, G, D*, K, V	G, A	Not very important. The apparent discrepancy between mutants and allowed substitutions is resolved by the properties of purified EA43. This protein has wild-type stability and DNA binding activity.
44	S	A	G	G, T	-	Not very important. Small size may be an essential feature at this position, however.
45	F	L	L*, Y	L*, Y	S, C	Structurally important/Partially buried. It is likely that the hydroxyl group of Tyr is able to escape to the solvent.
46	K	S	R, I, M,Y*	R, I, M, Y*	T	Structurally important/Partially buried. All allowed residues contain a long hydrophobic segment. The mutant does not.
47	K	K	M, S, N, T, E	M, S, N, T, R, E	-	Not very important.
48	E	P	D	D, V	K	Functionally important.

49	G	S	R	R	-	Structurally important. Out of 39 sequences only one substitition, Arg, was obtained at this position in the randomization experiment. Consequently, Gly does seem to be important. How Arg manages to substitute is not clear, although the substitution was isolated in combination with Asp at position 48 and Leu at 53, raising the posibility of compensating changes. Alternatively, the long side chain of Arg provides enough flexibility so that it can find a way to compensate for an otherwise destabilizing situation.
50	R	P	-	G, H, V	P	Functionally important.
51	I	V	V,L,P*,M,K	V,L,P*,M,K,S,F	-	Not very important.
52	G	T	V, D, E, C, R	V, D, E, C, R	-	Not very important.
53	A	G	S,G,P,L,T,E	S,G,P,L,T,E,V	-	Not very important.

§ Underlined residues are conserved in Arc and Mnt. Doubly underlined residues indicate identities in the two proteins.

† Reference 3

¥ Reference 1, unless indicated otherwise

* Isolated as a single mutant at least once.

all mutations at this position were isolated as singles at least once.

¶ Isolated as single in randomization experiment and has defective phenotype *in vivo*

References

1. Vershon, A. K., Bowie, J. U., Karplus, T. M. & Sauer, R. T. (1986) *Proteins: Structure Function and Genetics* **1,** 302-311.

2. Bowie, J. U. & Sauer, R. T. (1989) *J. Biol. Chem.* in press.

3. Bowie, J. U. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 2152-2156.

4. Bowie, J. U. & Sauer, R. T. (1989) *Biochemistry* submitted.

APPENDIX II

Using Neutral Amino Acid Substitutions in Arc to Detect Distant Sequence Relationships

Genetic and biochemical studies indicate that Arc and Mnt repressors are members of a new class of DNA binding proteins, using an extended structure to make major groove contacts (1; 2; 3; 4; 5). It seems unlikely that Arc and Mnt are the only members of this class, although no detectable homology to other DNA binding proteins can be identified using standard methods of sequence comparison. It is generally straightforward to detect sequence similarities at the level of 25% or greater, but it can be difficult to detect more distant relationships (6). One problem with standard methods of comparison is that every residue in the sequence is given equal importance in the alignment score even though many residues may not be important elements of the protein sequence. A knowledge of the substitutions that are allowed in a given protein structure can help to identify residues that are not crucial to the protein fold so that more weight can be given to the important positions. This principle is illustrated in Figure 1. In the first case, Figure 1a, a comparison is made between two sequences. These sequences appear to align poorly. If the substitutions shown in Figure 1b were known to be allowed, however, the alignment might be considered favorable. The entire sequence of Arc repressor has now been probed for allowed substitutions (1). Consequently, we sought to combine this knowledge into a more appropriately weighted sequence homology search.

Gribskov et al. have introduced a method for combining sets of related sequences in a homology search (7). A position-specific scoring matrix called a profile is first constructed from a set of aligned sequences. In this profile, each position has a different scoring matrix that is related to the types of allowed substitutions seen at each position. The score for

finding amino acid **a** at position **p** is given by $M(\mathbf{p},\mathbf{a}) = \sum_{b=1}^{20} W(\mathbf{p},\mathbf{b})xY(\mathbf{a},\mathbf{b})$, where Y(**a**,**b**) is the value from Dayhoff's mutation distance matrix (8) and W(**p**,**b**) is the weight for the appearance of amino acid **b** at position **p**, which is proportional to the number of occurrances of residue **b**.

A profile of the Mnt and Arc sequences along with the allowed substitutions found in Arc repressor was constructed and used to search the PIR[1] database for related sequences. Figure 2 shows a histogram of the number of protein sequences that give a particular score and table I lists the 20 top scoring proteins out of 4580 tested. Arc and Mnt are, as expected, the top scoring proteins. Most of the remainder are very long sequences. Since the probability of finding a significant match is greatly increased in long sequences relative to short ones, the sequence similarity found in these proteins is probably not significant. The third highest scoring protein, however, is TraY from the F plasmid in *E. coli*, a protein of only 131 amino acids. Confidence in the possible significance of the alignment is enhanced by indications that TraY, like Arc and Mnt, is likely to bind to DNA in a sequence specific manner.

The *tra* operon is found on F plasmids of *E. Coli* and is required for bacterial conjugation (9). During conjugation, DNA is transferred from the donor, F containing cell, to a recipient bacterial cell. Both TraY and a TraI protein are required for nicking at the origin, *ori*T, of the F plasmids, although the nicking activity appears to reside in the TraI protein (10; 11). TraI is also a helicase and is required for subsequent strand

---

[1] Protein Identification Resource, National Biomedical Research Foundation, Washington, DC

displacement[1]. Nicking occurs in a conserved region in *ori*T adjacent to a less conserved region. Because TraI nicks DNA within the conserved region, *tra*I genes from one plasmid can complement *tra*I mutants from a variety of otherwise incompatible plasmids. TraY, however, is much less promiscuous leading to suggestions that it binds to a site outside the conserved region. There is indeed a palindromic site whose sequence variation correlates with the ability of *tra*Y from other plasmids to complement *tra*Y mutants, although there are undoubtedly other sequences where similar arguments could be made. Figure 3 shows a highly speculative model of the nicking complex at *ori*T. While this model must be viewed with a high degree of skepticism at this point, it does appear likely that TraY binds specifically to a DNA site. *Tra*Y genes have now been sequenced for several F-like relatives, R1, R100 and ColB4 (11; 12; 13; 14). The TraY proteins from these related plasmids are each 75 amino acids long and thus similar in size to Arc and Mnt.

Figure 4 shows the sequence alignments of the TraY proteins with Arc and Mnt. The alignment can be made without gaps, and there are only two positions (marked with an asterisk) where there are apparent violations of the allowed substitution patterns seen in Arc. The significance of the alignment shown was tested by generating alignment scores for 100 randomized sequences of the TraY sequence from plasmid F. The mean of the random scores was 5.6 with a standard deviation of 0.9. The alignment with the natural TraY sequence is thus 6 standard deviations above the mean.

---

[1] The nicking activity was previously thought to reside in a separate gene, *traZ*.

REFERENCES

1. Bowie, J. U. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* in press.

2. Knight, K. L., Bowie, J. U., Vershon, A. K., Kelley, R. D. & Sauer, R. T. (1989) *J. Biol. Chem.* **264,** 3639-3642.

3. Knight, K. L. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci., USA* **86,** 797-801.

4. Vershon, A. K., Bowie, J. U., Karplus, T. M. & Sauer, R. T. (1986) *Proteins: Structure Function and Genetics* **1,** 302-311.

5. Zagorski, M. G., Bowie, J. U., Vershon, A. K., Sauer, R. T. & Patel, D. J. (1989) manuscript.

6. Taylor, W. R. (1988) *Protein Engineering* **2,** 77-86.

7. Gribskov, M., Mclachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355-4358.

8. Dayhoff, M.O. (1979) in *Atlas of Protein Sequence and Sructure*, eds. Schwartz, R.M. & Dayhoff, M.O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 353-358

9. Ippen-Ihler, K. A. & Minkley, E. G., Jr. (1986) *Ann. Rev. Gent.* **20,** 593-624.

10. Everett, R. & Willetts, N. (1980) *J. Mol. Biol.* **136,** 129-150.

11. Traxler, B. A. & Minkley, E. G., Jr. (1988) *J. Mol. Biol.* **204,** 205-209.

12. Inamoto, S., Yoshioka, Y. & Ohtsubo, E. (170) .

13. Finlay, B. B., Frost, L. S. & Paranchych, W. (1986) *J. Bacteriol.* **168,** 132-139.

14. Finlay, B. B., Frost, L. S. & Paranchych, W. (1986) *J. Bacteriol.* **166,** 368-374.

Name	Score	Documentation
RGBPA2	30.292	! Regulatory protein arc - Bacteriophage P22
RGBPM2	18.144	! Regulatory protein mnt - Bacteriophage P22
BVECTY	11.027	! traY protein - Escherichia coli plasmid F
QQBE11	10.720	! 140K ribonucleotide reductase - Epstein-Barr virus (strain B95-8)
HVMST7	10.339	! Ig heavy chain precursor V region - Mouse TEPC 1017
QQBE34	10.093	! Hypothetical BBLF4 protein - Epstein-Barr virus (strain B95-8)
GNWVY	10.033	! Genome polyprotein - Yellow fever virus (strain 17D)
SYECGU	9.999	! GMP synthase (glutamine-hydrolyzing)- Escherichia coli
GNNY2P	9.863	! Genome polyprotein - Poliovirus (type 1, strain Mahoney)
GNNY3P	9.856	! Genome polyprotein - Poliovirus (type 1, strain Sabin)
GNNY1P	9.856	! Genome polyprotein - Poliovirus (type 1, strain Mahoney)
GNNY5P	9.830	! Genome polyprotein - Poliovirus (type 2, strain Lansing)
DEECXA	9.778	! Pyridine nucleotide transhydrogenase, alpha chain - Escherichia coli
GNNY4P	9.777	! Genome polyprotein - Poliovirus (type 3, 2 strains)
HMXRS3	9.735	! Sigma 1 protein precursor - Reovirus (type 3)
ACCHG1	9.666	! Acetylcholine receptor protein, gamma chain precursor - Chicken
IVHO21	9.612	! Interferon alpha-II-1 precursor - Horse
WMTM8T	9.484	! 180K protein - Tomato mosaic virus (strain L)
NUMS	9.481	! Glucose-6-phosphate isomerase (EC 5.3.1.9) - Mouse
ACMSD1	9.478	! Acetylcholine receptor protein, delta chain precursor - Mouse

Table I. The twenty protein sequences that align most favorably with Arc/Mnt profile. The profile was constructed using logarithmic weighting of residue occurrences. Gap opening and gap extension penalties were 3.0 and 0.1, respectively.

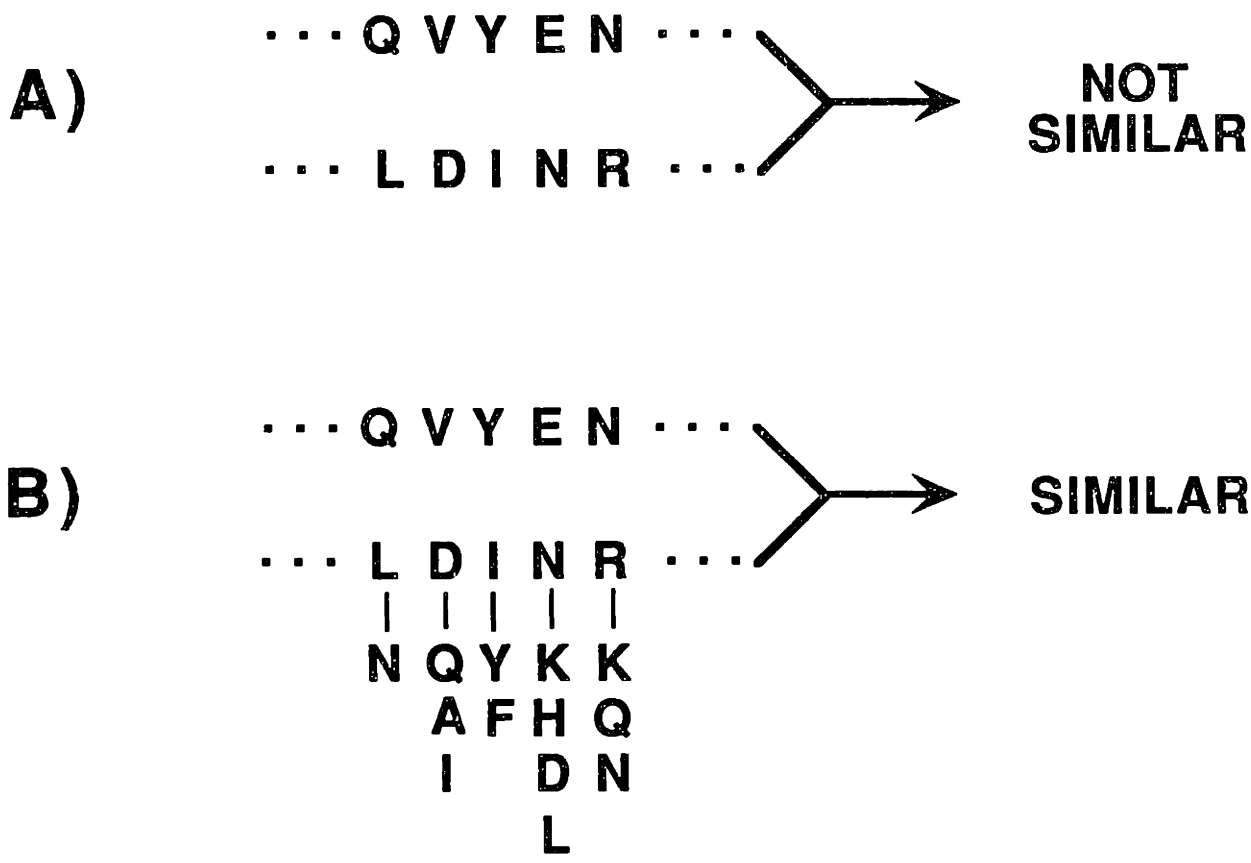


Figure 1 Comparison of sequence alignments with (A) and without (B) allowed substitution data.

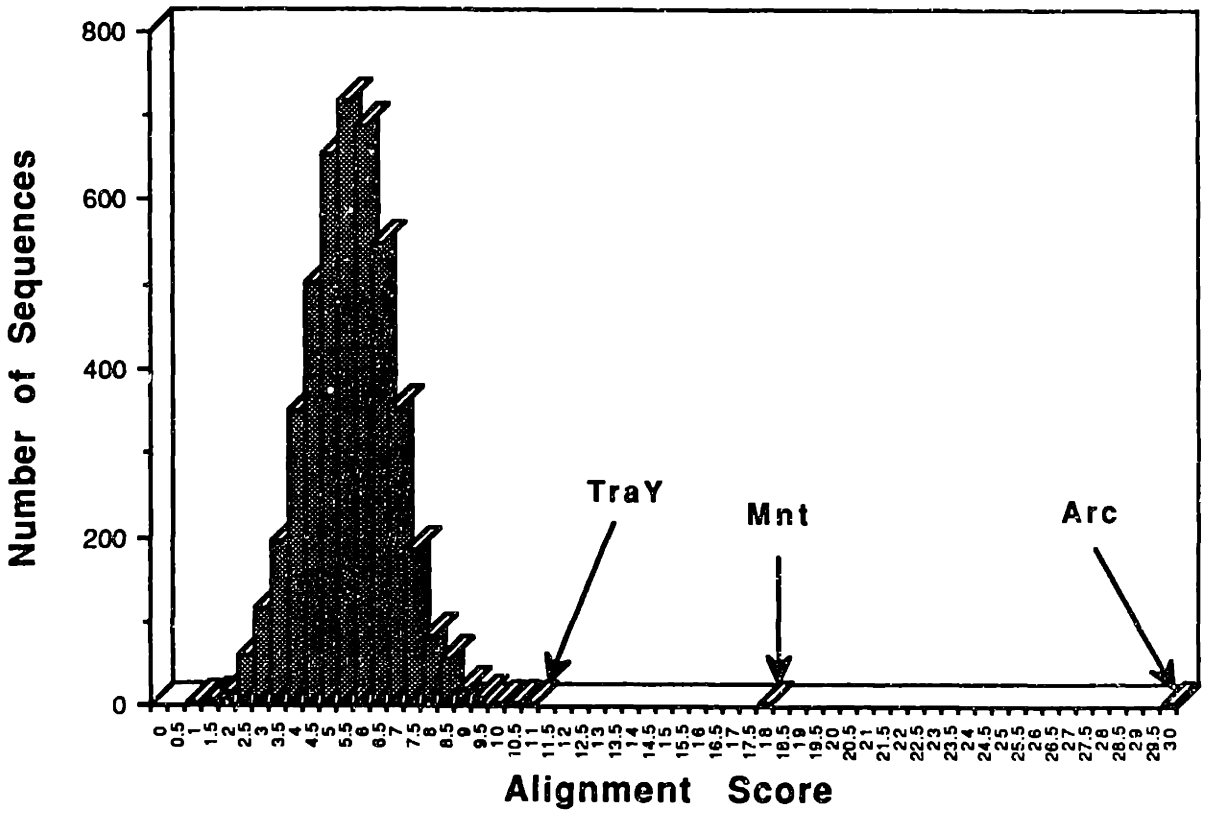


Figure 2. Distribution of alignment scores using the Arc/Mnt profile.

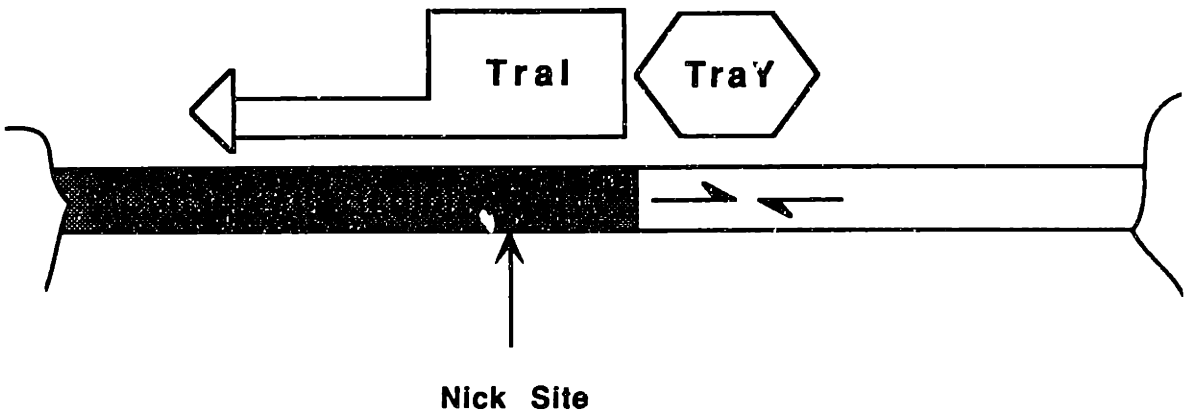


Figure 3. A possible arrangement of proteins required for nicking at oriT.

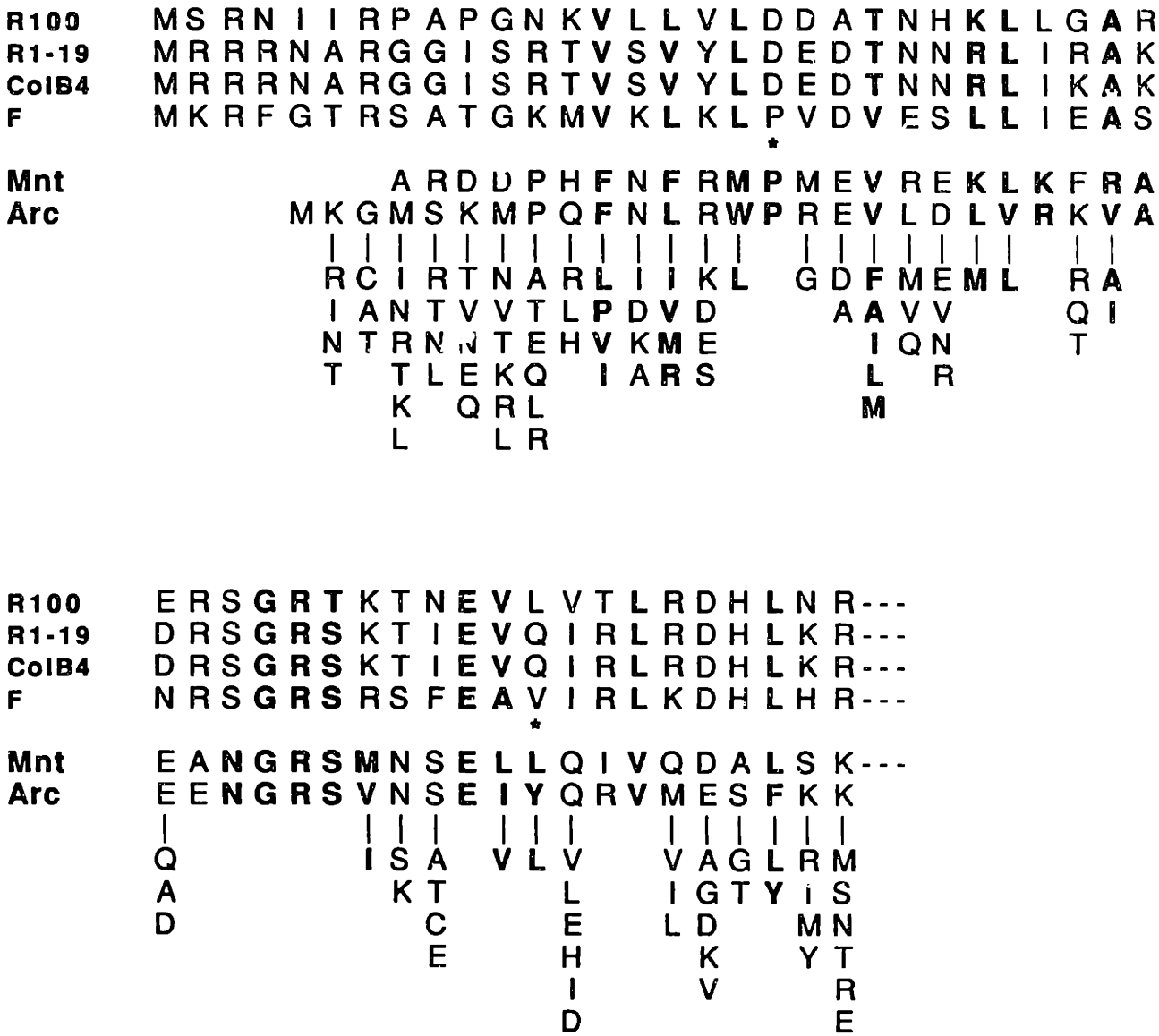


Figure 4. Sequence alignments of the TraY proteins (upper group) with Mnt and Arc (lower group). Amino acid substitutions that are allowed in Arc repressor are shown below the Arc repressor sequence. Positions that are conserved in Mnt and Arc variants are highlighted in bold. The TraY sequences are also highlighted at these positions if obvious conservation is seen. If the allowed substitution pattern is clearly violated, the position is marked by an asterisk.

CHAPTER 3

Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets with Solvent Accessibility Patterns of Known Structures

INTRODUCTION

Residues buried in the interior of a protein are almost invariably hydrophobic, and removing these hydrophobic residues from water is a major factor in protein folding and stability (1; 2). The importance of these residues is clearly demonstrated by the conserved hydrophobic character of buried residues in related proteins (3; 4; 5; 6; 7). Surface residues are more variable than buried residues; while they often are hydrophilic, hydrophobic amino acids also seem to be accommodated at many positions. Because of this difference, comparison of the sequences of related proteins allows one to identify residues that are likely to be buried and residues that are likely to be on the surface (6; 8). Such an analysis is most reliable when the sequences of many related proteins are known or when large numbers of neutral amino acid substitutions have been generated and analyzed (8; 9).

In this paper, we test whether patterns of conserved hydrophobic residues, inferred from a comparison of related protein sequences, are sufficiently characteristic of the three-dimensional fold to allow the identification of the correct structure in a database of known protein structures. If, in general, the preference for hydrophobic amino acids at particular sequence positions indicates that these residues are buried, then the pattern of hydrophobic residues should be reflected in a corresponding pattern of buried residues in the folded protein. We find that, in many cases, the similarity of these patterns is indeed sufficient to correctly identify the tertiary fold adopted by a set of protein sequences.

MATERIALS AND METHODS

Sequences, Sequence Alignments and Protein Structures: The following sets of sequences were used in this study: (i) the sequences of nine globins that were aligned by Lesk and Chothia (4) from the crystallographically determined structures; (ii) nine M class cytochromes from an alignment based on the tuna cytochrome c structure (10) and one additional sequence, rice cytochrome c, whose structure has recently been found to closely resemble that of tuna cytochrome c (11); (iii) the CheY protein family including the protein sequences of CheY, CheB, SfrA, OmpR, SpoA, SpoF, $NtrC_{Kp}$, $NtrC_{Bp}$, VirG, DctD, PhoB and ORF2 (J. Stock, personal communication); and (iv) the EF-Hand sequences from Szebenyi et al. (12).

The following 103 protein structures from the Brookhaven databank (13) were used in this study (Brookhaven designation): 156B, 1ABP, 1ACX, 1APR, 1BP2, 1CC5, 1CCR, 1CRN, 1CTF, 1CTX, 1ECD, 1FB4, 1FBJ, 1FC2, 1FDX, 1GCN, 1GCR, 1GP1, 1GPD, 1HIP, 1HMG, 1HMQ, 1INS, 1LZ1, 1MBD, 1MLT, 1NXB, 1PCY, 1PPD, 1PPT, 1PYP, 1RHD, 1RN3, 1RNS, 1SBT, 1SN3, 1TGN, 1TIM, 1TPA, 1UBQ, 2ABX, 2ACT, 2ADK, 2ALP, 2APP, 2AZA, 2B5C, 2CAB, 2CCY, 2CDV, 2CGA, 2CNA, 2CYP, 2EBX, 2EST, 2FD1, 2GCH, 2GN5, 2GRS, 2HHB, 2LH4, 2LZM, 2MDH, 2MT2, 2OVO, 2PAB, 2PKA, 2PTC, 2RHV, 2SGA, 2SNS, 2SOD, 2SSI, 2STV, 2TAA, 2TBV, 351C, 3C2C, 3CPV, 3CTS, 3FXC, 3ICB, 3PGK, 3PGM, 3RP2, 3SGB, 3TLN, 3WGA, 4ADH, 4APE, 4ATC, 4CYT, 4DFR, 4FXN, 4LDH, 4SBV, 5API, 5CHA, 5CPA, 5PTI, 5RXN, 8CAT, 9PAP. Our database also includes: Rop (14), catabolite gene activating protein (15), λ Cro repressor (16), the N-terminal domain of

bacteriophage 434 repressor (17) and the N-terminal domain of bacteriophage λ repressor (18). Our "database of known structures" consists of the unique subunits of these 108 proteins.

Hydrophobicity Patterns Derived from Sets of Aligned Sequences: Sets of aligned sequences were used to identify positions that require hydrophobic residues, and positions that can accommodate hydrophilic side chains. Each position in the set of aligned sequences was characterized by picking (with the steps described below) one of the most hydrophilic residues allowed at that position. Specifically: 1) The amino acids found at each position in the aligned sequences were ranked by hydrophobicity, using the scale of Fauchere and Pliska (19). 2) To account for possible alignment errors, sequence errors and occasional exceptions, the most hydrophilic residue found at each position was discarded (unless it was observed more than once), and 3) the most hydrophilic of the remaining residues was then picked as a measure of the extent to which non-hydrophobic residues are tolerated at this position. Arg and Lys residues were not included in this ranking unless they were the only residues found at a particular position. (Inspection of several globin structures showed that the long, aliphatic parts of the Arg and Lys side chains can effectively substitute for hydrophobic amino acids at some buried positions and still permit the charged groups at the ends of the side chains to reach the solvent.) This algorithm selects a relatively hydrophilic amino acid from the set of observed residues at each position and thus provides a rough measure of the "allowed hydrophilicity" at each site.

The steps listed above pick one amino acid to represent each position in the aligned sequences. This representation was further simplified by classifying these residues into three categories. Thus, each position in the aligned sequences is represented as H_1 (high hydrophobicity), H_2 (medium hydrophobicity), or H_3 (low hydrophobicity). H_1 is used if the amino acid which represents the allowed hydrophilicity at that position is Trp, Ile, Phe, Leu, Met, Val or Cys. H_2 is used if the amino acid which represents the allowed hydrophilicity at that position is Tyr, Pro, Ala, Thr, His, Gly, or Ser. H_3 is used if the amino acid which represents the allowed hydrophilicity at that position is Gln, Asn, Glu, Asp, Lys or Arg. The rationale for using these particular classifications is explained below.

Solvent Accessibility Patterns Determined from Known Structures: For each unique subunit in our database of known structures, a string was created to represent the solvent accessibility at each of the consecutive residues along the polypeptide chain. The solvent accessibility of each position was designated as B_1 (buried), B_2 (partially buried), or B_3 (exposed). Solvent accessibilities were calculated using the Lee and Richards algorithm (20) as implemented in the program ACCESS by Handschucher and Richards. The fractional solvent exposure for each residue in the structures was determined by calculating the solvent exposed area of the Cα and side chain atoms, and then dividing by the solvent exposed area of the same atoms in an extended Ala-X-Ala tripeptide (20). When determining which residues in a particular chain were buried, we ignored ions, all prosthetic groups except heme and the atoms of other subunits. We decided to use isolated subunits because our method of

determining a single hydrophobicity pattern introduces a bias toward monomeric structures. Although a multimeric protein (such as hemoglobin) might have hydrophobic residues at the subunit interfaces, a related monomeric protein (such as myoglobin) probably would have hydrophilic residues at the corresponding positions. Since each of our sequence sets includes monomeric proteins, and since our hydrophobicity analyses are heavily biased by the most hydrophilic residues in the set of aligned sequences, the hydrophobicity pattern for the set of sequences should most closely match the accessibility pattern calculated for the monomeric proteins and for the isolated subunits of the multimeric proteins.

Aligning Hydrophobicity and Accessibility Strings: The hydrophobicity string determined from aligned sequences was compared with each of the accessibility strings for the known structures, and the optimal alignment was determined using the Needleman and Wunsch algorithm (21). The score for pairing a particular accessibility group, B_i, and a particular hydrophobicity group, H_j, is given by the information value, $I(H_i:B_j)$, for that pairing (22). The information value is determined by $I(H_i:B_j) = \ln (P(H_i:B_j)/P(H_i))$, where $P(H_i:B_j)$ is the probability (i.e, the observed frequency) that the hydrophobicity group of a residue position is H_i if the accessibility group of that position is B_j and $P(H_i)$ is the probability of any residue being a member of hydrophobicity group H_i.

The range of values that define the three fractional accessibility categories and the three hydrophobicity categories were chosen by comparing the fractional solvent accessibility values for every residue in

five globin chains (sperm whale myoglobin, human hemoglobin α and β chains, root nodule leghemoglobin and larval insect erythrocrourin) with the residues that characterize the allowed hydrophilicity at each of these positions. The cutoffs used in defining each accessibility and hydrophobicity category were adjusted to maximize the total information, I_{total}, according to:

$$\mathbf{Max}\left(I_{total}\right) = \mathbf{Max}\left(\sum_{i=1}^{3}\sum_{j=1}^{3} N(H_i,B_j)\,\ln\left(\frac{P(H_i:B_j)}{P(H_i)}\right)\right)$$

where H_i is the i^{th} hydrophobicity grouping, B_j is the j^{th} solvent accessibility grouping and $N(H_i, B_j)$ is the total number of residues that are both in hydrophobicity group H_i and accessibility group B_j. The optimal accessibility groups were found to be 0-10%, 11-39% and greater than 39% solvent exposure. The optimal hydrophobicity groupings were those listed above. Figure 1 shows the data from the globin alignments and the scoring table derived from these data. Gaps in the alignment were prohibited within alpha-helices or beta-sheets. (Helices and sheets were identified by the method of Kabsch and Sander (23)). Gaps in other areas were allowed, but a gap opening penalty of 3.0 and a gap extension penalty of 0.2 were applied. These gap penalties were determined empirically by maximizing the difference between the scores for the globin structures and the score for the highest scoring non-globin structure. In every case, the significance of the alignment scores was assessed by generating 50 random rearrangements of the solvent accessiblity string, determining the best alignment score for each of these randomly rearranged sequences, and

comparing the score of the correct solvent accessibility sequence with the mean and standard deviation of the scores for the randomized strings. Scores are reported as the number of standard deviations above or below the mean of the scores for the randomized strings.

RESULTS

The goal of this work was to determine whether sequence information, from multiple aligned sequences, can be used to find a correct tertiary fold. We tried to find a known structure, with its characteristic pattern of solvent accessibility values along the polypeptide chain, that matches the hydrophobicity pattern seen in a set of sequences. Accessibility and hydrophobicity clearly are correlated (24), but the correlation is generally weak in single sequences because it is not uncommon to find hydrophobic residues on the surface of a protein. However, it is far less common for a surface hydrophobic residue to remain hydrophobic throughout the course of evolution. In a set of evolutionarily related sequences, most surface residues will accommodate hydrophilic substitutions, and the pattern of hydrophobic positions inferred from a set of sequences will be more informative than the pattern of hydrophobic positions in any single sequence.

Figure 2 summarizes our approach. Each protein subunit in the database is represented as a string of solvent accessibility values, with B_1 used to represent fully buried position, B_2 used to represent partially buried positions, and B_3 used to represent exposed positions. For a set of

sequences, a single string is generated to represent the extent to which hydrophilic residues are tolerated at each position (see Materials and Methods). Then each structure in the database is used as a "trial structure", and the best alignment between this sequence hydrophobicity pattern and the solvent accessibility pattern of the trial structure is determined. The alignments are found using the algorithm of Needleman and Wunsch (21) and the scoring system uses weights that represent the information value for each possible pairing of a hydrophobicity group and an accessibility group. This procedure gives the best alignment with each of the trial structures, and the structure with the best overall score is assumed to be the most plausible structure.

As described in Materials and Methods, we divided both the hydrophobicities and the solvent accessibilities into three groups . Each of the nine possible pairings between a solvent accessiblity group and a hydrophobicity group was assigned a weight based on how strongly the two groups were correlated in alignments of globin sequences and globin structures. Cutoffs for the hydrophobicity and accessibility groups were chosen by maximizing the ability of the program to distinguish the five globin structures from all other structures in the database.

Figure 1 shows the final scoring table used in evaluating the alignments. In the scoring table, positive numbers indicate that the two properties being compared are positively correlated in the globins; negative numbers indicate that the properties are negatively correlated. The most heavily weighted elements occur at the corners of the table. For example, a position that remains in the highly hydrophobic group, H_1, over the course

of evolution is likely to be a member of the buried group, B_1, and is very unlikely to be in the exposed group, B_3. A position that accommodates hydrophilic residues (H_3) is fairly likely to be exposed (B_3) and is very unlikely to be buried (B_1). Because insertions or deletions in evolutionarily related proteins usually do not occur within alpha-helices or beta-sheets, gaps were not allowed in these regions (3; 4; 5; 7). Gap penalties for insertions or deletions in other regions were determined empirically to optimize scores for the globins.

Figure 3 shows the distribution of scores when the globin hydrophobicity pattern is matched with each protein structure in the database. The five highest-scoring proteins, which are clearly separated from all other structures, are the five globin structures that were present in the database. This shows that the pattern of hydrophobicity extracted from a set of aligned sequences can be used to identify the correct protein fold among a large collection of possible structures. To determine whether this method works on protein sequences and structures that were not used to adjust the scoring system, the algorithm was tested on several sequence sets, with the same scoring table and gap penalties used in the globin alignments. So that the method could be evaluated, sequence sets were chosen in which the crystal structure of at least one member had been solved.

Cytochrome C: The cytochrome c proteins are a relatively diverse family of heme proteins that can be grouped into four structural classes of different sizes. We considered sequences of the M class of cytochromes since two of these structures were in our database. Figure 5a summarizes the results of

searching the database of known structures for the accessibility pattern that best matches the M class cytochrome hydrophobicity pattern. The two M class cytochromes in the structural database are the two highest scoring protein structures.

CheY Protein: The CheY protein is involved in chemotactic signal transduction in *E. coli* and shows sequence similarity to a number of other bacterial regulatory proteins (25). The crystal structure of the CheY protein has recently been determined (25) and is reportedly quite similar to that of flavodoxin (J. Stock, personal communication). Both proteins contain mixed α and β secondary structure elements. This is an interesting test case for our algorithm since no statistically significant homology to flavodoxin is observed by conventional sequence alignment methods (data not shown). As shown in Figure 4b, the flavodoxin structure is identified by our search procedure as the structure in our database which is most consistent with the hydrophobicity pattern for CheY and its homologues.

Calcium Binding Proteins: These proteins have a helix-loop-helix super-secondary structure, referred to as the EF-hand, and each domain of these proteins typically contains two such EF-hands. We used the combined hydrophobicity pattern of EF-hand sequences to search the structural database and the results are summarized in Figure 4c. One of the two calcium-binding proteins in the database, bovine intestinal calcium binding protein (ICBP), score . significantly higher than any other protein structure in the database. The other calcium-binding protein in the database, parvalbumin, did not score highly. Although these two calcium binding

proteins contain qualitatively similar helix arrangements, they do differ substantially in both interhelix angles and helical lengths (12). Clearly, ICBP and parvalbumin are sufficiently different that the solvent accessibility pattern for ICBP is well matched by the consensus hydrophobicity pattern while that of parvalbumin is not.

It is instructive to examine the two structures that score most highly after ICBP. Figure 5 shows the alignment of the EF-hand hydrophobicity pattern with ICBP, T4 lysozyme, and catabolite gene activating protein (CAP). (This figure also illustrates how a hydrophobicity pattern is generated). As indicated in the figure, the segment of T4 lysozyme which was identified by the EF-hand hydrophobicity pattern shares the same secondary structure features as the EF-hand. Furthermore, these two helices are oriented with respect to each other in a way reminiscent of the helices in the EF-hand (26). Along with the CheY structure identification discussed above, this is a good example of the ability of our method to identify similar structural motifs in proteins that are not obviously related in sequence.

On the other hand, the structure identified in CAP is clearly unrelated to the EF-hand (there is no helix in the first 18 residues of the identified segment) even though the alignment score is similar to that in the other two proteins. Because there always is some chance of such spurious alignments, a high score cannot be considered proof of a correct structural identification. We also note that any structural model identified in this type of search must only be considered an approximate representation of the actual fold. For example, although each of the globin

structures scores quite highly with the globin hydrophobicity pattern, individual helix packings in these structures can differ by up to 3 Å in interaxial distance and up to 30° in interhelix angle (4).

Discussion

Ultimately, one would like to be able to use sequence information to predict protein structures. In this paper, we have taken the preliminary step of testing whether we can use sequence information to pick out the correct structure from a database of known structures. We find that the correlation between the hydrophobicity of amino acid side-chains and their solvent accessibility in folded proteins is sufficiently good that a sequence pattern can be used to directly identify a structure.

The hydrophobicity information in a set of amino acid sequences was summarized by a single string of hydrophobicity values (high, medium, or low). Similarly, the information from the three-dimensional structures of proteins was summarized by representing the structures as strings of solvent accessibility values (buried, partially buried or exposed). The results discussed in this paper show that these simple patterns can be surprisingly selective. The tertiary fold adopted by a set of sequences was correctly identified in several different test cases. Most notably, the flavodoxin structure was correctly identified as being most compatible with the hydrophobicity pattern of the CheY protein family, even though no significant sequence homology has been found.

It is striking that simple hydrophobicity patterns allow the identification of structural folds. This extends earlier observations that patterns of hydrophobicity can be a distinctive feature of the sequences of particular classes of protein. Sweet and Eisenberg have shown that the degree of correlation of hydrophobicity values between pairs of aligned sequences is a good criterion for the structural similarity of the two proteins (27). In a similar vein, Bashford et al. have created a sequence template that uses only residue hydrophobicity and size criteria to select globin sequences from a sequence database (3). Here, we have shown that a hydrophobicity pattern can be correlated directly with structural features of the appropriate protein fold(s).

Our results are consistent with the idea that burying hydrophobic residues is of paramount importance in determining the conformation and stability of proteins. A recent mutagenic study of hydrophobic core residues of λ repressor suggests that the most important feature of interior residues is their hydrophobic character, and that the precise identity of the hydrophobic residue (i.e., its size and shape) is less important (28). Similar conclusions have been reached from examination of homologous protein structures (4; 5). Apparently, protein structures are relatively tolerant of small changes in the shape and volume of amino acids in their interiors, but only in rare instances can they tolerate hydrophilic residues in the core (5). This intolerance of hydrophilic residues in solvent-inaccessible positions is an important factor in the effectiveness of our alignment procedure.

The requirement that solvent-inaccessible residues be hydrophobic necessarily imposes sequence constraints on any polypeptide that adopts a particular fold. Our results suggest that the converse also is true: A given pattern of amino acid hydrophobicity imposes constraints on the number and kinds of conformations which can be adopted by a particular amino acid sequence. Theoretical studies using strings of "polar" and "non-polar" elements in a simple lattice model support the same conclusion (29). Continued studies of how the pattern of polar and non-polar groups in an amino acid sequence can define a tertiary structure should help us to understand how protein sequence determines three-dimensional structure.

Fortran programs used in this work are available from the authors upon request.

Acknowledgements

We thank Jeff Stock for generously providing his aligned sequences for proteins in the CheY family and Will Gilbert and Upul Obeysekare for help with the computing. Program development was assisted by the University of Wisconsin Genetics Computer Group procedure library (30).

REFERENCES

1. Kauzmann, W. (1959) *Adv. Protein Chem.* **14,** 1-63.

2. Privalov, P. L. (1979) *Adv. Protein Chem.* **33,** 167-241.

3. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196,** 199-216.

4. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225-270.

5. Lesk, A. M. & Chothia, C. (1982) *J. Mol. Biol.* **160,** 325-342.

6. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669-678.

7. Taylor, W. R. (1986) *J. Mol. Biol.* **188,** 233-258.

8. Bowie, J. U. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 2152-2156.

9. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241,** 53-57.

10. Dickerson, R. E. (1980) *Sci. Amer.* **242,** 136-153.

11. Oochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S. & Morita, Y. (1983) *J. Mol. Biol.* **166,** 407-418.

12. Szebenyi, D. M. E., Obendorf, S. K. & Moffat, K. (1981) *Nature* **294,** 327-332.

13. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tsumi, M. (1977) *J. Mol. Biol.* **112,** 535-542.

14. Banner, D. W., Kokkinidis, M. & Tsernoglou, D. (1987) *J. Mol. Biol.* **196,** 657-675.

15. McKay, D. B., Weber, T. A. & Steitz, T. A. (1982) *J. Biol. Chem.* **257,** 9518-24.

16. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. (1981) *Nature* **290,** 754-758.

17. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. (1988) *Science* **242,** 899-907.

18. Jordan, S. R. & Pabo, C. O. (1988) *Science* **424,** 893-899.

19. Fauchere, J. -. & Pliska, V. (1983) *Eur. J. med. Chem.-Chim. ther.* **18,** 369-

20. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55,** 379-400.

21. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443-453.

22. Gibrat, J. -., Garnier, J. & Robson, B. (1987) *J. Mol. Biol.* **198,** 425-443.

23. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.

24. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985) *Science* **229,** 834-838.

25. Stock, A. M., Mottenen, J. M., Stock, J. B. & Schutt, C. E. (1989) *Nature* **337,** 745-748.

26. Tufty, R. M. & Kretsinger, R. H. (1975) *Science* **187,** 167-169.

27. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171,** 479-488.

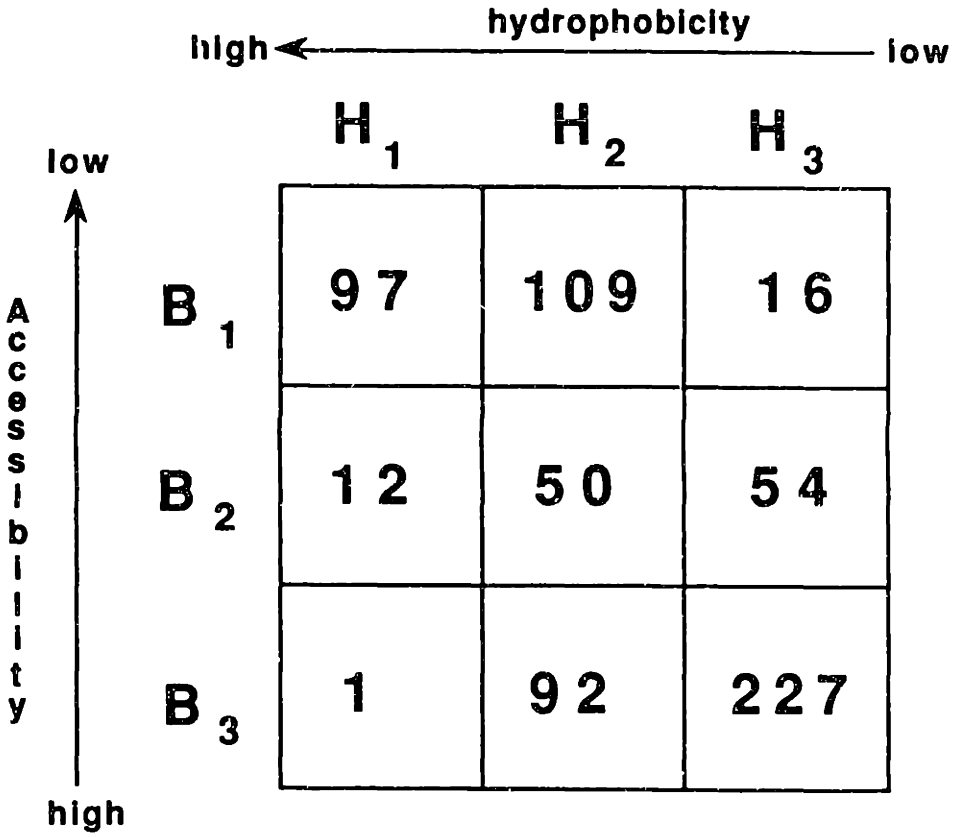
28. Lim, W. A. & Sauer, R. T. (1989) *Nature* in press.

29. Lau, K. F. & Dill, K. A. (1989) *Macromolecules* in press.

30. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387-395.

Figure 1. a) Occurrence of each hydrophobicity/accessibility pairing in the globin alignments. b) Scoring table (derived from the globin data) for evaluating alignments of hydrophobicity strings and solvent accessibility strings.

a)


hydrophobicity
high
low
H 1
H 2
H 3
low
Accessibility
high
B 1
B 2
B 3
97
109
16
12
50
54
1
92
227

b)

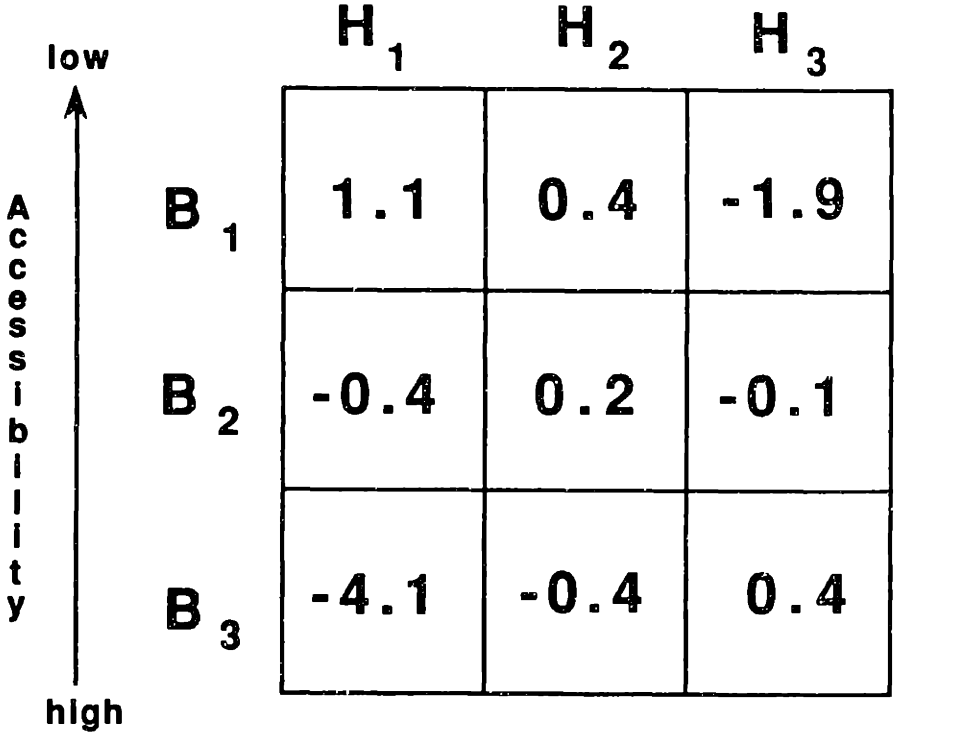
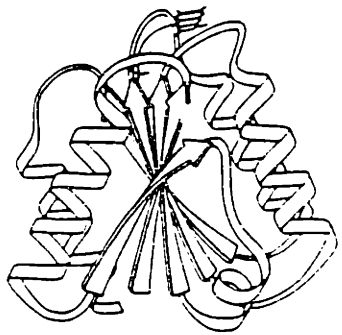
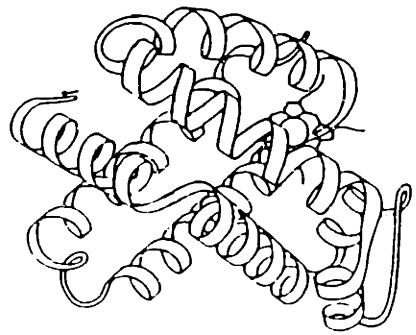
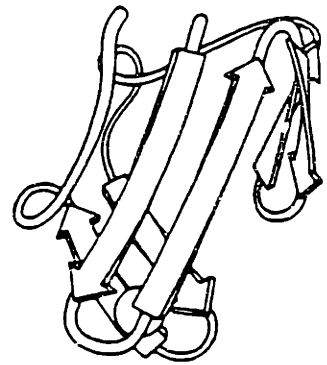

H 1
H 2
H 3
low
Accessibility
high
B 1
B 2
B 3
1.1
0.4
-1.9
-0.4
0.2
-0.1
-4.1
-0.4
0.4

Figure 2. Summary of pattern matching algorithm. Each protein in the structure database (represented here by a few of Jane Richardson's schematic drawings) is converted to a linear string representing the solvent accessibility of each residue in the folded protein. The arrows from the structures point to a part of the string derived for each of the structures. Shown at the bottom of the figure is a set of aligned protein sequences and part of the single hydrophobicity string generated from this set of sequences. This sequence string is then aligned as well as possible with each of the solvent accessibility strings. The structure which shows the highest quality alignment is circled. (The particular values shown here are for illustrative purposes only, and the position and length of the alignments will vary from one case to the next).







. . . $B_1 B_1 B_2 B_3 B_1 B_1 B_3 B_3 B_2$ . . .

. . . $H_2 H_1 H_2 H_1 H_3 H_3 H_1$ . . .

. . . $B_2 B_1 B_2 B_1 B_3 B_3 B_1 B_2$ . . .

. . . $H_2 H_1 H_2 H_1 H_3 H_3 H_1$ . . .

. . . $B_2 B_1 B_2 B_1 B_3 B_3 B_1 B_2 B_1$ . . .

. . . $H_2 H_1 H_2 H_1 H_3 H_3 H_1$ . . .

. . . $H_2 H_1 H_2 H_1 H_3 H_3 H_1$ . . .

. . . ALIGNED SEQUENCE 1 . . .

. . . ALIGNED SEQUENCE 2 . . .

. . . ALIGNED SEQUENCE 3 . . .

.

.

.

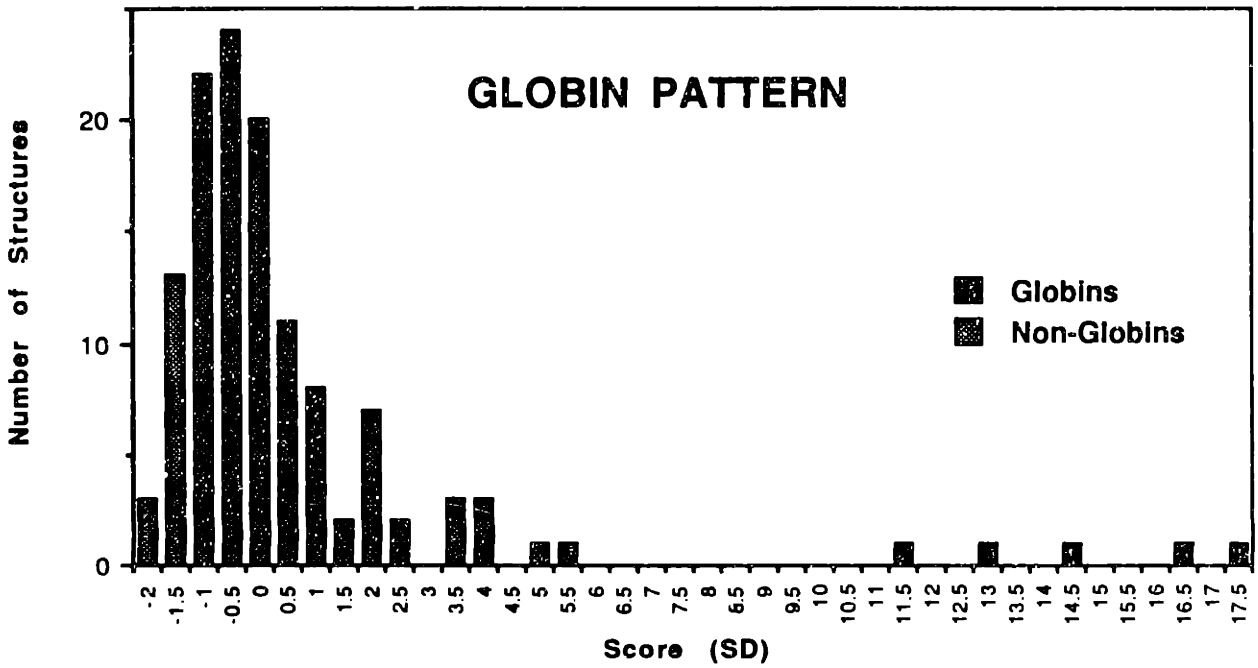
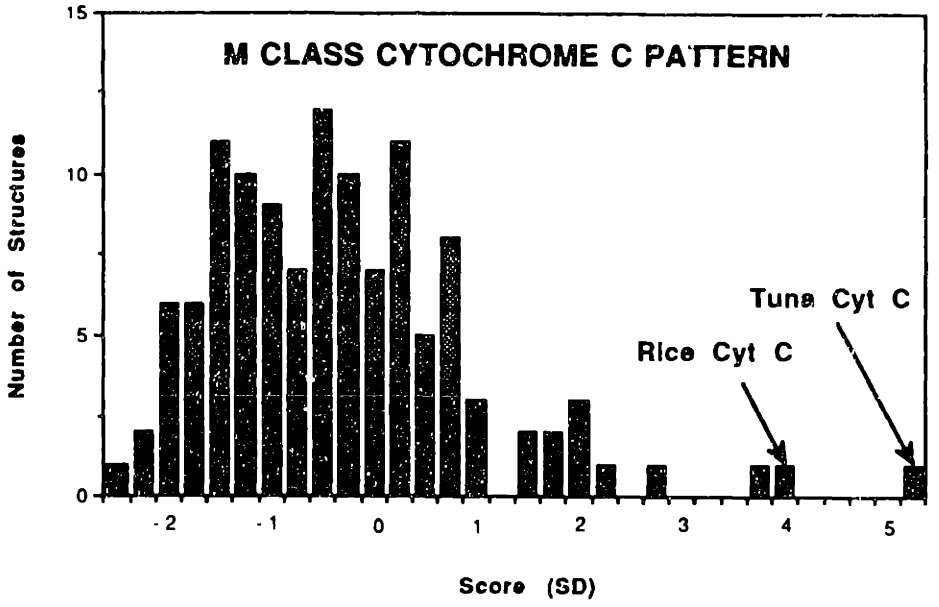


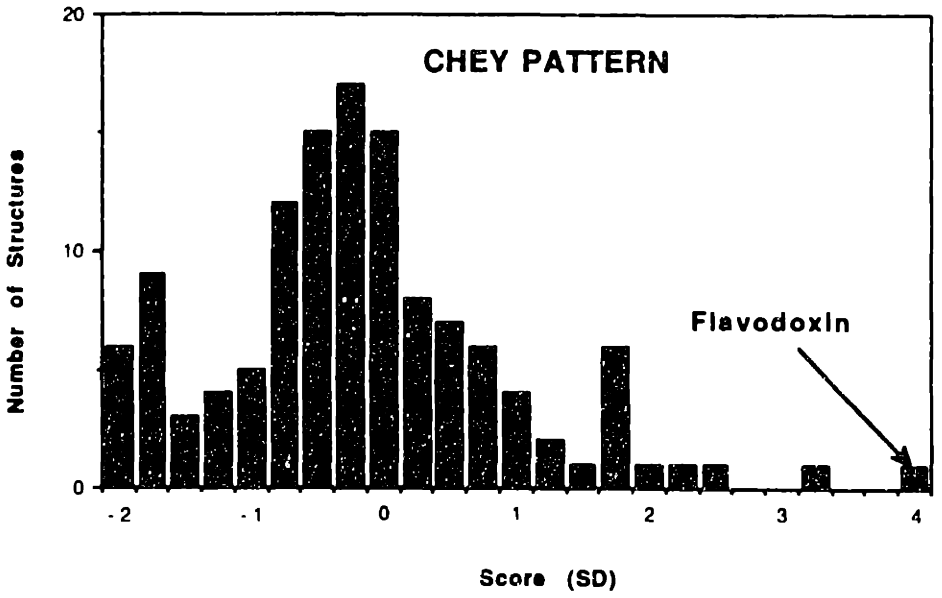
Figure 3. Alignment scores for the globin hydrophobicity pattern with accessibility patterns calculated from the structure database. Scores are given in standard deviation units and refer to the alignment quality of a particular accessibility pattern, relative to the mean quality of 50 random strings of the same composition. The top five scoring protein structures (in order) are myoglobin, the a subunit of human hemoglobin, leghemoglobin, erythrocrourin, and the b subunit of human hemoglobin.

Figure 4. Alignment scores (see the legend for figure 3) for the M class cytochromes c, the CheY protein family, and the EF-hand. The names, in order, of the top five scoring proteins for each histogram are A) tuna cytochrome c, rice cytochrome c, cytochrome B562, superoxide dismutase, flavodoxin; B) flavodoxin, T4 lysozyme, cytochrome B562, yeast phosphoglycerate kinase, chymotrypsinogen A; C) the EF-hand: vitamin D dependent calcium binding protein, T4 lysozyme, catabolite gene activating protein, 434 repressor, arabinose binding protein.

A)


M CLASS CYTOCHROME C PATTERN
Number of Structures
15
10
5
0
-2
-1
0
1
2
3
4
5
Tuna Cyt C
Rice Cyt C
Score (SD)

B)


CHEY PATTERN
Number of Structures
20
10
0
-2
-1
0
1
2
3
4
Flavodoxin
Score (SD)

C)

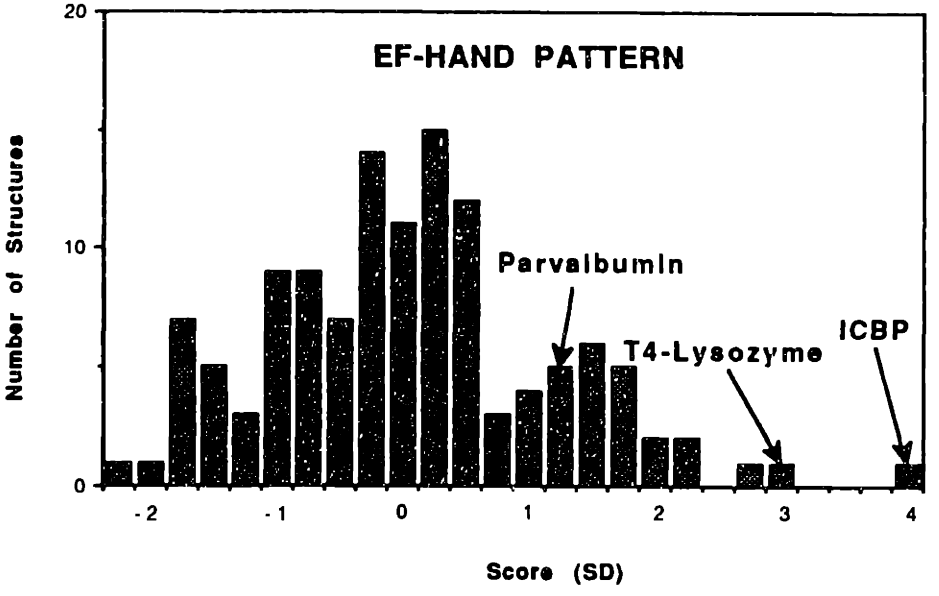
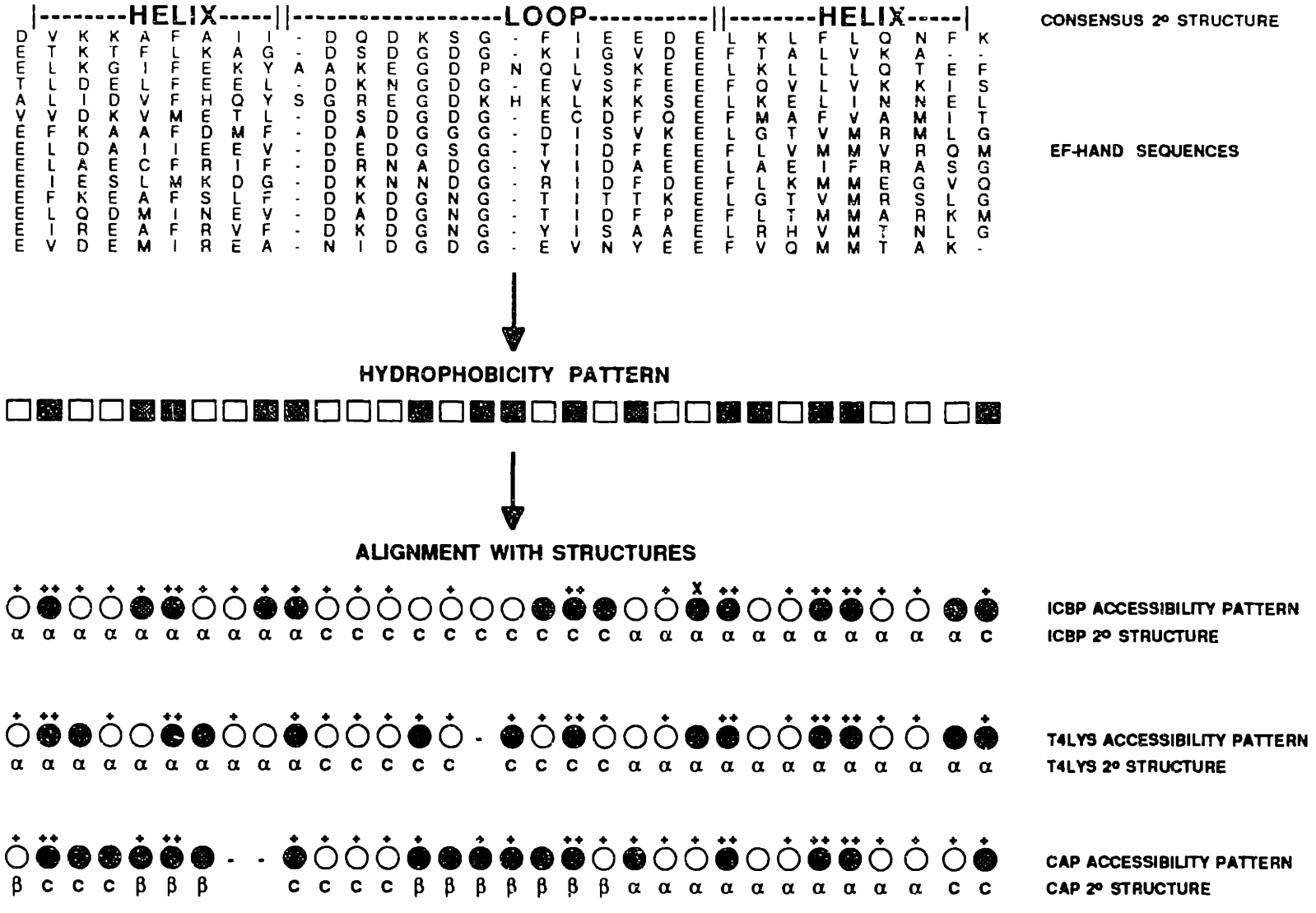

EF-HAND PATTERN
Number of Structures
20
10
0
-2
-1
0
1
2
3
4
Parvalbumin
T4-Lysozyme
ICBP
Score (SD)

Figure 5. Generation of the hydrophobicity pattern for EF-hand sequences and the alignments with the solvent accessibility patterns for the three proteins that gave the top scores. The amino acid sequences and consensus secondary structure are from Szebenyi et al. (12). The hydrophobicity pattern is shown below the sequences. Filled-in boxes correspond to members of the H_1 hydrophobicity group, shaded boxes to members of the H_2 group and open boxes to members of the H_3 group. The lower part of the figure shows the aligned solvent accessibility patterns of the proteins that gave the top three scores: vitamin D dependent calcium binding protein from bovine intestine [ICBP; residues 5 through 36], T4 lysozyme [T4-Lys; residues 41 through 71] and catabolite gene activating protein [CAP; residues 291 through 320]. Filled-in circles correspond to members of the B_1 group, shaded circles to members of the B_2 group and open circles to members of the B_3 group. Gaps in the alignment are indicated by dashes. Symbols above the accessibility patterns indicate the quality of the matches. ++ indicates a score of greater than 0.5, + a score between 0 and 0.5 and an X, a score less than 0.5. The secondary structure of the three proteins is shown below each accessibility pattern. α denotes an α helix, β a β-strand, and C a coil region.


HELIX
LOOP
HELIX
CONSENSUS 2° STRUCTURE
EF-HAND SEQUENCES
HYDROPHOBICITY PATTERN
ALIGNMENT WITH STRUCTURES
ICBP ACCESSIBILITY PATTERN
ICBP 2° STRUCTURE
T4LYS ACCESSIBILITY PATTERN
T4LYS 2° STRUCTURE
CAP ACCESSIBILITY PATTERN
CAP 2° STRUCTURE

CHAPTER 4

Identification of C-terminal Extensions That Protect Proteins from Intracellular Proteolysis

INTRODUCTION

Proteins that are thermally unstable are often subject to rapid proteolytic degradation in the cell. Intracellular turnover of these proteins can, however, be prevented in several ways. In some cases, degradation of unstable proteins can be suppressed by host mutations that affect the proteolytic machinery of the cell. For example, some rapidly degraded proteins are stabilized in *Escherichia coli* hosts that are *lon*$^-$, *htp*R$^-$, or *hfl*$^-$ (1-3). Degradation of specific proteins can also be prevented or slowed by intragenic mutations that increase the ratio of folded to unfolded protein in the cell. This can be accomplished in a direct fashion by sequence changes that increase the melting temperature of the protein or, in an indirect fashion, by changes that affect ligand binding reactions that are thermodynamically coupled to the protein folding reaction (4). In this paper, we describe a new class of intragenic mutations that suppress intracellular degradation and show that these mutations act by a mechanism that is independent of the thermodynamic stability of the target protein.

The subject of this paper is P22 Arc, a small dimeric DNA binding protein which acts as a transcriptional repressor of the bacteriophage P22 P*ant* promoter (for review, see ref. 5). Mutations at many of the 53 residue positions in Arc repressor lead to an Arc-defective phenotype and dramatically reduced intracellular levels of the mutant proteins (6). As part of a general study of Arc structure and function, we have isolated intragenic revertants of several different Arc-defective mutants. Surprisingly, we find that most revertants contain frameshift mutations

near the C-terminal portion of the coding sequence. These frameshift mutations result in the addition of extra C-terminal residues because of readthrough past the normal termination codon. We show that these C-terminal extensions stabilize Arc against intracellular degradation without affecting the oligomeric form, thermal stability, or DNA binding properties of the purified protein. We also show that addition of one of these C-terminal extensions to an unrelated protein prevents intracellular degradation of this heterologous protein.

MATERIALS AND METHODS

Strains and Arc Mutants: *Escherichia coli* strain UA2F is a derivative of strain US3 (*str*A, *thi*$^-$, *his*$^-$, *lac*Z$^-$, *lac*Y$^+$, *sup*^{0}, *rec*A$^-$) containing an F' episome (*lac*I^Q, *lac*Z::Tn5[*kan*^{R}] *pro*$^+$) and a λAC201 prophage which bears a P_{ant} promoter fusion to *cat* (6). X9T is a derivative of X90 (7) which was selected for resistance to T1 phage present in our laboratory fermentor.

The defective *arc* alleles reverted in these experiments, MI4, VG18 and DN20 were isolated previously (6,8). The mutations are named with the first letter referring to the wild-type amino acid and the second indicating the mutant amino acid. The number indicates the position in the amino acid sequence. Both MI4 and DN20 have been purified and studied biochemically (6). The structure and stability of MI4 is unaffected by the mutation, but DNA binding activity is significantly reduced. DN20, on the other hand, has only mildly reduced DNA binding activity but has lower stability. The VG18 mutant protein was too rapidly degraded to be purified suggesting that it has severely reduced stability (6).

Buffers and Media: The following buffers and cell growth media were used. Lysis buffer: 100 mM Tris-HCl [pH 8.0], 200 mM KCl, 1 mM EDTA, 2 mM $CaCl_2$, 10 mM $MgCl_2$; PCB: 50 mM Tris-HCl [pH 8.0], 0.1 mM EDTA, 5% glycerol, 1.4 mM 2-mercaptoethanol; CB: 10 mM Tris-HCl [pH 7.5], 50 mM KCl; TE: 10 mM Tris-HCl [pH 8.0], 1 mM EDTA; GBA: 10 mM Tris-HCl [pH 7.5], 3 mM $MgCl_2$, 0.1 mM EDTA, 100 mM KCl, 100 mg/ml bovine serum albumin and 0.02% Nonidet P-40. GBA2 is GBA buffer with 200 mM KCl. Recipes for LB medium and E medium can be found in references 9 and 10 respectively.

General Molecular Biology: Plasmids were constructed by established procedures (11). DNA sequencing was performed by the dideoxy method (12) following the protocols of Boehringer Mannheim Biochemicals. Single stranded plasmid DNA for sequencing was prepared as described (8). Oligonucleotides were synthesized on a Systec Microsyn 1450A. Cleavage of the DNA from the solid support and the removal of protecting groups was accomplished by suspending the support in 1 ml of ammonium hydroxide and heating at 55 °C overnight. The ammonium hydroxide supernatant was evaporated to dryness in a Savant Speed Vac Concentrator and the synthetic DNA was purified by electrophoresis on 20% polyacrylamide-urea denaturing gels. The band containing the oligonucleotide was excised and eluted into TE by diffusion. The DNA was then applied to a Sep-pack C18 cartridge (Waters) and after washing with TE, eluted with 1.5 ml of 70% acetonitrile/water. The solvent was removed by evaporation and the oligonucleotide was redissolved in TE.

Plasmid Construction: Plasmid pUS100 is a derivative of pBR322 which contains the *str*^{S} gene and an M13 origin of replication. This plasmid was constructed by replacing the small *Nde*I to *Pvu*I fragment of pNO1523 (13) with the corresponding fragment which contains the M13 origin of replication from plasmid pZ150 (14). The small *Hpa*I-*Nde*I fragment of pUS100, which contains the wild-type *str*^{S} promoter (*Hpa*I cuts after the fourth codon of the *str*^{S} gene; *Nde*I cuts within pBR322 sequences upstream of the promoter), was then replaced with a synthetic cassette which recreated the *str*^{S} coding sequence and ribosome binding site and introduced a unique *Bst*EII restriction site immediately upstream of the gene. The resulting plasmid, pUS200, no longer conferred streptomycin sensitivity to a resistant strain, confirming the loss of promoter activity. The small *Nde*I-*Bst*EII fragment of pUS200 was then replaced with a synthetic cassette that contained the wild-type *arc* operator, a P_{ant} promoter bearing a mild down mutation (T→A at position -7 of the promoter; ref. 15), and an upstream *Eco*RI site to yield plasmid pUS301. Insertion of the cassette destroyed the NdeI site. Wild-type or mutant *arc* genes were subsequently cloned into pUS301 by ligating the *Pvu*I-*Pvu*II fragment (containing the *arc* gene) from pTA200 or mutant derivatives (8) into the *Pvu*I to *Eco*RV backbone fragment of pUS301. A map of the resulting plasmid, pUS405, is shown in Figure 1. Plasmid pUS350 is a derivative of pUS301 in which the *Eco*RI sites were destroyed by the insertion of linkers; some *arc* genes were cloned into this vector to yield pUS406 using the strategy described for construction of pUS405.

The *lt*1 tail sequence was fused to the gene for the LP57 mutant of the N-terminal domain of λ repressor by inserting a synthetic DNA cassette

into the *Sph*I site of plasmid pRB104/LP57 (4). The resulting gene, LP57/*lt*1 contains λ repressor sequence up to amino acid 102 followed by the *lt*1 tail sequence, RKVE.....NQHE (see Figure 2 for the complete *lt*1 sequence).

Mutagenesis: Random primer mutagenesis was performed as described (8) except that mutagenesis was performed on single stranded DNA from plasmid pUS405. Hydroxylamine mutagenesis was performed on mutant derivatives of pTA200 as described by Nelson and Sauer (16). Hydroxylamine treated plasmids were digested with *Eco*RI and *Eco*RV and the *arc* containing fragment was cloned into the unmutagenized *Eco*RI to *Eco*RV backbone fragment of pUS406/MI4, a derivative of pUS406 which contains the MI4 Arc-defective allele.

Selection for Arc Activity: Our selection for Arc activity depends on Arc-mediated repression of a dominant str^S gene in the presence of a constitutively expressed but recessive str^R gene. For these studies, we use *E. coli* strain UA2F (6,8) containing plasmids pUS405 or pUS406. Strain UA2F is resistant to streptomycin, kanamycin, and chloramphenicol, and also carries a $lacI^Q$ allele. Plasmids pUS405 and 406 encode ampicillin resistance, contain mutant versions of the *arc* gene under control of the IPTG-inducible promoter P*tac*, and contain the Arc-repressible str^S gene. Transformation of strain UA2F with mutagenized plasmid DNA was performed as described (17) except that competent cells were prepared after growth in LB supplemented with 20 mM $MgSO_4$ and 25 mg/ml each of chloramphenicol, kanamycin and streptomycin. After the heat pulse, cells were outgrown in 5 ml LB plus 20 mM $MgSO_4$ for 1 h; harvested by centrifugation; and resuspended in 0.5 ml LB containing 20 mM $MgSO_4$

and 100 mg/ml ampicillin. Selections for Arc^+ transformants were then performed by plating 100 ml portions of cells on LB plates supplemented with 100 mg/ml ampicillin, 25 mg/ml kanamycin, 50 mg/ml streptomycin and 2 mg/ml IPTG. In strains UA2F/pUS405 or UA2F/pUS406, the streptomycin resistant phenotype is somewhat leaky. However, we have found that the selection is generally reliable as long as it is applied to freshly transformed cells. To ensure that candidates that passed the streptomycin selection were indeed Arc^+, an independent screen for Arc activity was also performed. The *cat* gene in UA2F is under control of the Arc-repressible promoter, P*ant* , and thus candidates that are Arc^+ should be sensitive to chloramphenicol. For this screen, candidates were picked and streaked on LB plates supplemented with 75 mg/ml chloramphenicol, 100 mg/ml ampicillin, 25 mg/ml kanamycin, and 2 mg/ml IPTG.

Degradation and Synthesis Rate assays: Rates of intracellular degradation of Arc and Arc variants were measured by pulse-chase experiments essentially as described (4). Overnight cultures of strains UA2F/pUS405 or UA2F/pUS406 were grown in E medium supplemented with 0.2% glucose, 0.01% of all amino acids except Met, Cys, Phe, Tyr, and Trp, 0.001% thiamine and 50 mg/ml ampicillin. Portions (0.1 ml) of the cultures were diluted to 5 ml in the same medium and grown at 37 °C until the A_{600} was approximately 0.5. The P*tac* promoter was then induced by the addition of IPTG (100 ml of a 10 mg/ml solution in water). After 15 min, 10 ml [^{35}S]-Methionine (>800 Ci/mmol, 15 mCi/ul) was added. One minute later, 200 ml of a 28 mg/ml solution of unlabeled methionine was added. At various times thereafter, 0.5 ml portions were removed, chilled on ice, and mixed with 35 ml of an ice-cold mixture of protease inhibitors consisting of 60 mM

phenylmethylsulfonyl fluoride, 30 mM N-ethylmaleimide and 80 mM sodium azide in 90 % ethanol. Cells were harvested by centrifugation, resuspended in 100 ml Laemmli sample buffer, and lysed by heating for 5 min at 90 °C. Following centrifugation of the lysed sample for 10 min in an Eppendorf centrifuge, 20 ml samples were electrophoresed on a 15% Laemmli gel. The gels were then treated with Autofluor (National Diagnostics), dried, and autoradiographed to visualize radiolabeled proteins. The Arc protein band is easily identified following autoradiography since it migrates in a position reasonably well removed from other major bands. Synthesis rate measurements were performed as described (4).

Protein Purification: Mutant Arc proteins were purified from *E. coli* strain X9T/pUS405 or X9T/pUS406 using several modifications of the previously described procedures (6,18). Unless indicated, all cell growth was at 37 °C, the initial steps of purification were performed at 4 °C, and column chromatography was performed at room temperature. Overnight cultures (5 ml) were grown in LB containing 100 mg/ml ampicillin and used to innoculate 2 liters of the same medium. This culture was grown to an A_{600} of approximately 1. At this time, expression of the *arc* gene from the P_{tac} promoter was induced by the addition of 0.2 g IPTG. Growth was then continued for 2.5 h, and cells were harvested and resuspended in cold lysis buffer (total volume = 20 ml)[1]. The cells were lysed by sonication; 0.2 ml polyethyleneimine (10% v/v) was added; the lysate was stirred for 30 min;

[1] In most cases, cells were harvested directly but in the purification of VG18/*st*1 and DN20/GD52, 20 ml of 100 mM N-ethylmaleimide, 100 mM phenylmethanesulfonyl fluoride and 0.1 g of sodium azide were added in order to inhibit proteolysis. It was not determined if the addition of protease inhibitors was required for the purification of these two proteins

and the precipitate was removed by centrifugation at 13,000 x g for 30 min. Ammonium sulfate (9 g) was added to the supernatant; the protein precipitate was collected by centrifugation at 13,000 x g for 30 min; and the pellet was dissolved in 25 ml PCB and dialyzed for 4 h against the same buffer. This material was filtered through a 0.45 mm filter and applied at a flow rate of 4 ml/min to a 1 x 10 cm Accell CM column (Waters) equilibrated in PCB containing 50 mM KCl. Most of the protein eluted in the flow through fraction. After the baseline stabilized, a linear gradient to 50% PCB plus 1 M KCl was run over 45 min. Depending on the allele, Arc eluted as a major peak from 30 to 40 minutes after starting the gradient. Fractions containing Arc were pooled, dialyzed against distilled water, and lyophilized. The dried protein was redissolved in 1 ml of CB and chromatographed at a flow rate of 0.5 ml/min on a 2.5 x 40 cm column of Sephadex G-75 (superfine) equilibrated in CB. Fractions containing Arc were pooled and stored at -20 °C.

Assays of Stability and DNA Binding: Denaturation of wild-type and mutant Arc proteins by guanidine hydrochloride (GuHCl) was monitored by changes in the the fluorescence emission intensity at 330 nm as described (6). These studies were performed at room temperature in a buffer containing 10 mM Tris-HCl [pH 7.5], 50 mM KCl, and Arc at an initial concentration of 16 mM. Thermal denaturation was assayed by circular dichroism by monitoring changes in the ellipticity at 222 nm as a function of temperature. These studies were performed in a buffer containing 10 mM potassium phosphate [pH 7.5], 100 mM KCl, and Arc at a concentration of 8 mM.

Operator and non-operator DNA binding by Arc or Arc variants was assayed by the gel mobility-shift assay described previously (6,19). These studies were performed at room temperature in GBA buffer. Plasmid pAO100 (20), which contains a synthetic arc operator insert in the *Cla*I site of pZ150 (13), was used as a source of both operator DNA (the *Eco*RI-*Eco*RV fragment) and non-operator DNA (the *Eco*RI-*Pst*I fragment). For non-operator DNA competition assays, operator binding affinity was measured in the presence or absence of 500 μg/ml of sonicated salmon sperm DNA using the gel mobility shift assay in GBA2 buffer. To compare non-operator dissociation constants measured in direct binding and competition assays, macroscopic dissociation constants were weighted by the appropriate statistical factors (21).

Molecular Weight Determinations: The elution positions of Arc and variant Arc proteins from a Sephadex G75 column were determined during the purification as described above. The column was calibrated by determining the elution positions of set of monomeric proteins (ovalbumin, myoglobin, carbonic anhydrase, cytochrome C) of known molecular weight. The exclusion and inclusion volumes were determined by the elution positions of thyroglobulin and vitamin B12.

For sedimentation velocity experiments, a mixture containing 120 mg of wild-type or mutant Arc and 100 mg each of bovine serum albumin, ovalbumin, carbonic anhydrase, and cytochrome C was prepared in 240 ml of 50 mM Tris-HCl [pH 7.5], 100 mM KCl. This was layered on top of a 10 ml glycerol gradient (15% to 40% in 50 mM Tris-HCl [pH 7.5], 100 mM KCl) and centrifuged in a Beckman SW40 rotor at 35,000 rpm for 66 hrs. The bottom

of the tube was punctured; 0.25 ml fractions were collected; and portions of each fraction were electrophoresed on 15% Laemmli gels. Proteins were detected by staining with Comassie blue and amounts were quantified by densitometry.

RESULTS

Strategy for Selection of Revertants: To isolate revertant sequence changes that restore the activity *in vivo* to Arc-defective mutants, we developed a positive selection for Arc activity. This selection takes advantage of the fact that a strain expressing both streptomycin sensitive and streptomycin resistant alleles of ribosomal protein S12 displays the sensitive phenotype (13). As described in the Methods section, the *str*^{S} gene in such a strain was brought under the control of an Arc repressible promoter. Thus, in the presence of active Arc, the *str*^{S} gene is repressed and the strain becomes streptomycin resistant.

Arc genes encoding the single mutations Met^{4}→Ile, Val^{18}→Gly, and Asp^{20}→Asn (MI4, VG18, and DN20, respectively) display an Arc-defective phenotype and fail to confer streptomycin resistance to the selection strain. Plasmids containing these genes were mutagenized and introduced into the selection strain by transformation. Streptomycin resistant colonies were then selected and tested for Arc activity using a screen for chloramphenicol sensitivity, as described in the Methods section. The *arc* genes from candidates that passed both the streptomycin selection and the chloramphenicol screen were sequenced.

Revertants Isolated: Second site revertants were isolated for each of the three primary mutations (Figure 2). One revertant, DN20/GD52, contained a simple missense mutation in addition to the primary mutation. The remaining revertants, surprisingly, contained frameshift mutations near the end of the gene in addition to the primary mutations. Four of the frameshift mutations were single base pair deletions at codons 50, 51 or 52; each of these resulted in extension of the protein reading frame by the same 25 codons. These -1 frameshifts were designated *lt*1, *lt*2, *lt*3, and *lt*4 (long tail). At the protein sequence level, these mutations differ only in the identity of Arc residues 51 and 52 (Figure 2). One frameshift mutation was a single base pair insertion mutant which resulted in a 8 residue extension of the protein reading frame. This +1 frameshift mutation was designated *st*1 (short tail). Because the -1 frameshifts encoded extremely similar protein sequences, only *lt*1 was investigated further. This mutation was introduced into a wild-type background and into the DN20 and VG18 mutant backgrounds. Each of these variants displayed an Arc^+ phenotype in the selection strain. The GD52 and *st*1 mutations were also transferred into wild-type backgrounds. These variants also conferred an Arc^+ phenotype to the selection strain.

Proteolysis Rate *In Vivo*: The intracellular half-lives of the primary mutants, and proteins bearing the *lt*1, *st*1 and GD52 mutations were determined by pulse-chase experiments. The results are shown in Table I. Each of the primary mutants is degraded more rapidly than wild-type Arc, with the VG18 mutant being the least stable. Addition of the *lt*1 mutation increases the intracellular stability of each of the mutants and also increases the stability of wild-type Arc. For example, as shown in the

pulse-chase experiment in Figure 3, the half-life of VG18 is increased from less than 5 minutes to greater than 2 hours by addition of the *lt*1 mutation. The GD52 mutation also causes a substantial stabilization of either wild-type Arc or the DN20 mutant; it increases the half-life of both proteins to approximately 2 hours. The *st*1 mutation has a smaller effect on intracellular turnover; it increases the half-life of VG18 to about 15 min but has little effect on the half-life of the wild-type protein. It is clear that each of the reverting mutations is able to significantly slow the intracellular degradation of at least one of the mutants. Moreover, we found that the revertants accumulate to higher steady-state levels than the primary mutants. These higher levels are solely a result of reduced degradation since we found in other experiments that the expression rates of the wild type, *lt*1, *st*1 and GD52 were identical (data not shown). These increased levels probably account for some, if not all, of the enhanced activity of the revertants *in vivo*.

Purification of the Arc Protein Variants: There are a number of mechanisms by which the *lt*1, *st*1, and GD52 reverting mutations might prolong the intracellular half-lives of the mutant proteins. To allow studies of the physical effects of the reverting mutations on the structure and properties of Arc, we purified the MI4/*lt*1, DN20/*lt*1, VG18/*lt*1, VG18/*st*1, and DN20/GD52 revertant proteins. We also purified proteins containing the *lt*1, *st*1, and GD52 mutations in otherwise wild-type backgrounds. All of these proteins were readily purified and each behaved in a manner similar to the wild-type Arc during purification. Purified proteins bearing the *st*1 and *lt*1 mutations migrated at the expected positions on Laemmli gels and amino acid analysis of the *lt*1 protein was consistent with a full length

product. In particular, the presence of the penultimate residue in the *lt*1 sequence could be easily verified since it is the only histidine in the sequence. The purified MI4 and DN20 mutants were available from previous studies (6). It is notable that the VG18/*lt*1 and VG18/*st*1 proteins could be purified easily, as we had been unable to purify the VG18 mutant by itself because of rapid degradation prior to and during purification.

Protein Stability and Structure: The equilibrium between the folded and unfolded forms of a protein can be a major determinant of its turnover rate (4). To observe the effect of the reverting mutations on Arc's folding equilibrium, the stability of the purified proteins to GuHCl denaturation was determined. Table I lists the concentration of GuHCl at which each of the proteins is half denatured. These results show that none of the reverting mutations alter the stabilities of the proteins to denaturation by GuHCl. For example, the DN20 mutant is half-denatured at 1 M GuHCl, as are the DN20/GD52, and DN20/*lt*1 revertants. In previous experiments, we found a good correlation between stabilities determined for Arc variants in thermal denaturation and GuHCl denaturation experiments (6). That is, Arc mutants that are more thermally stable than wild-type arc also more stable to GuHCl denaturation and *vice versa*. Thus, the reverting mutations are not likely to cause significant changes in the thermal stabilities of the proteins in which they reside. To verify this, we performed thermal denaturation experiments on wild-type Arc and Arc containing the *lt*1 mutation in an otherwise wild-type background. Figure 4a shows that the thermal denaturation profiles of these two proteins are almost identical. Clearly then, the reverting mutations do not increase the stability of Arc to chemical or thermal denaturation. These data indicate that

thermodynamic stabilization of the folded protein structure is not the mechanism by which these sequence changes increase the intracellular half-lives of the proteins in which they reside.

Figure 4b shows that the circular dichroism spectra of wild-type and the *lt*1 Arc proteins are extremely similar despite the fact that the *lt*1 protein contains 25 additional amino acids. Thus, there is no evidence that the additional 25 amino acids in the *lt*1 protein possess any regular secondary structure. These residues are presumably disordered. Moreover, we found that the NMR spectrum of *lt*1 shows a dispersion of aromatic resonances that is very similar to the wild-type protein (data not shown). This indicates that the additional C-terminal residues in the *lt*1 protein do not perturb the tertiary structure of the core Arc protein.

DNA Binding Properties: Unstable repressors can be protected from proteolysis *in vivo* by mutations that enhance binding to cellular DNA and thereby stabilize the folded form of the protein (4). To test if this mechanism operates in the case of the *arc* reverting mutations, we determined the operator and non-operator DNA binding activities of each of the purified proteins by a gel mobility shift assay. In addition, the non-operator DNA binding affinities of the wild type, *lt*1 and GD52 Arc proteins were also measured using a competition assay. The two assays gave similar results. For example, the dissociation constants for the wild type, *lt*1 and GD52 proteins binding to a single non-operator site were found to be 0.3 mM by the gel binding assay and 0.4 mM by the competition assay. The data presented in Table I show that the ability of the Arc proteins to bind operator or non-

operator DNA is generally unaffected by the reverting mutations. For example, the DNA binding properties of wild-type Arc are not changed by addition of the *lt*1 or *st*1 tail sequences or by the GD52 missense mutation. In one case, addition of the *lt*1 tail does appear to cause a small increase in the affinity of a mutant protein for operator DNA (*cf.*, MI4 and MI4/*lt*1) but this difference may reflect the fact that the revertant protein was freshly purified while the original mutant had been purified and kept frozen for over a year. Furthermore, the *lt*1 tail had no effect on the DNA binding properties in three other protein sequence backgrounds tested. The overall DNA binding data suggest strongly that none of the reverting mutations causes a substantial increase or decrease in the affinity of the protein for operator or non-operator DNA. As operator DNA binding requires a rather precise set of three-dimensional contacts, these findings also support our conclusion that the reverting mutations do not alter the tertiary structure of the core Arc protein.

Oligomeric Form of the Purified Proteins: Wild-type Arc is a dimer in solution (18). To determine if the reverting mutations alter the oligomeric form of Arc, we determined the apparent molecular weights of the purified proteins by gel-filtration chromatography (Figure 5). Proteins containing either the GD52 or *st*1 mutations eluted at the positions expected for dimers. Thus, it seems clear that these mutations do not change the oligomeric state of Arc. Proteins containing the *lt*1 mutation, by contrast, eluted at an apparent molecular weight (*ca.* 36 kd) higher than that expected for dimers (*ca.* 18 kd). This could indicate either that these proteins form tetramers or that they have an oblong shape resulting from the addition of 25 amino acids. The latter possibility is reasonable if, as suggested by the CD

spectrum, the *lt*1 tail residues are disordered since the tail may then be less compact than is typical for a globular protein. If this were true, the *lt*1 protein should migrate more slowly than expected in a sedimentation velocity centrifugation experiment, and this is indeed observed. Figure 6 shows that the *lt*1 protein sediments slightly slower than cytochrome C, a roughly globular protein with a molecular weight of 12.5 kd. Consequently, it is likely that the *lt*1 protein is a dimer with a relatively extended shape. It seems clear that none of the reverting mutations change the quaternary structure of Arc.

Effect of the *lt*1 Tail on Degradation of a Heterologous Protein: The data presented thus far show that the reverting mutations do not affect the thermodynamic stability, tertiary or quaternary structure, or DNA binding properties of Arc. This raises the possibility that these mutations act directly at the level of primary sequence, to slow proteolytic degradation. If this were true, then it might be possible to inhibit the turnover of a different protein by the addition of the *lt*1 tail to its C-terminal end. The LP57 mutant of the 1-102 N-terminal fragment of λ repressor has reduced thermal stability and a reduced half-life in the cell (4). We fused the *lt*1 tail to the carboxyl terminal end of the LP57 protein to produce the protein LP57/*lt*1 and assayed intracellular degradation by the pulse-chase experiment. As shown in Figure 7, addition of the *lt*1 tail increased the half-life of the LP57 mutant from approximately 1 hour to greater than four hours. Thus, the *lt*1 tail is able to inhibit proteolysis of a heterologous protein. The finding that the same set of C-terminal amino acids can inhibit the degradation of two different proteins suggests that the influence of these residues on

proteolytic susceptibility is unlikely to be a result of specific effects on the structure or activity of these proteins.

DISCUSSION

We have described the isolation and characterization of reverting mutations that restore activity *in vivo* to Arc defective mutants. Five of the six reverting mutations were frameshifts near the end of the gene which resulted in proteins extended by readthrough past the normal termination codon. Four of these frameshifts (*lt*1, *lt*2, *lt*3 and *lt*4) resulted in 25 residue extensions and one (*st*1) resulted in a 8 residue extension. Only one of the suppressors, GD52, was a simple missense mutant. All of the reverting mutations decrease the intracellular turnover rate of Arc repressor and the *lt*1 tail sequence was also found to inhibit proteolytic degradation of an different protein, the N-terminal domain of λ repressor.

Since numerous factors can alter proteolysis rates, we tested possible mechanisms by which the reverting mutations may act to inhibit intracellular degradation. Results presented in the companion paper suggest that mutations that alter the fraction of a protein in the folded form, can affect its degradation rate (4). Since structural and functional properties of a protein can change the fraction in the folded form, we looked for an effect of the reverting mutations on the structure, stability, or DNA binding activity of the wild-type or mutant Arc proteins. The *lt*1, *st*1, and GD52 mutations were purified in wild-type backgrounds and at least one mutant background. Despite the rather extensive protein sequence

changes, we could detect no effect of these mutations on the biochemical properties of the wild-type or mutant proteins. None of the reverting mutations altered the oligomeric form, stability, structure, or DNA binding properties of the purified proteins. Negative results are, of course, difficult to interpret since it is always possible that we might have seen effects under different conditions or that physical properties we did not test are altered. However, the *lt*1 tail was also found to significantly slow the degradation of a heterologous protein, a mutant form of the N-terminal domain of λ repressor. The finding that the same tail sequence can protect two different and unrelated proteins from proteolysis suggests strongly that the tail does not slow degradation by altering physical properties specific to the protein in which it resides.

We do not wish to suggest that the thermodynamic stability of the native protein plays no role in the rate at which Arc or its variants are degraded. On the contrary, stability appears to play a significant role in Arc turnover since Arc mutants such as VG18 and DN20, which are less thermally stable than wild-type Arc, are degraded more rapidly in the cell. This suggests that unfolded Arc is likely to be the primary target for proteolytic degradation. However, a simple increase in the fraction of Arc in the folded form does not appear to account for the protection against proteolysis conferred by the reverting mutations.

Another general way in which the reverting mutations could prevent degradation would be to cause aggregation of the protein to an insoluble and thus protease-resistant form. This is presumably the mechanism by which trpE-fusion proteins escape proteolysis since these proteins are recovered

from the cell in an insoluble form (22). However, there are several reasons to doubt that an aggregation mechanism is relevant here. First, the Arc proteins bearing the reverting mutations were freely soluble in crude lysates. Second, when purified, these proteins had the expected dimeric structure and showed no tendency to precipitate, even at concentrations of 1 mM or more. Third, aggregation would reduce the free Arc concentration and therefore decrease activity in the cell, but the reverting mutations actually enhance activity in the cell.

It is also possible that the reverting mutations create binding sites for another soluble protein which sterically blocks protease action. However, no other protein or proteins were observed to to co-purify with any of the revertant proteins. Moreover, if this mechanism were general then this putative factor would need to bind to Arc proteins bearing the *lt*1, *st*1, and GD52 sequence changes, and to the N-terminal domain of λ repressor bearing the *lt*1 tail. This seems unlikely, especially as the *lt*1, *st*1, and GD52 mutations have no obvious features in common.

How then do the reverting mutations prevent intracellular degradation? Since the sequence changes cause no obvious changes in the physical properties of the proteins which bear them, the primary sequence may be an important factor. We note that each of the reverting mutations affects the C-terminus of Arc. C-terminal sequence changes have also been found to alter the proteolysis rate of another protein, the N-terminal domain of λ repressor (4). These findings may provide a clue about the mechanism of intracellular degradation. It is possible that an important step in the normal degradation of Arc involves a recognition event at the C-terminus

which is blocked by the reverting mutations. Since essentially all well characterized proteases exhibit substrate specificity (even so-called non-specific proteases) it is reasonable to expect that the intracellular proteases of *E. coli* will also have site preferences (23). Consequently it is possible that the degradation of Arc, the N-terminal domain of λ repressor, and possibly other proteins is mediated by a protease which prefers particular residues at the C-terminus. Although no proteases having this property have yet been characterized from *E. coli*, many have been identified from other sources. For example, carboxypeptidases digest progressively from the C-terminus and a few endoproteinases have been characterized whose activities depend upon the nature of C-terminal residues (24). We do not know which residues of the C-terminal extensions are important for inhibiting proteolysis, although we do note that all the revertants introduce charged side chains within one or two residues of C-terminus of the protein. The importance of these charged residues has yet to be determined, but it should now be possible to use directed mutagenesis to identify the sequence determinants within the C-terminal extensions that confer protection against proteolysis.

The finding that certain sequences at the C-terminus of a protein can enhance stability *in vivo* may aid in the design of vectors for the high level production of desired proteins in *E. coli*. For example, C-terminal fusions of the *lt*1 tail or similar sequences might stabilize and allow purification of unstable or marginally stable proteins in soluble form. On the other hand, once we understand the sequence characteristics that inhibit proteolysis in greater detail, it may be possible to improve the intracellular stability of

proteins dramatically by simply altering one or a few residues at the C-terminal end of the protein.

ACKNOWLEDGEMENTS

I would like to thank Dawn Parsell for performing the expression rate experiments and providing the plasmid pRB104/LP57, Drew Vershon for help in the purification of the defective Arc mutants, Dinshaw Patel and Mike Zagorsky for the NMR spectra, Peter Kim for the use of the CD spectropolarimeter and Har Gobind Khorana for the use of the Laser Densitometer. Discussions with Dawn Parsell and Andrew Pakula were invaluable to the progress of this work.

REFERENCES

1. Gottesman, S., and Zipser, D. (1979) *J. Bacteriol.*, **133**, 844-851

2. Baker, T. A., Grossman, A. D., and Gross, C. A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 6779-6783

3. Hoyt, M.A., Knight, D. M., Das, A., Miller, H.I., and Echols, H. (1982) *Cell*, **31**, 565-573

4. Parsell, D. A., and Sauer, R. T. (1988) *J. Biol. Chem.*, submitted as companion paper

5. Susskind, M. M., and Youderian, P. (1983) in *Lambda II* (Hendrix, R., Roberts, J., Stahl, F., Weisberg, R., eds), Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 347-363

6. Vershon, A. K., Bowie, J. U., Karplus, T. M., and Sauer, R. T. (1986) *Proteins*, **1**, 302-311

7. Amman, E., Brosius, J., and Ptashne, M. (1983) *Gene*, **25**, 167-178

8. Vershon, A. K., Blackmer, K., and Sauer, R. T. (1986) in Protein Engineering (Inouye, M., and Sarma, R., eds), Acedemic Press, Orlando, Florida, pp. 243-256

9 Miller, J. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

10. Davis, R., Botstein, D., and Roth, J., (1980) *Advanced Bacterial Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

11. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

12. Sanger, F., Nicklen, S., and Coulson, A. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467

13. Dean, D. (1981) *Gene*, **15**, 99-102

14. Zagursky, R., and Berman, M. (1984) *Gene*, **27**, 183-191

15. Youderian, P., Bouvier, S., and Susskind, M. (1982) *Cell*, **30**, 843-853

16. Nelson, H. C. M., and Sauer, R. T. (1985) *Cell*, **42**, 549-558

17. Hanahan, D. (1983) *J. Mol. Biol.*, **166**, 557-580

18. Vershon, A. K., Youderian, P., Susskind, M. M., and Sauer R. T. (1985) *J. Biol. Chem.*, **260**, 12124-12129

19. Vershon, A. K., Liao, S., McClure, W., and Sauer, R. T. (1987) *J. Mol. Biol.*, **195**, 323-331

20. Vershon, A. K. (1986) Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA

21. Woodbury, C.P Jr., and von Hippel, P.H. (1983) *Biochemistry*, **22**, 4730-4737

22 Spindler, K. R., Rosser, D. S., and Berk, A. J. (1984) *J. Virol.*, **49**, 132-141

23 Perlmann, G. E., and Lorand, L., eds. (1970), Methods in Enzymology, **19**, Academic Press, New York, New York

24. Hersh, L., and Morihara, K. (1986) *J. Biol. Chem.*, **261**, 6433-6437

Table I. Properties of Arc and Arc variant proteins. $t_{1/2}$ is the half-life of each protein in strain X9T. $[GuHCl]_{1/2}$ is the concentration of GuHCl at which the protein is half denatured at 25 °C, pH 7.5. DNA binding affinities determined by the gel mobility-shift assay are shown relative to the wild-type affinities (smaller numbers indicate stronger binding). Under the conditions used, wild type Arc shows half-maximal operator binding at a protein concentration of 0.4 nM and and half-maximal non-operator DNA binding at 0.3 mM. The macroscopic non-operator dissociation constant measured by the gel binding assay is dependent on the length of the non-specific fragment and is related to the dissociation constant for binding to a single non-operator site as described (21).

Protein	$t_{1/2}$(min)	$[GuHCl]_{1/2}$(M)	Relative DNA Affinity	
			Operator	Non-operator
Wild Type	30	1.2	1	1
*st*1	45	1.2	1	1
GD52	60	1.2	1	1
*lt*1	>120	1.2	1	1
VG18	< 5	-	-	-
VG18/*st*1	15	0.6	27	3
VG18/*lt*1	>120	0.5	27	3
DN20	5	1.0	3	3
DN20/GD52	60	1.0	3	3
DN20/*lt*1	>120	1.0	3	3
MI4	15	1.2	27	3
MI4/*lt*1	>120	1.2	9	3

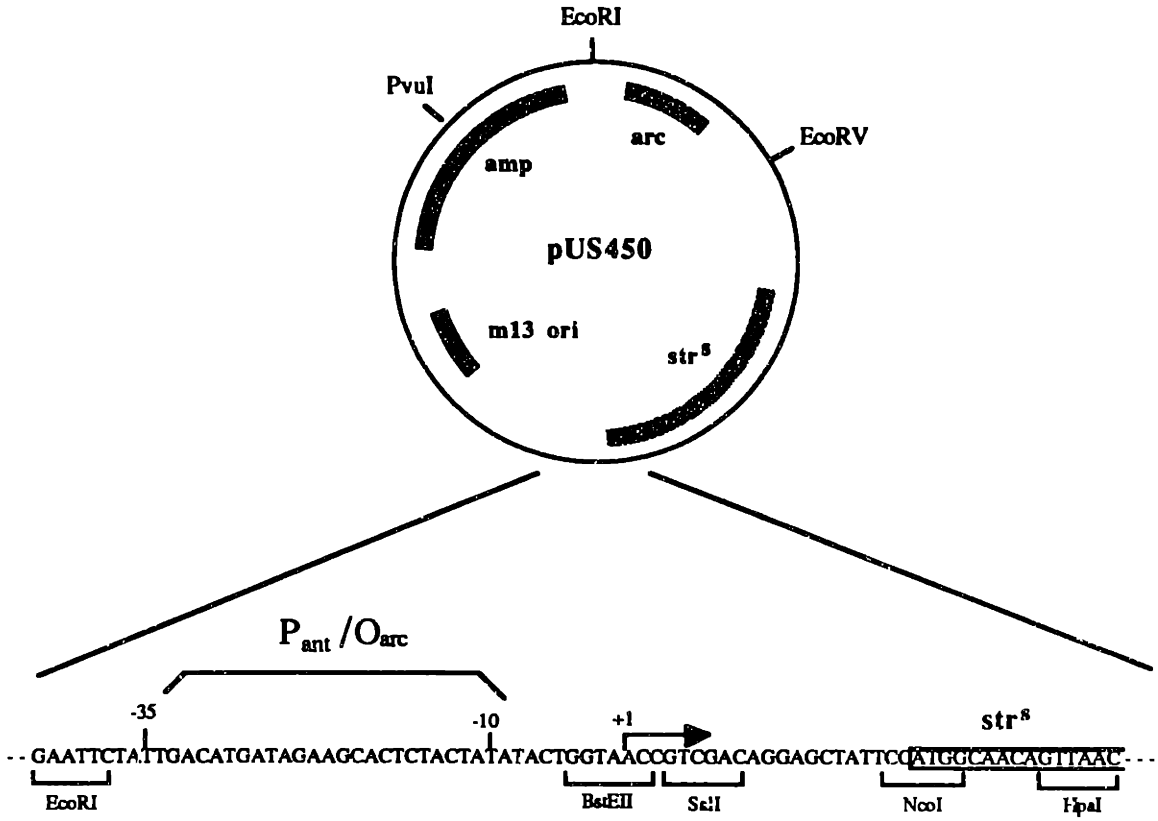


Figure 1. Map of plasmid pUS450. The DNA sequence of the promoter/operator region controlling the *str*^{S} gene is shown below. The *arc* operator is contained within the P_{ant} promoter (19).

A.

REVERTANT SEQUENCE CHANGES

	Gly^{49}	Arg^{50}	Ile^{51}	Gly^{52}	Ala^{53}	stop
	GGG	CGC	ATT	GGC	GCG	TAA ...
GD52	---	---	---	-A-	---	---
*st*1	---	---	---	GAC-	---	---
*lt*1	---	---	---	---	--▲	---
*lt*2	---	---	---	---	▲--	---
*lt*3	---	---	-▲-	---	---	---
*lt*4	--A	--▲	---	---	---	---

B.

C-TERMINAL PROTEIN SEQUENCES

Wild Type	...G R I G A^{53}
GD52	...G R I D A^{53}
*st*1	...G R I D R V K L K P Q L R^{61}
*lt*1	...G R I G R K V E A P T A V T V R A S V V S K S L E K N Q H E^{78}
*lt*2	...G R I G "^{78}
*lt*3	...G R M A "^{78}
*lt*4	...G R L A "^{78}

Figure 2. Revertant sequence changes. (A) DNA sequence changes in the revertants. Dashes in the sequences indicate no change, triangles indicate deletion of the base pair, and letters indicate particular base changes. The *st*1 sequence contains a G base insertion between codons 51 and 52. (B) C-terminal protein sequences resulting from the revertant sequence changes are shown in single letter code. The deletion and insertion mutations cause translation past the normal termination codon as a result of the shifted reading frame. The four deletion mutations result in identical amino acid sequence extensions past residue 52.

Figure 3. Pulse-chase analysis of the degradation of the VG18 mutant and the VG18/lt1 revertant proteins in *E. coli* strain X9T. The position of each protein is indicated in the figure. The VG18 protein is rapidly degraded with a half-life of less than 5 min, while the VG18/lt1 protein is not substantially degraded after 120 min. The extracts were electrophoresed on separate gels which were run to different extents.

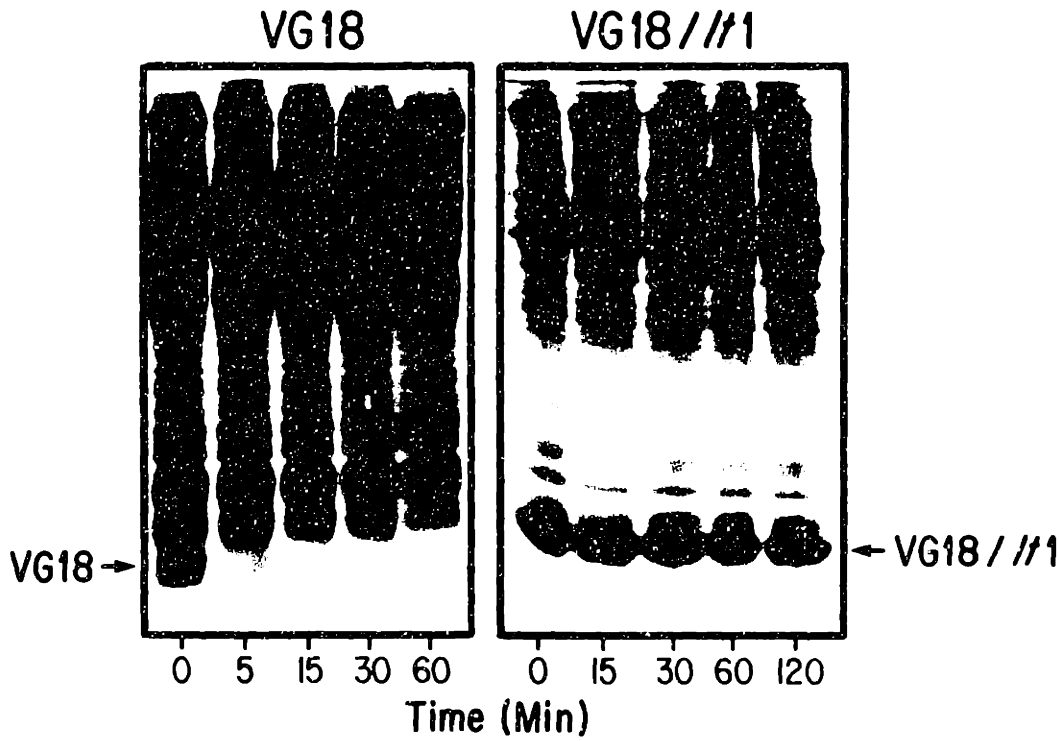
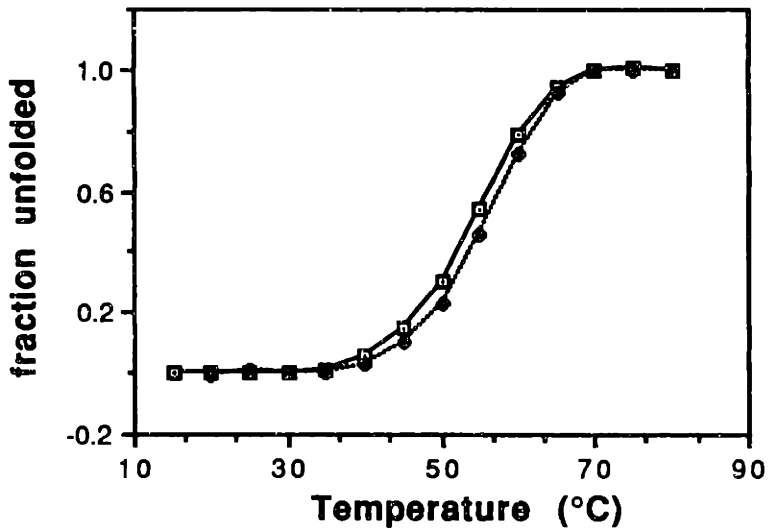
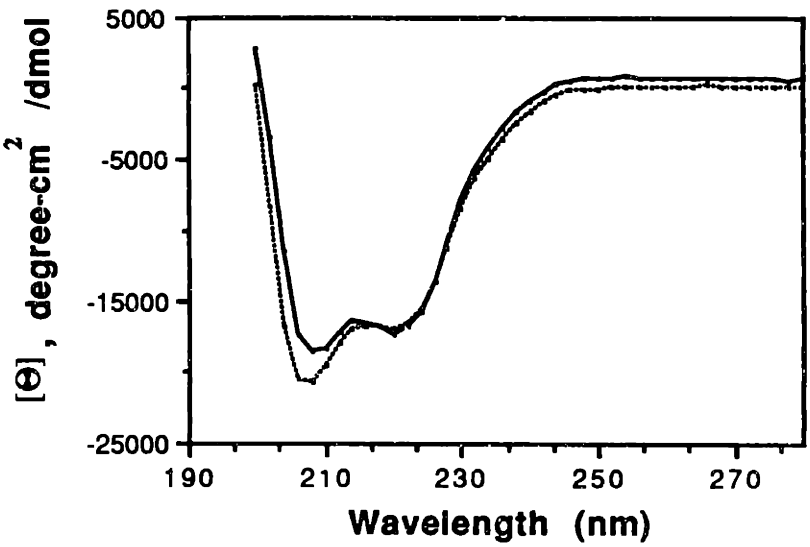

VG18
VG18//1
VG18
VG18//1
0 5 15 30 60
0 15 30 60 120
Time (Min)

Figure 4. Comparison of the structure and stability of the Arc and *lt*1 proteins. (A) Thermal denaturation profiles for wild-type Arc (solid line) and the *lt*1 protein (dashed line) at pH 7.5. F_{app} is the apparent fraction of the protein which is unfolded at each temperature. (B) Circular dichroism spectra of wild-type Arc (solid line) and *lt*1 protein (dashed line) at 25 °C, pH 7.5.

A)


fraction unfolded
1.0
0.6
0.2
-0.2
10
30
50
70
90
Temperature (°C)

B)


[Θ], degree-cm 2 /dmol
5000
-5000
-15000
-25000
190
210
230
250
270
Wavelength (nm)

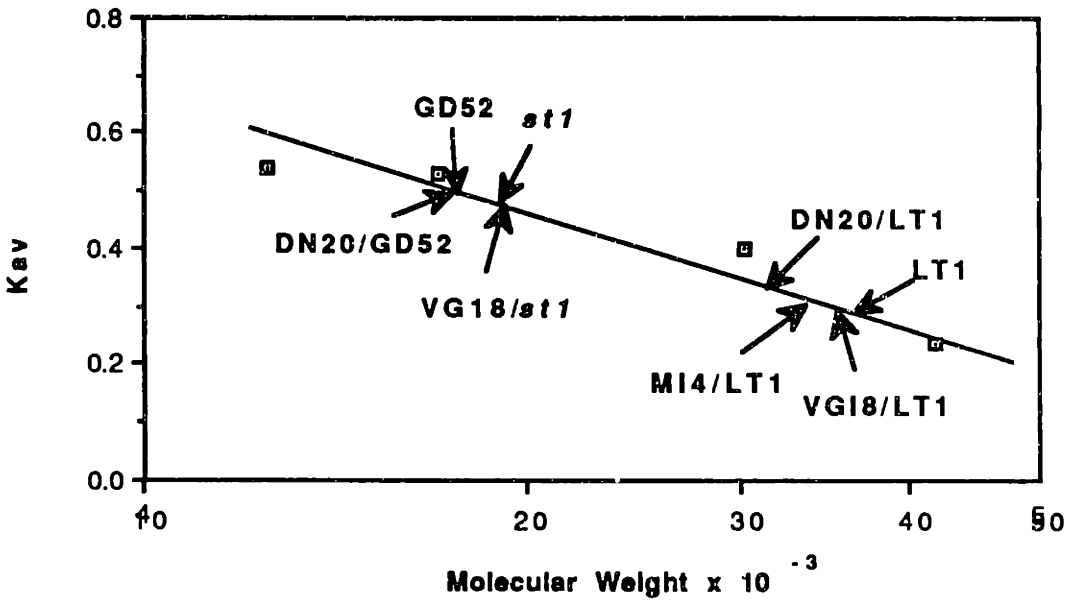


Figure 5. Elution position of variant Arc proteins on Sephadex G75. Kav is $(V_e-V_o)/(V_I-V_o)$, where Ve is the elution volume of the protein, Vo is the void volume and V_I is the inclusion volume. The column was calibrated as described in the text.

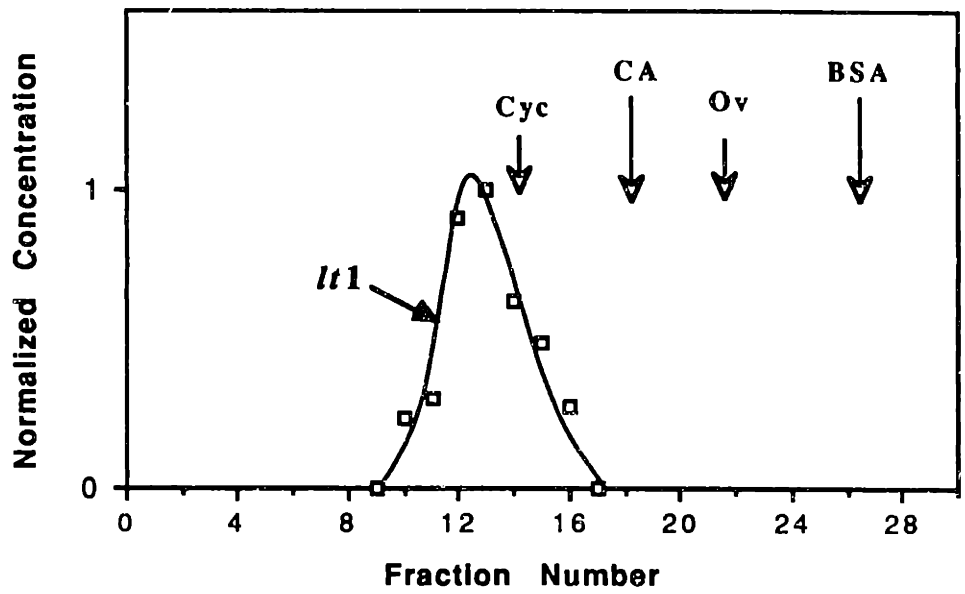
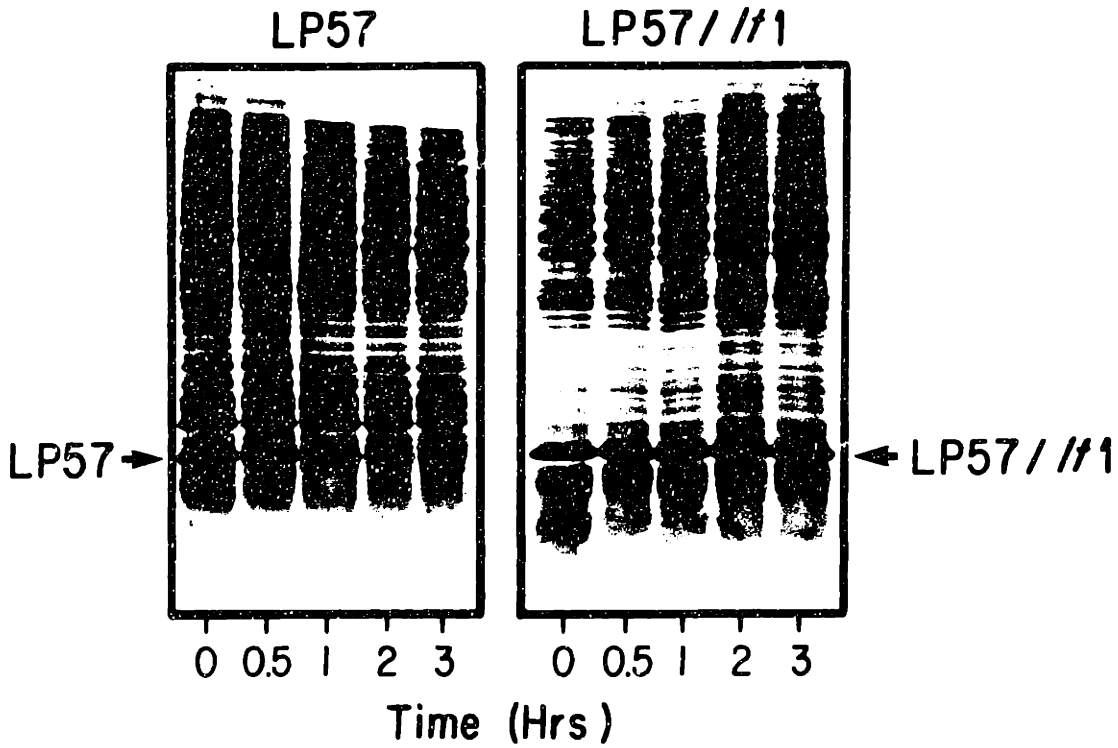


Figure 6. Sedimentation of the *lt*1 protein in a glycerol density gradient. Protein concentrations in each fraction were normalized to the concentration in the peak fraction. The peak fractions of several standard proteins are also shown: cytochrome C (Cyc), carbonic anhydrase (CA), ovalbumin (Ov), and bovine serum albumin (BSA). The wild-type Arc protein also migrates slightly slower than cytochrome C (not shown).

Figure 7. Intracellular turnover rate of the LP57 mutant in the N-terminal domain of λ repressor and the LP57/*lt*1 protein. The position of the band corresponding to each protein is indicated. The half-life of LP57 is about 1 h while that of the LP57/lt1 protein is greater than 3 h. The extracts were electrophoresed on separate gels, run to different extents.


LP57
LP57/ /f1
LP57
LP57/ /f1
0 0.5 1 2 3
0 0.5 1 2 3
Time (Hrs)

CHAPTER 5

Equilibrium Dissociation and Unfolding of the Arc Repressor Dimer

INTRODUCTION

The bacteriophage P22 Arc repressor is a small protein (53 residues) that binds to DNA in a sequence specific manner (1). Because the *arc* gene is small and selections for and against activity have been designed (2, 3), the Arc system is well suited to genetic analysis. For example, the phenotypic effects of over 200 different missense mutations have been determined and, as a set, these mutations affect every residue of Arc (3, 4). The Arc system is also amenable to structural studies. Crystallographic work is in progress (5) and the secondary structure of the protein has been determined by NMR studies (6). Residues 6-14 are part of a β-sheet region while residues 16-28 and 35-47 form amphipathic α-helices. Finally, Arc is of interest from the standpoint of DNA binding. Genetic and biochemical experiments make it clear that residues in the β-sheet region are responsible for the specificity of operator DNA binding (1, 3, 4, 7), indicating that Arc uses a novel structural motif for DNA recognition.

Arc is tetrameric when bound to operator DNA (B. Brown, unpublished) but dimeric or monomeric in solution, depending upon conditions (8). In this paper, we examine the dissociation and unfolding reactions of the Arc dimer. We show that these reactions are closely coupled, so that denaturation can be described in terms of the equilibrium populations of folded dimers and unfolded monomers. This two-state behavior suggests that folded monomers are thermodynamically unstable. Our results also suggest that the operator binding reaction *in vitro* proceeds, at least in part, from unfolded monomers to DNA bound

tetramers. Thus, mutations which affect protein stability will also affect DNA binding because the folding and dimerization reactions are an integral part of the overall DNA binding reaction.

MATERIALS AND METHODS

Protein and Reagents: Arc was purified as described previously (3, 8), and stored frozen in a buffer containing 10 mM Tris-HCl [pH 7.5], 50 mM KCl and 0.1 mM EDTA. Protein concentrations were determined using an extinction coefficient of 7800 $M^{-1}cm^{-1}$. An 8 M stock solution of guanidine hydrochloride (GuHCl) was obtained from Pierce Chemical Company. Stock 8 M urea solutions were prepared freshly each day using high purity grade urea obtained from United States Biochemical Corporation, and distilled, deionized water.

Urea and GuHCl Unfolding Monitored by Fluorescence: For denaturation experiments, protein samples were incubated for at least one hour at each GuHCl or urea concentration prior to taking fluorescence measurements. For urea denaturation experiments at different temperatures, solutions were incubated for at least 30 min in a water bath at the appropriate temperature (maintained to +/- 1°C) prior to fluorescence intensity measurements. This 30 min time period was more than sufficient to achieve equilibrium since no spectral changes were observed in the samples after a 5 min incubation at each temperature. In fact, folding and refolding of Arc was generally found to be complete within seconds under a

variety of conditions. Urea or GuHCl denaturation curves obtained starting with fully folded protein were identical to renaturation curves starting with fully unfolded protein, indicating that these reactions are fully reversible.

Fluorescence measurements were made using a Perkin-Elmer MPF-3 spectrofluorimeter in the early phases of this work, and a Greg PC spectrofluorimeter for the majority of the studies. For collection of emission spectra, the samples were excited at 280 nm and the emission intensity was adjusted to about 80,000 photons detected per second at λ_{max}. Spectra were recorded in 1 nm wavelength increments and the signal acquired for 5 sec at each wavelength. The spectrum of each sample was corrected by subtraction of the buffer alone. For urea and guanidine denaturation experiments, the samples were excited at 280 nm and the emission intensity recorded at 327 nm. The emission intensity of the protein solution in the absence of denaturant was adjusted to about 80,000 photons per second and the emission intensity of each sample was acquired for 15 sec.

Thermal Denaturation Monitored by Circular Dichroism: Circular dichroism measurements were performed using an AVIV model 60DS spectropolarimeter. Thermal denaturation was monitored by changes in circular dichroism at 222 nm in 2.5 °C steps. The samples were equilibrated for 2.5 min at each temperature and the signal recorded for 1 min. This rate of heating was found adequate to achieve equilibrium at each temperature. The melts were found to be more reversible at low pH than at higher pH. At pH 4, the ellipticity at 222 nm returned to better than

95% of its starting value while at pH 7.3, the ellipticity returned to 80 to 90% of its starting value. When melts were held at the T_m for extended periods of time at either pH, the ellipticity at 222 nm remained constant. Hence, the partial irreversibility of the unfolding transitions appears to result from protein damage that occurs at high temperature. For spectra, multiple scans (10 or more) were obtained for each sample and averaged. Each spectrum was recorded in 1 nm wavelength increments and the signal acquired for 1 s at each wavelength.

RESULTS

Probes of the Folding Equilibrium: Two spectroscopic methods can be used to monitor the unfolding reaction of Arc (2, 3, 4, 8). Figure 1a shows the circular dichroism spectra of native Arc and heat denatured protein. The peak of negative ellipticity at 222 nm in the native spectrum is characteristic of a high degree of α-helical character, and changes in this signal provide one method of following the unfolding reaction. Figure 1b shows the fluorescence emission spectra of native Arc and the protein denatured in urea. Arc contains a single tryptophan residue at position 14. The emission spectrum of the unfolded protein has a λ_{max} at 347 nm which is similar to that of tryptophan as a free amino acid. In contrast, λ_{max} of the folded protein is significantly blue shifted to 327 nm and this shift is accompanied by a large increase in intensity. This suggests that the tryptophan is buried in a hydrophobic environment in the folded protein (9). The fluorescence intensity change at 327 nm provides a second probe of Arc unfolding.

Denaturation is two-state: At the concentrations used in this work, Arc is dimeric under physiological conditions (8). Thus, the overall unfolding reaction must start with the folded dimer (A_2) and end with two unfolded monomers (2U). There are, however, several possible descriptions of the unfolding reaction, depending on the relative stabilities of dimers and folded monomers (A). If both the folded monomer and folded dimer are significantly populated states in the denaturation transition zones, the overall unfolding reaction will be described by:

$$A_2 \overset{K_1}{\rightleftharpoons} 2A \overset{K_2}{\rightleftharpoons} 2U \qquad (1)$$

where

$$K_1 = \frac{[A]^2}{[A_2]} \quad \text{and} \quad K_2 = \frac{[U]}{[A]}$$

In this case, one might see biphasic denaturation curves or non-superimposible transitions if the spectral probes used are differentially sensitive to the various species. As shown in Figure 2, however, this is not the case. The unfolding reaction follows a single and coincident transition, when followed either by changes in fluorescence or circular dichroism. This result suggests that either both CD and fluorescence respond in equivalent manners to the various equilibrium species or that only two states are significantly populated in the denaturation transition zone. The latter possibility is discussed below.

If the Arc dimer dissociates prior to the transition zone of the denaturation curves, only the equilibrium between folded and unfolded monomer would be observed. In this case, however, the observed reaction would be unimolecular and the equilibrium populations of folded and unfolded protein would be independent of protein concentration. This possibility is ruled out by the experiments of Figures 3a and 3b, which show that the stability of Arc increases with higher concentrations. The dimer state must, therefore, be significantly populated in the transition zone.

If the folded monomer state is essentially unpopulated, unfolding could then be described as a concerted reaction

$$A_2 \overset{K_u}{\rightleftharpoons} 2U \qquad (2)$$

where

$$K_u = \frac{[U]^2}{[A_2]} = 2\,(P_t)\,\frac{f_u^{\,2}}{1-f_u} \qquad (3)$$

with P_t being the total protein concentration and f_u the fraction of unfolded protein. If this model provides a reasonable thermodynamic description of the denaturation reaction, then one should calculate the same value of K_u or ΔG_u (calculated as -RT ln [Ku]) from experiments performed at different protein concentrations.

Figure 4 shows unfolding free energies, ΔG_u, calculated for different protein concentrations as a function of urea or GuHCl concentration. The free energies calculated at different protein concentrations assuming the two-state model are the same to within experimental error. Thus, the data are consistent with the two state assumption (equation 3) in the transition zone of the urea and GuHCl denaturation curves. The data points for both denaturants were least squares fit to the equation:

$$\Delta G_u = \Delta G_u{}^{H_2O} + m\,[\text{denaturant}] \qquad (4)$$

where $\Delta G_u{}^{H_2O}$ is the extrapolated unfolding free energy in the absence of denaturant. Both GuHCl and urea denaturation gave $\Delta G_u{}^{H_2O}$ values of about 11 kcal/mole ($K_u = 10^{-8}$ M). Thermal denaturation is also well approximated by the two-state model. The square symbols in Figure 5 show ΔG_u values calculated for the two thermal denaturation experiments shown in Figure 3b using the two state assumption.

Temperature dependence of the unfolding free energy: The values of the thermodynamic parameters ΔG_u, ΔH_u, and ΔS_u for denaturation vary as a function of temperature depending upon the value of ΔC_p, the constant pressure heat capacity change for the reaction (10). To estimate the value of ΔC_p for Arc denaturation, we have followed the recent suggestion of Pace and Laurents (11) and used urea denaturation curves to determine unfolding free energies at temperatures well below the thermal transition

zone. Urea denaturation curves were determined at 5°C, 15°C, and 25°C and analyzed as discussed above to determine $\Delta G_u^{H_2O}$ values at each temperature. These values are plotted in Figure 5. Assuming that ΔC_p is approximately constant over the temperature range of the experimental data, the unfolding free energy, ΔG_T, at temperature T will be given by:

$$\Delta G_T = \Delta H_1 - (\frac{T}{T_1})(\Delta H_1 - \Delta G_1) + \Delta Cp\ (T - T_1 - T \ln (\frac{T}{T_1})) \qquad (5)$$

where ΔH_1 and ΔG_1 are the unfolding enthalpy and the unfolding free energy at some temperature T_1 in the thermal transition zone where ΔG_1 can be measured (10). The data shown in Figure 5 were least squares fit to equation 5 using T_1 = 54°C and ΔG_1 = 7.3 kcal/mole. The best fit was obtained with ΔC_p = 1.6 kcal/mole-deg and ΔH_1 = 71 kcal/mole. A theoretical curve using these values is also shown in Figure 5.

The data shown in Fig. 5 indicate a slight (0.4 kcal/mole) decrease in Arc stability at 5°C relative to 25°C. To confirm the stability decrease at low temperature, we determined the fluorescence spectrum at 2°C and 20°C of an Arc sample under conditions where the protein is about half-denatured (1.6 μM Arc, 1.5 M urea, 10 mM Tris [pH 7.3] and 100 mM KCl) . The spectrum at the low temperature was significantly red-shifted and had diminished intensity. Control experiments showed that this change could not be caused by temperature dependent changes in the fluorescence

intensity of the folded or unfolded protein (which increase with decreasing temperature) without a shift from folded to unfolded Arc. Thus, Arc indeed appears to be destabilized at the lower temperature.[1]

The pH and salt dependence of Arc stability: The stability of Arc is both pH and salt dependent. As shown in Figure 6, the thermal stability of Arc decreases significantly as the pH is lowered below 6. At a concentration of 32 μM, Arc was found to be completely unfolded at pH 2 and room temperature (data not shown).

Figure 7a shows melting curves for Arc at different concentrations of KCl. The protein is clearly stabilized by increasing salt concentration. Since potassium and chloride ions have relatively minor effects on hydrophobic forces (12), it would appear that salt stabilizes Arc because ions bind preferentially to the folded form of the protein. The difference in the number of ions bound to the folded state relative to the unfolded state can be obtained from the slope of a plot of ln (**a**) vs. ln (K_u), where **a** is the KCl activity and K_u is the apparent equilibrium constant for denaturation (11, 13). Such a plot is shown in figure 7b for K_u values at 40°C. The slope of the line is 2.1 indicating that two ions are bound to the folded dimer.

---

[1] A worst case estimate of the error in ΔC_p can be obtained from the finding that ΔG_u reaches a maximum above 5°C. Since $d\Delta G/dT = -\Delta S$, ΔS_u, the entropy of unfolding, must also reach zero above 5°C. At 54°C, ΔS_u = 0.2 kcal/mole-deg and the dependence of ΔS on temperature is given by $d\Delta S/dT = \Delta C_p/T$. If ΔS reaches zero at 5°C, then ΔC_p (worst case) =1.2 kcal/mole-deg.

Table 1 lists the effect of various salts on the melting temperature of Arc. Because sulfate and phosphate ions can affect hydrophobic forces (12), it is not possible, in these cases, to assume that salt stabilization is caused solely by differential ion binding. Nevertheless, since Arc has a significant net positive charge and binds strongly to DNA and negatively charged ion-exchange columns, but not to positively charged ion-exchange columns, it would be surprising if the protein did not bind anions such as chloride, phosphate, and sulfate.

DISCUSSION

A number of small DNA binding proteins appear to be folded only as dimers. These include the λ Cro protein and the Trp repressor. The crystal structure of Cro shows two distinct subunits, with a dimer interface formed by a region of antiparallel β-sheet (14). While the structure of the Cro dimer appears to contain discrete domains, dissociation and denaturation of Cro dimers shows two-state behavior (15), indicating that the folded Cro monomer must be unstable. For Trp repressor the situation is different. In this case, the structural elements from each monomer are interwoven in the dimer and cannot be obviously separated into distinct folding units (16). As might be expected, Trp repressor also shows two-state dissociation and unfolding (C.R. Matthews, personal communication). In this case, however, unlike the λ Cro protein, it makes no sense to speak of folded monomers in a structural sense.

The results presented here show that the Arc dimer dissociates and unfolds in a concerted two-state reaction. Thus, the equilibrium constant measured in the transition zone of the denaturation curves reflects the overall stability of the Arc dimer relative to the unfolded monomer. It is possible that the Arc dimer consists of two unstable domains as seen in Cro repressor or as a single domain containing two polypeptide chains as seen in Trp repressor. In either case, the finding that folding and dimerization in Arc are tightly coupled suggests that Arc needs to form dimers to maintain a stably folded structure.

The dissociation-unfolding constant for Arc, K_u ,has a value of about 10^{-8} M at 25 °C in 10 mM potassium phosphate [pH 7.3] and 100 mM KCl. Thus, the protein is expected to be largely dissociated below 10^{-8} M. Under similar buffer conditions, half-maximal binding of Arc to operator DNA is observed at a protein concentration of about 5×10^{-10} M (17). Thus Arc is almost certainly monomeric and probably unfolded at the concentrations where operator binding is measured. When bound to the operator under these conditions, Arc is tetrameric (B. Brown, unpublished). Hence, operator binding is a complex reaction that involves folding and dimer formation, tetramer formation, and DNA binding. Clearly, effects on any of these coupled reactions will affect the level of binding observed, and interpreting the effects of solution conditions or amino acid substitutions on operator binding will be complicated. For example, Arc binding to operator DNA changes as a function of temperature, salt, and pH (18), but as we have shown here, so does the stability of the dimer. In some cases, these effects are in opposite directions. For example, operator binding becomes

weaker as the KCl concentration is raised (18), whereas the dimer becomes more stable. Hence, the intrinsic salt dependence of dimer-operator binding must be greater than is indicated by the overall salt dependence of the binding reaction. The consequences of changes in pH, salt, and temperature on Arc stability reported here should be quite useful in helping to understand the overall operator binding reaction.

ACKNOWLEDGEMENTS

We would like to thank C.N. Pace and D.V. Laurents for making their work available to us prior to publication, Kevin Shoemaker for pointing out the possibility of Hofmeister effects and Peter Kim for the use of the spectrofluorimeter and the CD spectropolarimeter.

REFERENCES

1. Knight, K. L., Bowie, J. U., Vershon, A. K., Kelley, R. D. & Sauer, R. T. (1989) *J. Biol. Chem.* **264**, 3639-3642.

2. Bowie, J. U. & Sauer, R. T. (1989) *J. Biol. Chem.* in press.

3. Vershon, A. K., Bowie, J. U., Karplus, T. M. & Sauer, R. T. (1986) *Proteins: Structure Function and Genetics* **1**, 302-311.

4. Bowie, J. U. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2152-2156.

5. Jordan, S. R., Pabo, C. O., Vershon, A. K. & Sauer, R. T. (1985) *J. Mol Biol.* **185,** 445-446.

6. Zagorski, M. G., Bowie, J. U., Vershon, A. K., Sauer, R. T. & Patel, D. J. (1989) *J. Mol. Biol.* submitted.

7. Knight, K. L. & Sauer, R. T. (1989) *Proc. Natl. Acad. Sci., USA* **86,** 797-801.

8. Vershon, A. K., Youderian, P., Susskind, M. M. & Sauer, R. T. (1985) *J. Biol. Chem.* **260,** 12124-12129.

9. Teale, F. W. J. (1960) *Biochem. J.* **76,** 381.

10. Privalov, P. L. (1979) *Adv. Protein Chem.* **33,** 167-241.

11. Pace, C. N. & Laurents, D. V. (1989) *Biochemistry* **28,** 2520.

12. von Hippel, P. H. & Wong, K. (1964) *Science* **145,** 577-580.

13. Record, M. T., Anderson, C. F. & Lohman, T. M. (1978) *Q. Rev. Biophys.* **11,** 103-178.

14. Anderson, W. F., Ohlendorf, D. H., Takeda, Y. & Matthews, B. W. (1981) *Nature* **290,** 754-758.

15. Pakula, A. A. & Sauer, R. T. (1989) *Proteins: Structure Function and Genetics* in press.

16. Schevitz, R. W., Otwinowski, Z., Jachimiak, A., Lawson, C. L. & Sigler, P. B. (1985) *Nature* **317,** 782-786.

17. Vershon, A. K., Kelley, R. D. & Sauer, R. T. (1989) *J. Biol. Chem.* in press.

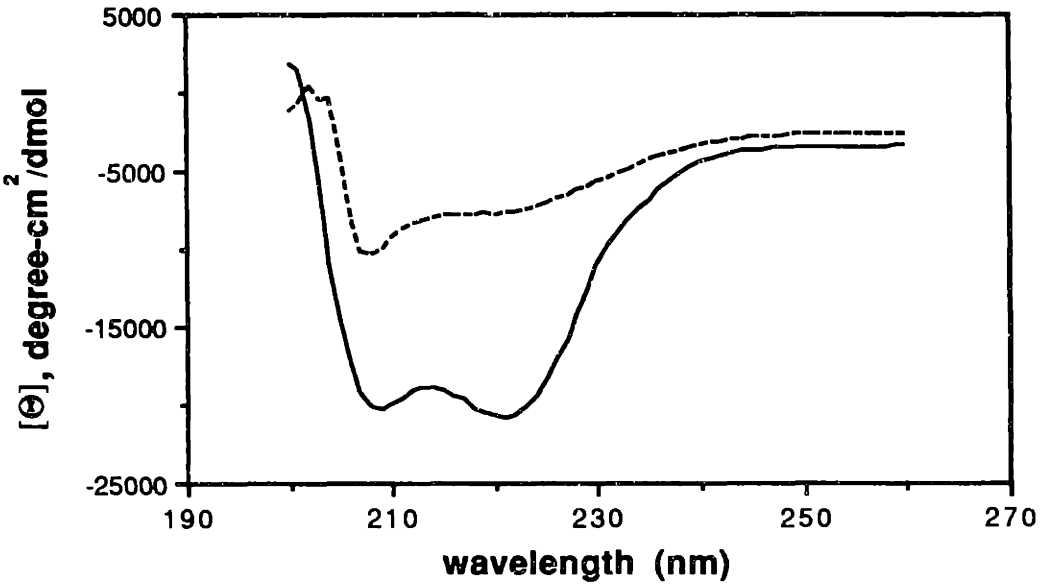
18. Vershon, A. K., Liao, S., McClure, W. R. & Sauer, R. T. (1987) *J. Mol. Biol.* **195,** 323-331.

Salt	Conc.(M)	T_m (°C)	
		pH 4.0	pH 7.3
KCl	0.1	33	54
KCl	0.25	38	56
KCl	0.5	43	60
KCl	1.0	51	63
NaCl	0.5	44	--
NH_4Cl	0.5	45	--
$MgCl_2$	0.5	46	--
KH_2PO_4	0.5	56	--
Na_2SO_4	0.5	61	--

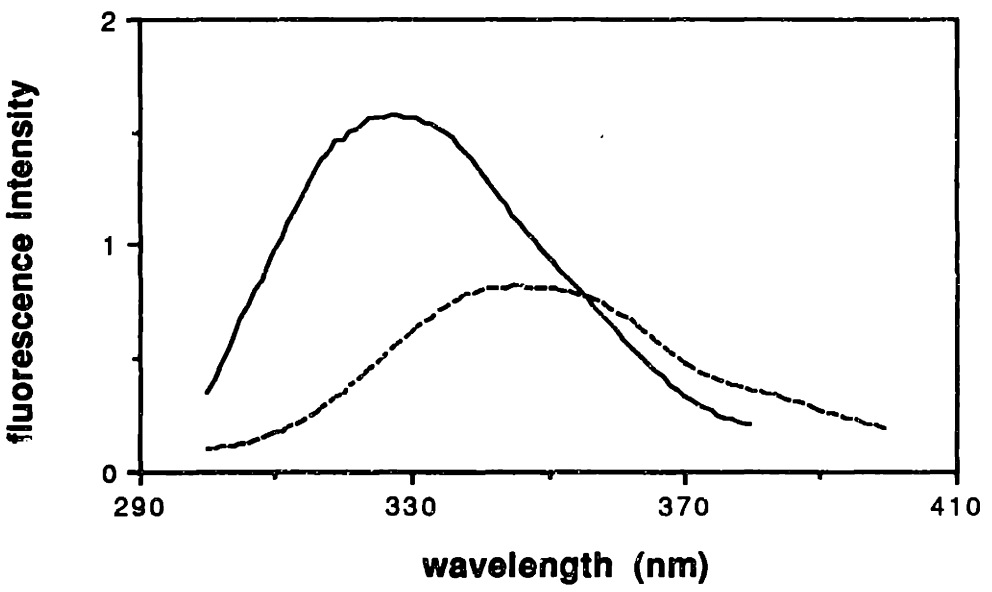
Table I. The effect of various salts on the thermal stability of Arc repressor. T_m is the temperature at which the protein is half denatured. These experiments were performed at Arc concentrations of 8 μM in 10 mM potassium acetate [pH 4.0] or 10 mM potassium phosphate [pH 7.3].

Figure 1. Circular dichroism and fluorescence spectra of native and denatured Arc. a) CD spectra at 25 °C (solid line) and 80 °C (dotted line) of 16 μM Arc in 10 mM potassium phosphate [pH 7.3] and 100 mM KCl. b) Fluorescence spectra of 1.6 μM Arc in 10 mM potassium phosphate [pH 7.3] and 100 mM KCl at 25 °C with (dotted line) or without (solid line) 4.8 M urea.

a)


5000
-5000
-15000
-25000
190
210
230
250
270
[Θ], degree-cm 2 /dmol
wavelength (nm)

b)


2
1
0
290
330
370
410
fluorescence intensity
wavelength (nm)

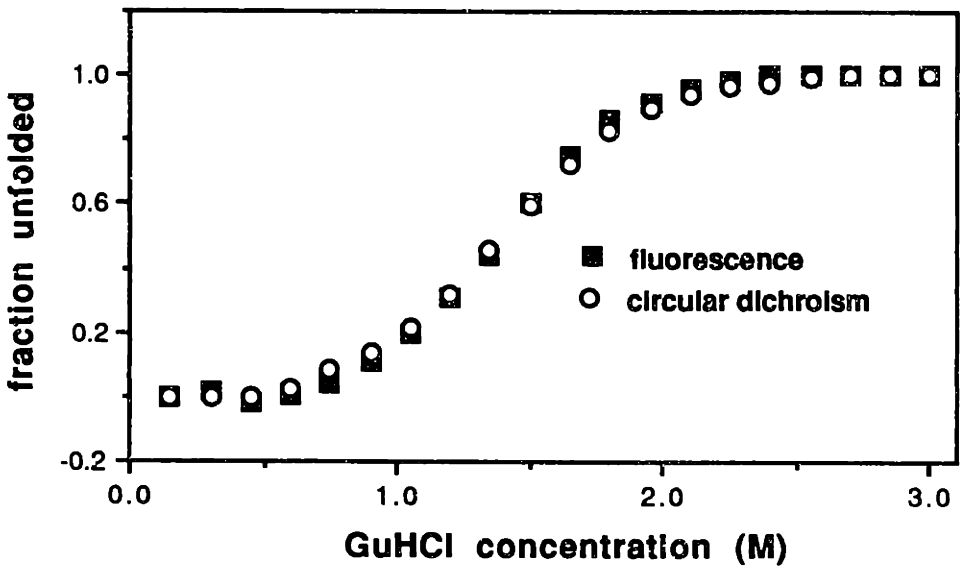
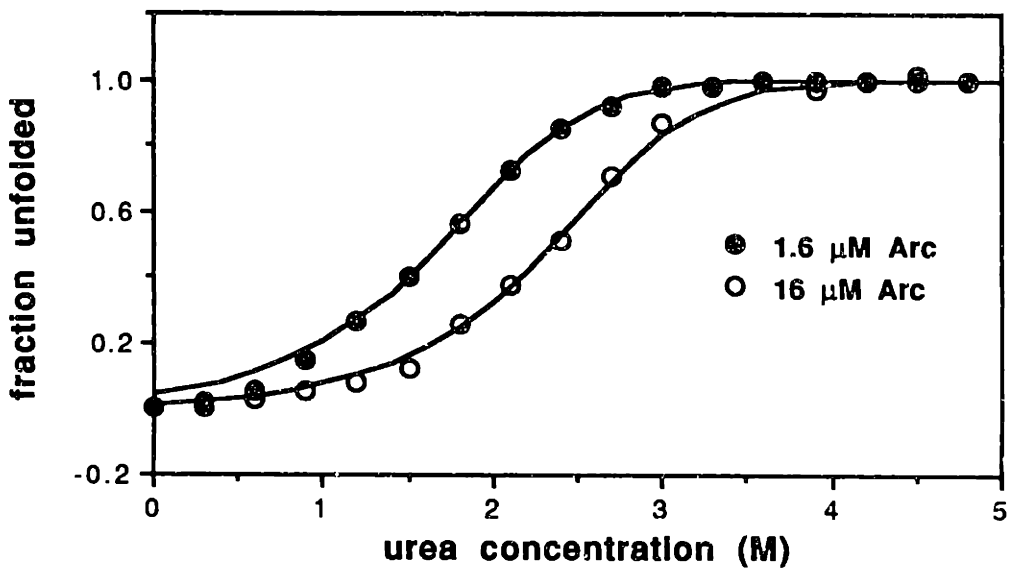


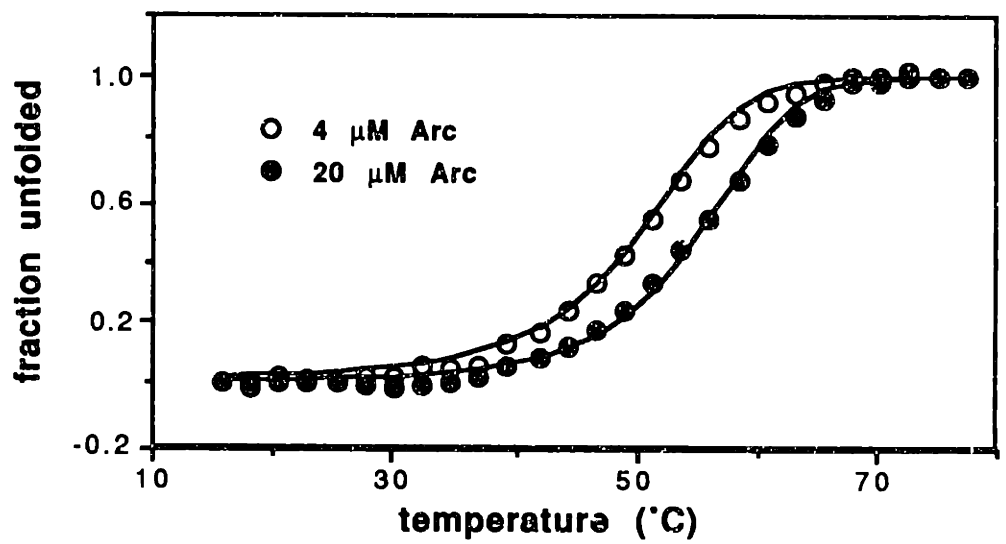
Figure 2. GuHCl denaturation curves monitored by fluorescence or circular dichroism. The data were obtained with 16 μM Arc in 50 mM Tris-HCl [pH 7.5] and 100 mM KCl, at 20°C.

Figure 3. Concentration dependence of Arc denaturation. a) Urea denaturation of 1.6 μM and 16 μM Arc in 10 mM potassium phosphate [pH 7.3] and 100 mM KCl at 25°C. Theoretical curves were calculated with equation 4 using ΔG_u^{H20} = 11.1 kcal/mole and m=-1.91 for 1.6 μM Arc and ΔG_u^{H20} = 11.0 kcal/mole and m= -1.91 for the 16 μM Arc concentration. These parameters were obtained from an least squared fit to equation 4, using ΔGu values determined in the transition zones of the two denaturation curves. b) Thermal denaturation curves of 4 μM and 20 μM Arc in 10 mM potassium phosphate [pH 7.3] and 100 mM KCl. Theoretical curves were calculated from equation 5 using T_1= 54°C, ΔG_1 = 7.3 kcal/mole, ΔH_1=71 kcal/mole and ΔCp = 1.6 kcal/mole-deg (see text).

a)


fraction unfolded
1.0
0.6
0.2
-0.2
0
1
2
3
4
5
urea concentration (M)
1.6 µM Arc
16 µM Arc

b)


fraction unfolded
1.0
0.6
0.2
-0.2
10
30
50
70
temperature (°C)
4 µM Arc
20 µM Arc

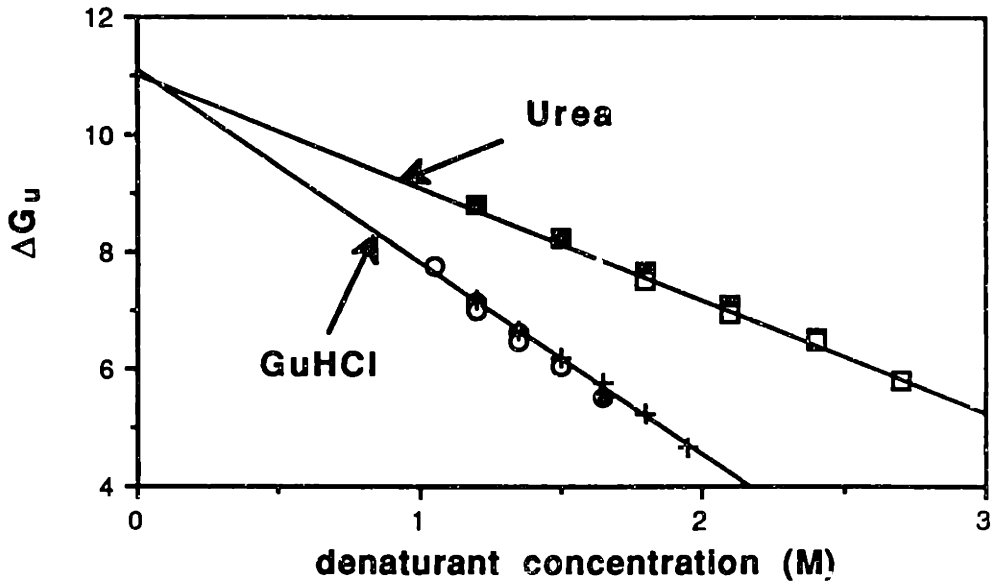


Figure 4. Unfolding free energy (ΔGu) of Arc repressor as a function of urea concentration or GuHCl concentration. ΔGu values were calculated using the two state assumption (equation 3). The urea denaturation free energies at Arc concentrations of 1.6 μM (filled squares) and 16 μM (open squares) were calculated using the data shown in figure 3. The least squares line has a slope of -1.91 kcal/mole-M and an intercept at $\Delta G_u^{H_2O}$= 11 kcal/mole. The data for GuHCl denaturation free energies were calculated from denaturation curves obtained with Arc concentrations of 4 μM (open circles), 16 μM (filled circles), and 32 μM (crosses) in 50 mM Tris-HCl [pH 7.5] and 100 mM KCl at 20°C. The least squares line has a slope of -3.27 kcal/mole-M and a intercept at $\Delta G_u^{H_2O}$= 11.1 kcal/mole.

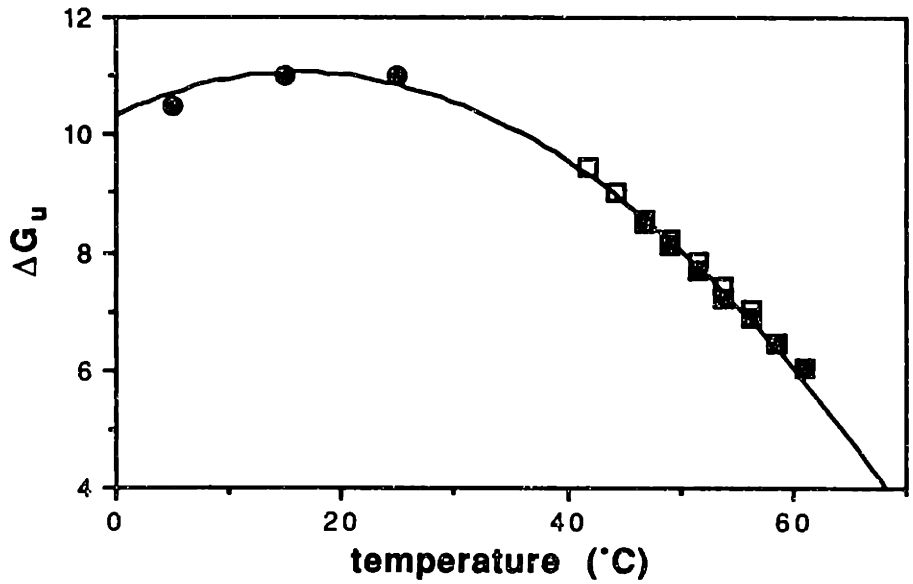


Figure 5. Unfolding free energy as a function of temperature. Square boxes represent unfolding free energies measured directly from the transition zones of the thermal denaturation curves shown in figure 3 at Arc concentrations of 4 μM (open boxes) and 20 μM (filled boxes). Circles represent $\Delta G_u{}^{H_2O}$ values determined from an analysis of urea denaturation curves determined at various temperatures. The urea denaturation curves were obtained at 1.6 μM Arc concentrations in the same buffer used for the thermal denaturation curves. The curve shows the least squares fit of the points to equation 5 as described in the text.

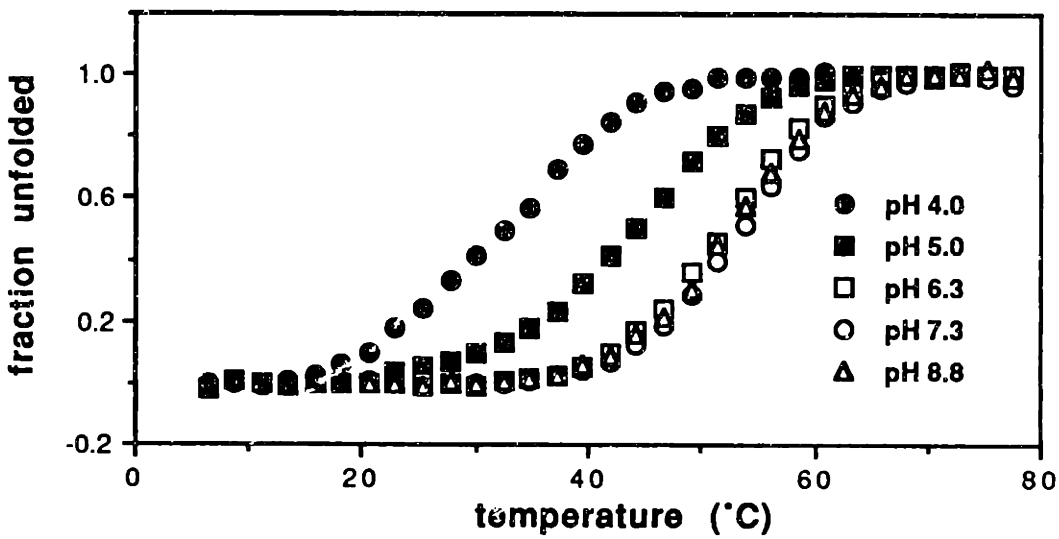
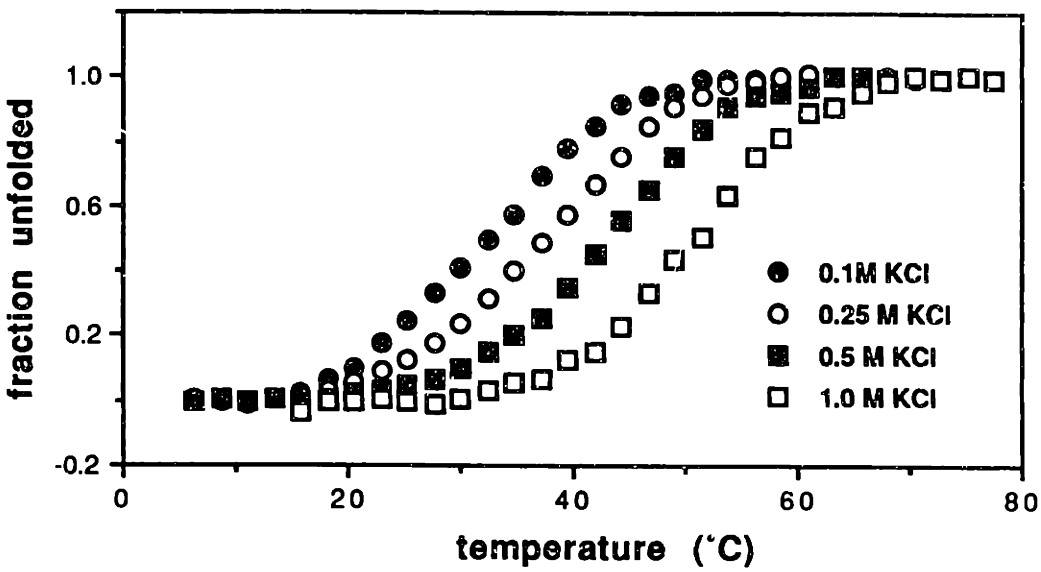


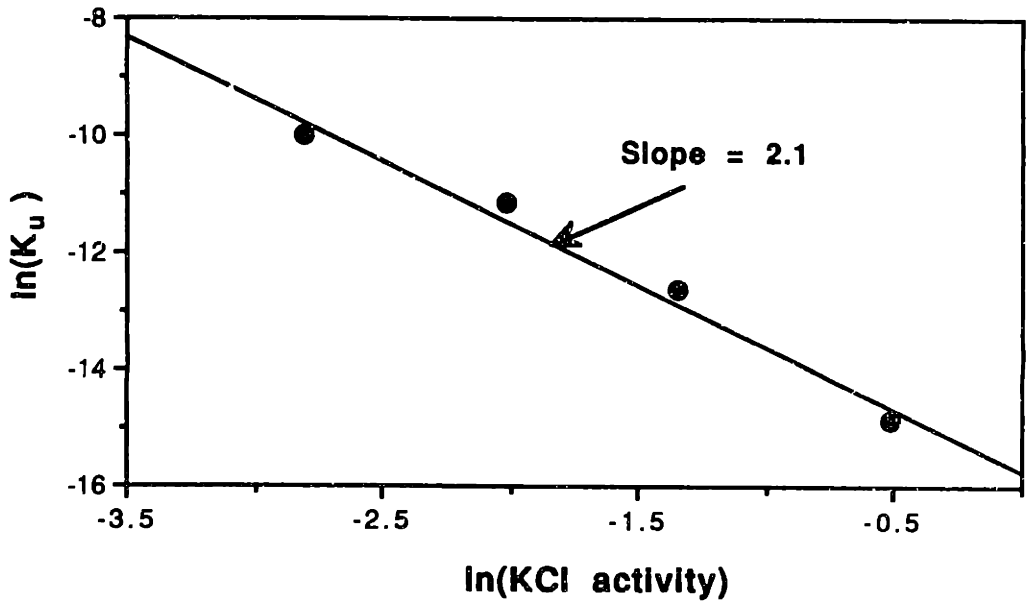
Figure 6. pH dependence of Arc stability. Thermal denaturation curves were obtained using protein concentrations of 8 μM in 100 mM KCl and; (i) 10 mM potassium acetate [pH 4.0]; (ii) 10 mM potassium acetate [pH 5.0]; (iii) 10 mM potassium phosphate [pH 6.3]; (iv) 10 mM potassium phosphate [pH 7.3]; (v) 10 mM glycine [pH 8.8].

Figure 7. Salt dependence of Arc stability. a) Thermal denaturation curves at protein concentrations of 8 μM in 10 mM potassium acetate [pH 4.0] and 0.1 M, 0.25 M, 0.5 M, and 1.0 M KCl. b) Plot of ln(Ku) vs. ln(KCl activity) at 40 °C. Equilibrium constants at 40 °C were estimated by linear extrapolation of van't Hoff plots of the data from the denaturation curves shown in figure 7a.

a)


fraction unfolded
1.0
0.6
0.2
-0.2
0
20
40
60
80
temperature (°C)
0.1M KCl
0.25 M KCl
0.5 M KCl
1.0 M KCl

b)


ln(K_u)
-8
-10
-12
-14
-16
-3.5
-2.5
-1.5
-0.5
ln(KCl activity)
Slope = 2.1