scan

Carrigan, John

05P1977

From: Sean Pitman [Seanpit@gmail.com]
Sent: Wednesday, November 30, 2005 8:31 AM
To: Library; Wood, Elizabeth
Subject: Photocopy and interlibrary loan request:

Importance: High

✓ You have an interlibrary loan request from:
Sean Pitman
Phone #: 63558
Department: Pathology
Account no: 999911969
Invoice for personal requests? no
fax #:
Medline UI:10.1002/prot.340070403
Journal title: Proteins: Structure, Function, and Genetics Year published: 1990
Volume: Volume 7, Issue 4 , Pages 306 - 316 Article title: Functionally
acceptable substitutions in two -helical regions of repressor
authors: John F. Reidharr-Olson, Robert T. Sauer
service: Regular
Deliver by:PDF_email

Type of User: Internal COH
IP Address: 151.152.74.150
Referer: http://library.coh.org/ill.asp?mode=confirm&sel=10
Browser Type: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)
Platform: WinNT
AOL?: False
TimeStamp: 11/30/2005 8:31:06 AM

PROTEINS: Structure, Function, and Genetics 7:306–316 (1990)

Functionally Acceptable Substitutions in Two α-Helical Regions of λ Repressor

John F. Reidhaar-Olson and Robert T. Sauer
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

ABSTRACT **A method of targeted random mutagenesis has been used to investigate the informational content of 25 residue positions in two α-helical regions of the N-terminal domain of λ repressor. Examination of the functionally allowed sequences indicates that there is a wide range in tolerance to amino acid substitution at these positions. At positions that are buried in the structure, there are severe limitations on the number and type of residues allowed. At most surface positions, many different residues and residue types are tolerated. However, at several surface positions there is a strong preference for hydrophilic amino acids, and at one surface position proline is absolutely conserved. The results reveal the high level of degeneracy in the information that specifies a particular protein fold.**

Key words: neutral mutations, random mutagenesis, protein structure, protein folding, protein families

INTRODUCTION

An important problem with regard to protein structure and function involves understanding how the sequence of a protein determines its three-dimensional structure. This problem is complicated in part by the fact that not all residues in proteins contribute equally to the structure; some residues are clearly more important than others. At present, it is not possible to tell, simply by examining a sequence, which residues are essential in determining the structure and which are less important. However, by examining a set of related sequences, each of which forms a functional protein with the same basic structure, it is often possible to make such a distinction. Such a family of sequences may be a set of phylogenetically related proteins generated by natural selection, or a set of variants of a single protein generated by random mutagenesis and identified by a functional selection. By examining the pattern of allowed substitutions at each position in the sequence, it is possible to assess the importance of each position to the structure or activity of the protein. Positions that tolerate few substitutions are high in informational content, whereas positions that are highly variable are low in informational content.

We have used a targeted random mutagenesis technique to generate a large number of neutral amino acid substitutions in the N-terminal domain of λ repressor. The crystal structure of the N-terminal domain is known, both for the isolated protein[1] and for its complex with operator DNA[2] (Fig. 1). The N-terminal domain consists of residues 1–92 of the intact protein, and is made up of five α-helices separated by loops of various lengths.[1] The protein binds to operator DNA as a dimer, with dimerization mediated by hydrophobic packing of α-helix 5 of one monomer against α-helix 5′ of the other monomer.[1] Previously, we reported the results of random mutagenesis of the part of helix 5 which forms the dimer interface region of the protein,[3] and found that the informational content at each position in the sequence correlates well with the extent to which the side chain at that position is buried in the structure. Here, we present an extension of that work to a much larger portion of the sequence, including the rest of helix 5 and all of helix 1. These regions of the protein are of interest for structural studies because they are the two longest helices in the N-terminal domain and, with the exception of two residues in helix 1, make no direct contacts with the operator DNA. We find that, in general, buried positions are high in informational content, whereas surface positions are low in informational content. However, a class of surface positions with restricted substitution patterns is also observed, which was not seen in the previous work.

MATERIALS AND METHODS

Strains and Plasmids

Escherichia coli K-12 strain X90 (*ara*Δ(*lac pro*) *nalA argE*am *rif thi-1*/F′ *lacI*^{Q} *lac*$^{+}$ *pro*$^{+}$) was used in this work.[4] Plasmids pJO103,[3] pDP160 (D. Parsell, unpublished), and pWL103 and pWL104 (W. Lim, unpublished) are pBR322-derived plasmids that contain the gene for residues 1–102 of λ repressor under control of a *tac* promoter, an ampicillin-

Received July 24, 1989; revision accepted December 29, 1989.

Address reprint requests to Robert T. Sauer, Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139.

Fig. 1. Stereo view of a ribbon representation of the N-terminal domain of λ repressor bound as a dimer to operator DNA.[2] The upper monomer is colored blue, with helices 1 and 5 colored yellow; the lower monomer is colored green with helices 1 and 5 colored orange. Helix 1 is part of the globular portion of the protein, whereas helix 5 extends out from the rest of the monomer to form the dimer interface. These graphics were produced using coordinates provided by Carl Pabo and Steve Jordan.

resistance gene, and an M13 origin of replication that allows production of single-stranded DNA for sequencing. The gene for residues 1–102 is a synthetic gene that contains several restriction sites to allow cassette mutagenesis. Plasmids pJO103 and pDP160 contain "stuffer" fragments which prevent uncut or singly-cut plasmid from passing the functional selection.

Mutagenesis

Combinatorial cassette mutagenesis was performed as described.[3] Oligonucleotide cassettes were designed to randomize the codons for residues 8–23 and 75–83 of the N-terminal domain of λ repressor. The oligonucleotides were synthesized with NN^G_C in place of a codon being randomized; that is, an equal mixture of all four bases (N) was included at the first two positions of the codon, and an equal mixture of G and C was included at the third position. Codons were randomized in the following combinations: 8/9; 10/11; 12/13; 14; 14/17; 15/16; 17/18; 19/20; 21/22; 23; 75/76; 77/78; 79/80; 81/82; 81/82/83; and 83. Cassettes were ligated[5] into the appropriate restriction sites of the synthetic N-terminal domain gene. Ligated DNA was transformed[6] into X90 cells, and the cells were plated on LB plates containing 0.1 mg/ml ampicillin and 10^9 each of the *c*I⁻ phages λKH54 and λKH54h80. Cells containing the wild-type N-terminal domain or variants with greater than 5–10% of wild-type activity survive this selection.[3] Single-stranded plasmid DNA was isolated from cells surviving the selection and was sequenced by the dideoxy method.[7]

Two mutations, Lys-83 and Val-84, were constructed by site-directed mutagenesis. The Lys-83 mutant was found to be active, i.e., resistant to phage λKH54; the Val84 mutant was found to be inactive.

Monte Carlo Simulations of Random Mutagenesis

Simulations of the mutagenesis experiments were performed using a computer program which randomly chooses amino acids from a set of allowed residues. The probability of choosing a particular residue is proportional to the number of codons that specify that amino acid. The user specifies the set of allowed residues and the number s of random sequences to be chosen. The program then performs 1000 trials in which s random amino acids are chosen from the allowed set, and calculates the number of trials in which 1, 2, 3, etc. different residues were observed. Simulations of this type were used to determine, at each position examined by random mutagenesis, the probability of observing $n + 1$ amino acids in a set of s sequences, where n is the number of residues actually observed at that position and s is the number of candidates actually sequenced. For these simulations, the set of allowed residues was comprised of the amino acids actually observed in the random mutagenesis experiments, plus one additional residue which was encoded by a single codon in the NN^G_C randomization.

RESULTS

Experimental Design

In previous work,[3] we used combinatorial cassette mutagenesis to investigate the informational content of helix 5 residues at the dimer interface of the N-terminal domain (residues 84–91). In this study, we have used the same technique to extend our analysis to a larger portion of the protein. Two separate

regions, highlighted in Figure 1, were examined. The first consists of residues 75–83. This includes the first half of helix 5 (residues 78–83), as well as the last three positions (residues 75–77) in the loop region between helices 4 and 5. The second consists of helix 1 (residues 9–23) and the residue immediately preceding it (residue 8).

Separate mutagenic cassettes were designed to randomize from one to three codons at a time. Oligonucleotides were synthesized with an equal mixture of all four bases at the first two positions of the codons being randomized and an equal mixture of G and C at the third position. This technique generates a pool of cassettes which encodes all 20 amino acids at those codons. The cassettes were ligated into the appropriate plasmid backbone, and the resulting plasmids were transformed into *E. coli*. Genes encoding active proteins were identified by a functional selection, and a number of candidates from each randomization were sequenced. The result is a list of functional sequences, from which one can determine the spectrum of allowed substitutions at each position.

Allowed Substitutions

A list of functional N-terminal domain sequences is shown in Figure 2. Because of the experimental design, many of the functional sequences contain more than one mutation. However, in general there appears to be no dependence of a change at one position on a change at another, as most changes were recovered in several different mutant backgrounds. Consequently, we treat the changes as if they were independent of one another.

A summary of acceptable substitutions at each position is shown in Figure 3. This figure also shows the acceptable substitutions identified at positions 84–91 in our previous study.[3] In general, there is a fairly high level of tolerance to amino acid substitution in both the helix 1 and helix 5 regions of the protein. However, from position to position there is a wide range in the number of substitutions tolerated. Of the 33 positions that have been examined, 13 are informationally very rich, accepting only one or two residues. In contrast, 14 positions are quite low in informational content, tolerating nine or more different residues. The remaining six positions are intermediate in informational content, accepting between four and seven different residues.

The informational content of a position is a function of the role that the residue at that position plays in the structure or activity of the protein. In our work on the dimer interface,[3] we found a good correlation between the number of functional residues observed at a given position and the fractional solvent accessibility of the residue at that position in the crystal structure. This general trend persists in the extended analysis (Fig. 4), but there are several interesting exceptions. We will discuss the informational content of individual positions by grouping them into three categories: buried positions, surface positions with restricted substitution patterns, and surface positions that are highly variable.

Buried Residues

As shown previously,[3] residues buried in the protein structure are highly conserved, and thus high in informational content. In helix 1, four residues are greater than 90% buried: Asp-14, Ala-15, Leu-18, and Tyr-22. At each of these positions, only the wild-type residue is recovered. The side chain of Asp-14 forms a charge-stabilized hydrogen bond with Arg-17 and a hydrogen bond with Ser-77. Since Asp-14 is conserved, one or both of these interactions must be important; as discussed below, the charge-stabilized hydrogen bond between Asp-14 and Arg-17 appears to be the more critical of these two interactions. Ala-15 is conserved, presumably because larger side chains at this position would cause unfavorable steric contacts that would disrupt the structure; replacement with the smaller residue glycine would cause loss of hydrophobic and van der Waals interactions and would also increase the entropy of unfolding. Leu-18, which forms part of the hydrophobic core of the protein, is also conserved. Maintaining proper packing in the core apparently requires the precise size and geometry of the leucine side chain. Other side chains, even if hydrophobic, apparently cause some disruption of the structure by causing steric overlaps or by leaving unfavorable holes. Tyr-22 appears to be conserved for both structural and functional reasons. The aromatic ring is completely buried, while the phenolic hydroxyl extends into solvent where it makes a hydrogen bond to the phosphate backbone of the operator DNA.[2] In intact λ repressor, replacing Tyr-22 with histidine disrupts the structure, resulting in a protein with a T_m 30°C lower than that of the wild-type protein.[8] In contrast, a mutant with phenylalanine at this position is as stable as wild-type but has DNA binding activity that is reduced 15-fold.[9]

Position 21, which is about 75% solvent-inaccessible, accepts only the wild-type isoleucine and the conservative substitution leucine. Given the degree of exposure of Ile-21, it is somewhat surprising that the tolerance to substitution is so low at this position. In the crystal structure, Ile-21 occupies a crevice near the surface of the protein. Smaller side chains would probably lead to loss of hydrophobic packing interactions, but it is not immediately clear why larger hydrophobic side chains, or side chains such as lysine and arginine, could not be accommodated.

In the residue 75–83 region, the buried positions are also highly conserved. Phe-76 is completely buried in the core of the protein and is conserved in the randomization experiments. The other buried positions, Ser-77, Ile-80, and Ala-81, are also high in

a

10	15	20	
T Q E Q L E D A R R L K A I Y E			
R Q E Q L E D A R R L K A I Y E			*1*
R E E Q L E D A R R L K A I Y E			*1*
R H E Q L E D A R R L K A I Y E			*1*
R I E Q L E D A R R L K A I Y E			*1*
R W E Q L E D A R R L K A I Y E			*1*
N N E Q L E D A R R L K A I Y E			*1*
N W E Q L E D A R R L K A I Y E			*1*
G K E Q L E D A R R L K A I Y E			*2*
K A E Q L E D A R R L K A I Y E			*1*
K N E Q L E D A R R L K A I Y E			*1*
S R E Q L E D A R R L K A I Y E			*4*
S G E Q L E D A R R L K A I Y E			*1*
S I E Q L E D A R R L K A I Y E			*1*
S K E Q L E D A R R L K A I Y E			*2*
S P E Q L E D A R R L K A I Y E			*3*
S T E Q L E D A R R L K A I Y E			*1*
S Y E Q L E D A R R L K A I Y E			*1*
S V E Q L E D A R R L K A I Y E			*1*
T R E Q L E D A R R L K A I Y E			*2*
T F E Q L E D A R R L K A I Y E			*1*
T P E Q L E D A R R L K A I Y E			*1*
T Q **A Q** L E D A R R L K A I Y E			*1*
T Q **R Q** L E D A R R L K A I Y E			*5*
T Q **N Q** L E D A R R L K A I Y E			*1*
T Q **C Q** L E D A R R L K A I Y E			*1*
T Q **Q Q** L E D A R R L K A I Y E			*2*
T Q **E R** L E D A R R L K A I Y E			*2*
T Q **E Q** L E D A R R L K A I Y E			*3*
T Q **E K** L E D A R R L K A I Y E			*3*
T Q **G Q** L E D A R R L K A I Y E			*1*
T Q **I Q** L E D A R R L K A I Y E			*2*
T Q **L Q** L E D A R R L K A I Y E			*1*
T Q **K Q** L E D A R R L K A I Y E			*3*
T Q **M Q** L E D A R R L K A I Y E			*3*
T Q **M E** L E D A R R L K A I Y E			*1*
T Q **S Q** L E D A R R L K A I Y E			*1*
T Q **T Q** L E D A R R L K A I Y E			*2*
T Q **V Q** L E D A R R L K A I Y E			*3*
T Q E Q **A A** D A R R L K A I Y E			*1*
T Q E Q **R A** D A R R L K A I Y E			*3*
T Q E Q **R D** D A R R L K A I Y E			*1*
T Q E Q **R E** D A R R L K A I Y E			*1*
T Q E Q **R M** D A R R L K A I Y E			*1*
T Q E Q **Q A** D A R R L K A I Y E			*1*
T Q E Q **Q Q** D A R R L K A I Y E			*2*
T Q E Q **I D** D A R R L K A I Y E			*1*
T Q E Q **I L** D A R R L K A I Y E			*1*
T Q E Q **L R** D A R R L K A I Y E			*1*

10	15	20	
T Q E Q L E D A R R L K A I Y E			
T Q E Q **L E** D A R R L K A I Y E			*1*
T Q E Q **L M** D A R R L K A I Y E			*1*
T Q E Q **L S** D A R R L K A I Y E			*2*
T Q E Q **K R** D A R R L K A I Y E			*1*
T Q E Q **K E** D A R R L K A I Y E			*2*
T Q E Q **K H** D A R R L K A I Y E			*1*
T Q E Q **M A** D A R R L K A I Y E			*2*
T Q E Q **M E** D A R R L K A I Y E			*1*
T Q E Q **F L** D A R R L K A I Y E			*1*
T Q E Q **T A** D A R R L K A I Y E			*1*
T Q E Q **V Q** D A R R L K A I Y E			*1*
T Q E Q **V E** D A R R L K A I Y E			*1*
T Q E Q L E **D** A R R L K A I Y E			*9*
T Q E Q L E **D** A **R** L K A I Y E			*32*
T Q E Q L E **D** A R **K** L K A I Y E			*1*
T Q E Q L E D **A R** R L K A I Y E			*27*
T Q E Q L E D **A Q** R L K A I Y E			*1*
T Q E Q L E D **A K** R L K A I Y E			*3*
T Q E Q L E D **A M** R L K A I Y E			*5*
T Q E Q L E D **A S** R L K A I Y E			*1*
T Q E Q L E D A R **R L** K A I Y E			*13*
T Q E Q L E D A R **K L** K A I Y E			*4*
T Q E Q L E D A R R L **K A** I Y E			*2*
T Q E Q L E D A R R L **K R** I Y E			*10*
T Q E Q L E D A R R L **K E** I Y E			*3*
T Q E Q L E D A R R L **K G** I Y E			*4*
T Q E Q L E D A R R L **K L** I Y E			*5*
T Q E Q L E D A R R L **K K** I Y E			*1*
T Q E Q L E D A R R L **K M** I Y E			*5*
T Q E Q L E D A R R L **K S** I Y E			*3*
T Q E Q L E D A R R L **K T** I Y E			*1*
T Q E Q L E D A R R L K A **I Y** E			*21*
T Q E Q L E D A R R L K A **L Y** E			*2*
T Q E Q L E D A R R L K A I Y **A**			*1*
T Q E Q L E D A R R L K A I Y **R**			*4*
T Q E Q L E D A R R L K A I Y **Q**			*2*
T Q E Q L E D A R R L K A I Y **E**			*1*
T Q E Q L E D A R R L K A I Y **G**			*1*
T Q E Q L E D A R R L K A I Y **L**			*7*
T Q E Q L E D A R R L K A I Y **K**			*1*
T Q E Q L E D A R R L K A I Y **M**			*2*
T Q E Q L E D A R R L K A I Y **S**			*3*
T Q E Q L E D A R R L K A I Y **V**			*2*

Fig. 2. Functional sequences for the helix 1 **(a)** and helix 5 **(b)** regions of the N-terminal domain of λ repressor obtained by combinatorial cassette mutagenesis. Sequences are grouped by cassette, with the randomized positions in boldface. Additional mutations are shown in italics. Numbers next to sequences indicate the number of times particular mutant sequences were obtained. The asterisk indicates a mutant generated by site-directed mutagenesis (see text). The wild-type sequence and the amino acid position numbers are at the top of each column.

b

75					80					85					90			
E	F	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		
A	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		6
D	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		2
Q	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		8
E	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		5
S	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		13
T	**F**	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	**A**	**P**	S	I	A	R	E	I	Y	E	M	Y	E	A	V		4
E	F	**S**	**P**	S	I	A	R	E	I	Y	E	M	Y	E	A	V		24
E	F	S	P	**A**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**A**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**A**	**M**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**R**	**I**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**R**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**R**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**N**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**D**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**C**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**Q**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**Q**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**E**	**I**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**E**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**G**	**I**	A	R	E	I	Y	E	M	Y	E	A	V		4
E	F	S	P	**G**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**G**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**H**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**H**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**K**	**C**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**K**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**K**	**L**	A	R	E	I	*S*	E	M	Y	E	A	V		1
E	F	S	P	**S**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**S**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**T**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	**T**	**K**	A	R	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	**Y**	**L**	A	R	E	I	Y	E	M	Y	E	A	V		2

75					80					85					90			
E	F	S	P	S	I	A	R	E	I	Y	E	M	Y	E	A	V		
E	F	S	P	S	I	**A**	**A**	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**R**	E	I	Y	E	M	Y	E	A	V		7
E	F	S	P	S	I	**A**	**Q**	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**E**	E	I	Y	E	M	Y	E	A	V		3
E	F	S	P	S	I	**A**	**G**	E	I	Y	E	M	Y	E	A	V		9
E	F	S	P	S	I	**A**	**L**	E	I	Y	E	M	Y	E	A	V		3
E	F	S	P	S	I	**A**	**K**	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**M**	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**S**	E	I	Y	E	M	Y	E	A	V		4
E	F	S	P	S	I	**A**	**T**	E	I	Y	E	M	Y	E	A	V		5
E	F	S	P	S	I	**A**	**Y**	E	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**V**	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	S	I	**S**	**R**	E	I	Y	E	M	Y	E	A	V		2
E	F	S	P	S	I	A	R	**R**	I	Y	E	M	Y	E	A	V		8
E	F	S	P	S	I	A	R	**Q**	I	Y	E	M	Y	E	A	V		2
E	F	S	P	S	I	A	R	**Q**	I	Y	E	M	*F*	E	A	V		1
E	F	S	P	S	I	A	R	**E**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	A	R	**G**	I	Y	E	M	Y	E	A	V		3
E	F	S	P	S	I	A	R	**H**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	A	R	**L**	I	Y	E	M	Y	E	A	V		2
E	F	S	P	S	I	A	R	**K**	I	Y	E	M	Y	E	A	V		*
E	F	S	P	S	I	A	R	**M**	I	Y	E	M	Y	E	A	V		2
E	F	S	P	S	I	A	R	**S**	I	Y	E	M	Y	E	A	V		4
E	F	S	P	S	I	A	R	**T**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	A	R	**V**	I	Y	E	M	Y	E	A	V		3
E	F	S	P	S	I	**A**	**R**	**N**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**R**	**E**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**E**	**R**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**A**	**E**	**S**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**S**	**A**	**E**	I	Y	E	M	Y	E	A	V		1
E	F	S	P	S	I	**S**	**R**	**Q**	I	Y	E	M	Y	E	A	V		2

Fig. 2b. Legend appears on page 309.

informational content. The crystal structure indicates that the hydroxyl of Ser-77 is 2.8 Å from the carboxylate of Asp-14, suggesting that these two residues are hydrogen bonded.[2] However, the fact that alanine is recovered at position 77 indicates that this is not a crucial interaction. Ile-80 is 85% buried; this position accepts five different residues, four of which (isoleucine, leucine, methionine, and cysteine) are fairly hydrophobic. One somewhat surprising substitution allowed at this position is lysine. Presumably, the long aliphatic portion of the lysine side chain can substitute for a hydrophobic residue, as long as the terminal ϵ-NH_2 can reach the surface.

Surface Positions With Restricted Substitution Patterns

Of the 16 surface positions examined here, 7 (positions 8, 11, 16, 17, 19, 75, and 78) accept only a limited set of residues. The degree of restriction ranges from absolute conservation of the wild-type residue to tolerance of up to six residues. In cases where several residues are tolerated, the tendency is for most of them to be fairly hydrophilic. Some of the restricted surface positions can be understood in structural or functional terms from the cocrystal structure. Lys-19 makes a contact with the phosphate backbone of the operator DNA.[2] This contact must be energetically significant, since no other residue was recovered at this position. Arg-17, as mentioned above, forms a charge-stabilized hydrogen bond with Asp-14; in this case, only lysine and arginine are observed. The importance of this interaction is underscored by the fact that positions 14 and 17 were randomized as a pair. These results suggest that there may be no other combinations of side chains at these positions that are able to make an energetically equivalent interaction.

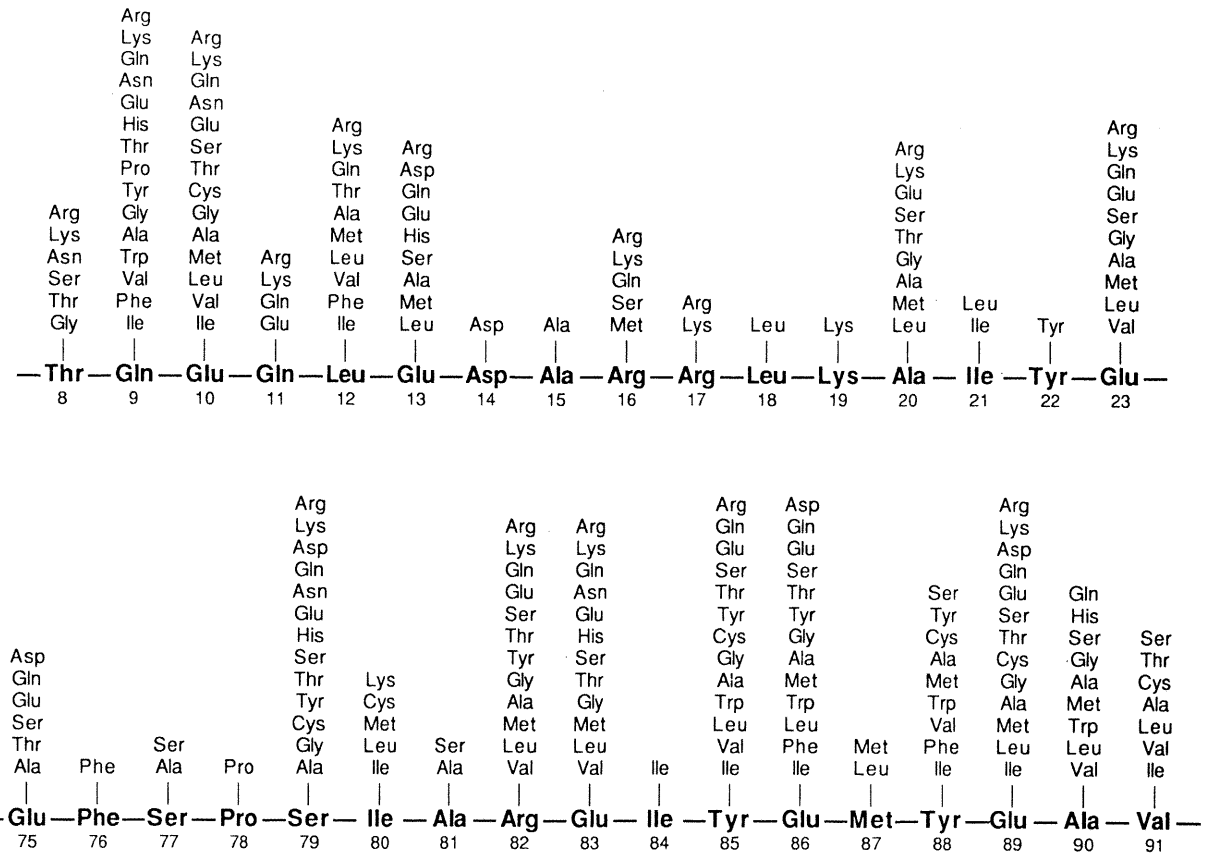

Fig. 3. Functionally acceptable residues in the helix 1 (top) and helix 5 (bottom) regions. The amino acids are listed from top to bottom in order of increasing hydrophobicity according to the scale of Eisenberg et al.[21] Data for positions 84–91 are from Reidhaar-Olson and Sauer.[3]

For the other surface positions with restricted substitution patterns there is no simple structural interpretation. The side chain of Pro-78 is over 80% exposed, and yet the wild-type residue is conserved at this position. At positions 8, 11, 16, and 75, the wild-type residue is over 50% exposed to solvent, but each of these positions accepts only a limited set of residues. There is a preference at each of these positions for charged and other polar residues. It is not immediately clear why these four positions should show such a preference for hydrophilic amino acids, since many other surface positions accept both polar and nonpolar residues (see below).

Surface Variable Positions

Nine surface positions accept highly variable substitutions. This class includes positions 9, 10, 12, 13, 20, 23, 79, 82, and 83. All of these positions accept at least nine different residues, and thus are fairly low in informational content. Position 9 accepts all residue types, and positions 79 and 82 accept all residue types except proline (positively and negatively charged, polar, nonpolar, and aromatic). However, at position 12, no negatively charged residues were observed, and at positions 10, 13, 20, 23, and 83, aromatics were not recovered. As discussed below, one must be cautious about drawing conclusions based on the few residues at these positions that are not seen; if more candidates were sequenced, additional residues might be found to be tolerated. For example, at position 83, substitution of the wild-type glutamate with lysine was not observed after sequencing 28 candidates. Since lysine is functional at position 83 in the intact protein,[10] this substitution was introduced into the N-terminal domain by site-directed mutagenesis. Subsequent tests confirmed that the mutant protein is functional. In general, the few residues that are not observed at highly variable positions would probably be functional or have only a weakly defective phenotype, since strongly defective mutations have never been observed at any of these positions, with the exception of a leucine to proline substitution at position 12.[11] In contrast, the buried positions, which are highly conserved in the randomization experiments, are often sites of defective mutations.[11]

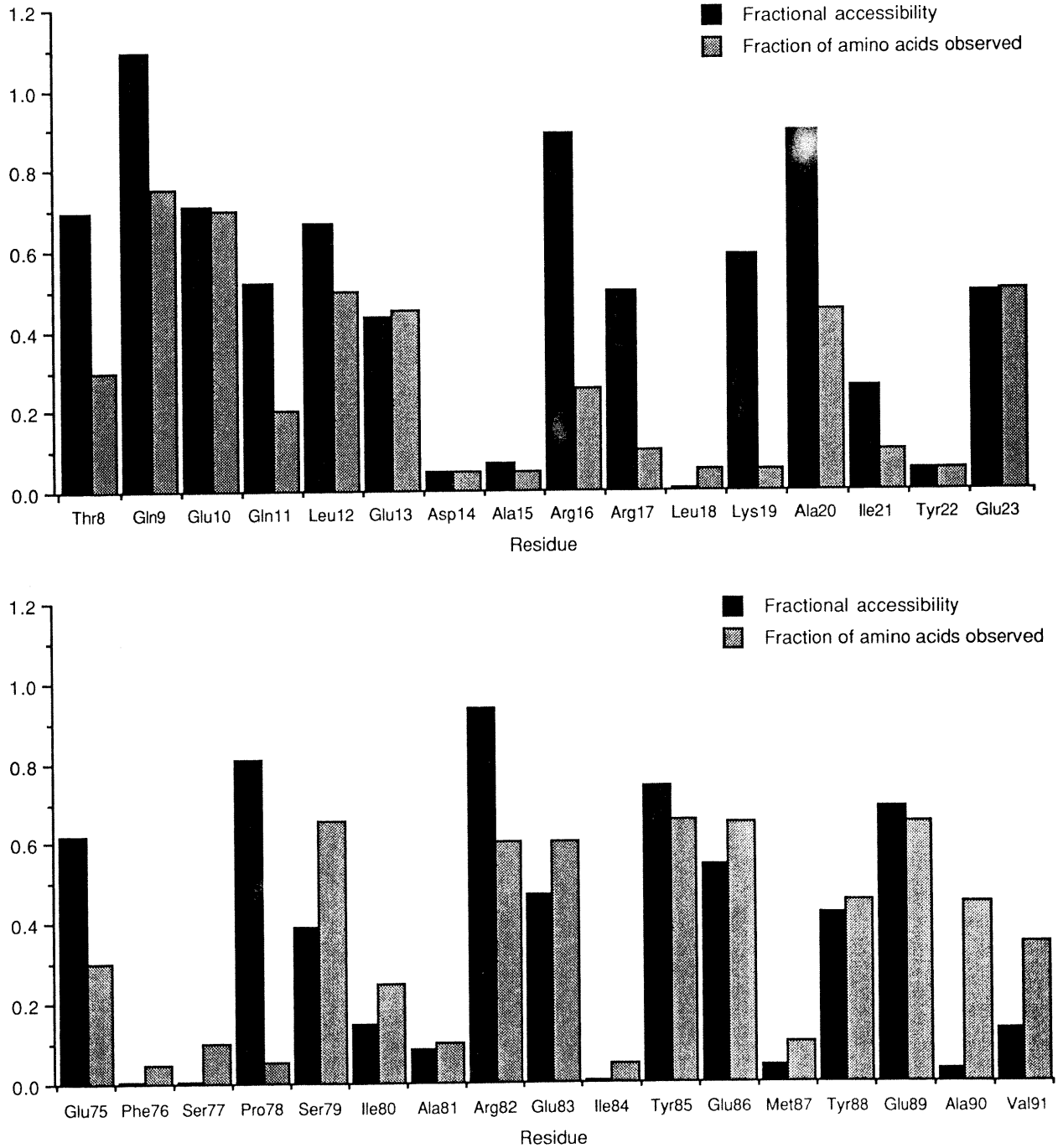

Fig. 4. Correlation between the solvent accessibility and the number of functional amino acids observed at positions in the helix 1 **(top)** and helix 5 **(bottom)** regions. Gray bars indicate the fraction of the 20 naturally occurring amino acids recovered at each position. Black bars indicate the fractional solvent accessibility of the wild-type side chain in the dimer. Solvent accessibilities were computed using the method of Lee and Richards,[22] with crystal coordinates provided by Carl Pabo and Steve Jordan.[2] Fractional accessibilities were obtained by dividing by the appropriate side chain accessibilities in the reference tripeptide Ala-X-Ala.[22] Data for positions 84–91 are from Reidhaar-Olson and Sauer.[3]

Number of Amino Acids Allowed at Each Position

In interpreting the results of our random mutagenesis experiments, we often draw inferences from the fact that many residues are not observed at a given position. It is possible, in this case, to reason from a negative result because the method of mutagenesis ensures that all possible sequence changes

are present in the mutagenized population. However, because of sampling statistics, it is important to be able to estimate the probability that the complete set of allowed amino acids has been observed at a given position. To estimate these probabilities we performed computer simulations of the random mutagenesis experiments. The simulation program was used to determine the likelihood that an additional allowed residue would have been observed in the set of functional candidates sequenced at that position. For example, at position 75, 36 candidates were sequenced and 6 functional amino acids were observed. If 7 amino acids were actually allowed at this position, what is the probability that all 7 would have been observed in a set of 36 sequences? If this probability were low, then it would be necessary to determine more sequences before concluding that the entire set of allowed changes had been recovered.

Table I shows the probability that an additional allowed residue would have been observed at each position examined in helix 1 and helix 5. At positions that accept a large number of different residues, it is clearly unlikely that the entire range of functional substitutions has been observed. This conclusion is supported by the finding, discussed above, that lysine is functional at position 83, even though it was not observed in the 35 functional candidates generated by random mutagenesis. However, at most of the highly restricted positions, the probability that all possible functional residues have been observed is quite high. For example, at position 84, each of the 17 functional sequences contains isoleucine. If an additional residue were functional at this position, the chance that it would have been observed in a set of 17 sequences is greater than 99%. This implies that amino acid substitutions at position 84 should result in proteins that do not pass the selection. To test this, a mutant was constructed in which Ile-84 was replaced by the conservative substitution valine. This mutant was found to be nonfunctional.

Although the simulations give estimates of the certainty that all allowed changes have been recovered, one must be cautious in interpreting these results, particularly at positions where one residue appears far more frequently than others. In generating the probabilities shown in Table I it was assumed that all functional amino acids are recovered with probabilities dictated solely by the number of codons for these residues. That is, all functional residues were assumed to have an equal chance of passing the imposed biological selection. However, as the list of functional sequences in Figure 2 indicates, at several positions one residue was recovered far more frequently than others. In most of these cases, the preference is for the wild-type residue. For example, in the randomization of position 21, the wild-type isoleucine was recovered 21 times, whereas leucine was recovered only twice. This clearly suggests that genes containing the overrepresented codons have some selective advantage, either because of differences in the intrinsic activities of the proteins or their levels of expression.[12] It is also possible that the nonrandom distribution in some cases reflects differential representation of the codons in the pool of mutagenic cassettes, although this cannot be a general problem since many different residues are recovered at a number of positions. In any case, it will never be possible to rigorously estimate the probability that additional residues would have been recovered without knowing whether these residues have a significant selective disadvantage.

TABLE I. Number of Candidates Sequenced and Number of Amino Acids Observed*

Position	s	n	$P_s(n+1)$
Thr-8	29	6	0.72
Gln-9	29	15	0.01
Glu-10†	35	14	0.05
Gln-11†	35	4	0.98
Leu-12	28	10	0.18
Glu-13	28	9	0.25
Asp-14	42	1	>0.99
Ala-15	37	1	>0.99
Arg-16†	37	5	0.90
Arg-17	50	2	>0.99
Leu-18	17	1	0.99
Lys-19	34	1	>0.99
Ala-20	34	9	0.46
Ile-21†	23	2	0.99
Tyr-22	23	1	>0.99
Glu-23	24	10	0.11
Glu-75	36	6	0.88
Phe-76	36	1	>0.99
Ser-77	28	2	>0.99
Pro-78	28	1	>0.99
Ser-79	38	13	0.17
Ile-80	38	5	0.97
Ala-81†	47	2	>0.99
Arg-82	47	12	0.42
Glu-83	35	11‡	0.24
Ile-84	17	1	>0.99
Tyr-85	38	13	0.16
Glu-86†	51	13	0.42
Met-87†	17	2	0.97
Tyr-88	52	9	0.88
Glu-89	51	13	0.39
Ala-90†	43	9	0.67
Val-91	46	7	0.86

*s = number sequenced; n = number of amino acids observed; $P_s(n+1)$ = probability that an additional allowed amino acid would have been observed in s sequences, if all $n+1$ amino acids are recovered with equal efficiency (i.e., if each functional amino acid is recovered at a frequency determined solely by the number of codons for that residue).

†At these positions, one amino acid was greatly overrepresented within the set of observed residues (its frequency was more than 3 standard deviations above the mean expected from a purely random distribution).

‡Lysine is also functional at position 83, but it was not observed in the 35 candidates sequenced following random mutagenesis.

DISCUSSION

By combining a targeted random mutagenesis procedure with a functional selection, we have been able to rapidly generate a large family of sequences for a single protein, each of which differs from the wild-type by at most three changes. At the level of selection we have employed, mutant proteins with greater than 5–10% of wild-type activity are recovered. By increasing or decreasing the stringency of the selection, different levels of tolerance to substitution would undoubtedly be observed at many positions. For example, randomization experiments on the hydrophobic core of the N-terminal domain of λ repressor have shown that when a lower level of activity is permitted, a greater number of substitutions are acceptable.[13] Furthermore, some changes may be acceptable in some sequence contexts but not in others. For example, in experiments requiring the same level of activity that we have used in this study, position 18 accepted only leucine when randomized alone, but tolerated alanine, valine, and isoleucine when other core residues were allowed to vary at the same time.[13]

A family of functional sequences if of enormous utility for studying various aspects of protein structure and function. For a protein of known structure, as in the present case, it is possible to gain additional insights into the roles of individual residues in the protein that are not possible with a single sequence. For example, the crystal structure of the N-terminal domain indicates that Asp-14 makes a charge-stabilized hydrogen bond with Arg-17 and a hydrogen bond with Ser-77. In the absence of a family of functional sequences, it is difficult to assess the importance of these interactions to the stability of the protein, or to infer the relative importance of the three residues. However, the results of the random mutagenesis experiments indicate that the interaction of Asp-14 and Arg-17 is important, since aspartate is conserved at position 14 and only arginine and lysine are observed at position 17. On the other hand, alanine as well as serine is allowed at position 77, suggesting that the hydrogen bond between Ser-77 and Asp-14 can be removed without seriously destabilizing the protein. Presumably, helix 1 is stabilized by the interaction between Asp-14 and Arg-17, which are spaced roughly one turn of the helix apart. In α-helical peptides, such stabilization has been observed with glutamate and lysine residues spaced one additional residue apart.[14]

Conserved Hydrophilic Surface Positions

We have now examined the informational content of over thirty residue positions in the N-terminal domain of λ repressor. We find that at positions that are buried in the structure there are severe limitations on which residues are tolerated. However, at surface sites, we observe two classes of positions: those that will accept essentially all types of residues, and those for which there is a much more limited pattern of substitution. Among the second class are positions that accept only highly hydrophilic residues. A similar class of surface residues, amounting to roughly one-third of the total residues, has been observed in the globins.[15] A particularly interesting example of this type in the N-terminal domain is position 11. Gln-11 is more than 50% exposed to solvent, and yet at this position only arginine, lysine, glutamine, and glutamate were recovered. This restriction in the substitution pattern at position 11 contrasts with the results at the two flanking positions. Positions 10 and 12 are also highly exposed to solvent but accept 14 and 10 different residues, respectively. In fact, the wild-type residue at position 12 is the highly hydrophobic leucine. Why should some surface positions show such a bias toward hydrophilic amino acids, while adjacent surface positions accept virtually any residue? One explanation for the tendency of some surface positions to accept only hydrophilic residues is that the overall surface must be reasonably polar for the protein to be soluble in aqueous solvent. However, some positions clearly must be more important than others in maintaining a polar surface; otherwise, one would expect that every surface position could accept a nonpolar side chain, provided most other surface positions remained hydrophilic. Perhaps certain positions, such as 11, are situated on the surface in such a way that they readily lead to aggregation of the protein when occupied by a suitably nonpolar side chain.

Another explanation for the tendency of some surface sites to remain hydrophilic involves the folding of the protein. If the specific pattern of hydrophobic and hydrophilic amino acids in a protein helps to determine whether or not the protein folds into a native-like structure, then changing this pattern by replacing certain hydrophilic with hydrophobic residues may present new folding pathways, leading to different structures or nonproductive folding intermediates.

Prolines in α-Helices

Prolines are rarely found within α-helices[16]; when they do occur, they cause a kink in the helix.[17] We find that proline is not tolerated in the middle of either helix 1 or helix 5, even at positions of low informational content. This failure to observe proline in the middle of these helices cannot be due to an underrepresentation of proline in the pool of random sequences, since proline appears frequently in unselected populations generated by random mutagenesis.[3]

Proline does occur frequently at the first position of α-helices,[18] and is allowed at the first position of each of the two helices examined here. In fact, position 78, the first residue in helix 5, tolerates only

proline. Why does this surface position require proline, whereas most surface positions tolerate several different residues? One possibility is that mutants at this position would be less stable than wild-type. The ϕ dihedral angle of proline is already constrained to an α-helical conformation; consequently, the entropic cost involved in restricting proline to a helical conformation would be lower than for other residues. In fact, a mutant in which Pro-78 is replaced by alanine is somewhat less thermally stable than wild-type (manuscript in preparation). However, this mutant is also unusually susceptible to intracellular proteolysis. This effect may reflect alterations in the kinetics of folding of the mutant protein or changes in the conformation of the folded or unfolded polypeptide chain.

Degeneracy in the Folding Process

Examination of over 30 residues in the N-terminal domain of λ repressor reveals that a surprisingly large number of positions are quite low in informational content. Nearly half of the positions examined in helix 1 and helix 5 will accept nine or more different residues, and only a few positions are absolutely conserved. This suggests that there is a high level of degeneracy in the folding process; that is, there are many possible sequences that will specify a protein that resembles the N-terminal domain of λ repressor. Moreover, if the criterion for neutral mutations were changed from the present requirement of 5–10% activity compared to wild-type, to the less stringent requirement that the protein simply be folded, the level of degeneracy would presumably be even higher.

Based on results presented here, it is possible to make a very rough estimate of the level of degeneracy of the folding information for the N-terminal domain. Making the simplifying assumptions that helix 1 and helix 5 are representative of the rest of the protein in terms of informational content, and that the changes observed at each position are independent of one another, we can estimate the number of allowed sequences for the entire protein. Multiplying together the numbers of functional amino acids observed at each position gives 4×10^{20} different sequences for the 30 residue positions examined thus far. On the one hand, this calculation overestimates the number of functional sequences, since changes at individual positions are less likely to be independent of one another as more positions are allowed to vary. Moreover, combining changes at several positions, each of which individually decreases the activity of the protein slightly, may result in a protein that is essentially nonfunctional. On the other hand, some changes which are not allowed when positions are randomized individually may be tolerated in other sequence contexts. Extrapolating to the rest of the protein indicates that there should be about 10^{57} different allowed sequences for the entire 92-residue domain. Clearly, this is an extraordinarily rough calculation, and we do not intend to suggest that we can accurately determine how many sequences would actually adopt a structure resembling the N-terminal domain of λ repressor. However, the calculation does indicate in a qualitative way the tremendous degeneracy in the information that specifies a particular protein fold. Nevertheless, the estimated number of sequences capable of adopting the λ repressor fold is still an exceedingly small fraction, about one in 10^{63}, of the total number of possible 92-residue sequences. A similar result has been obtained for cytochrome *c* based on phylogenetic sequence comparisons.[19]

The high level of degeneracy involved in protein folding suggests that the most fruitful approaches to structure prediction will concentrate on those residues that are informationally rich. Such residues are readily identified by examining families of related sequences. Information derived from sets of sequences has been useful in predicting regions of secondary structure for Arc repressor, a protein of unknown structure.[20] Furthermore, a knowledge of which residues are informationally rich may help in tertiary structure prediction, either as a starting point for the prediction process or as a means of evaluating possible structures generated by algorithms that use other criteria. Experiments that determine the informational content of residues in a protein of known structure serve as an important control for evaluating the success of such predictive schemes.

ACKNOWLEDGMENTS

We thank Dawn Parsell and Wendell Lim for providing plasmids used in this work, Steve Jordan and Carl Pabo for providing the coordinates of the repressor–operator complex, Jim Bowie for help with the solvent accessibility calculations, and Paul Schimmel for the use of his graphics system. Discussions with Dawn Parsell, Jim Bowie, and Wendell Lim contributed greatly to the progress of this work. This work was supported by NIH Grant AI-15706.

REFERENCES

1. Pabo, C.O., Lewis, M. The operator-binding domain of λ repressor: Structure and DNA recognition. Nature (London) 298:443–447, 1982.
2. Jordan, S.R., Pabo, C.O. Structure of the λ complex at 2.5 Å resolution: Details of the repressor-operator interactions. Science 242:893–899, 1988.
3. Reidhaar-Olson, J.F., Sauer, R.T. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. Science 241:53–57, 1988.
4. Amann, E., Brosius, J., Ptashne, M. Vectors bearing a hybrid *trp-lac* promoter useful for regulated expression of cloned genes in *Escherichia coli*. Gene 25:167–178, 1983.
5. Maniatis, T., Fritsch, E.F., Sambrook, J. "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor, New York: The Cold Spring Harbor Laboratory, 1982.
6. Hanahan, D. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. 166:557–580, 1983.
7. Sanger, F., Nicklen, S., Coulson, A.R. DNA sequencing

with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463–5467, 1977.
8. Hecht, M.H., Sturtevant, J.M., Sauer, R.T. Effect of single amino acid replacements on the thermal stability of the NH_2-terminal domain of phage λ repressor. Proc. Natl. Acad. Sci. U.S.A. 81:5685–5689, 1984.
9. Hecht, M.H., Hehir, K.M., Nelson, H.C.M., Sturtevant, J.M., Sauer, R.T. Increasing and decreasing protein stability: Effects of revertant substitutions on the thermal denaturation of phage λ repressor. J. Cell. Biochem. 29:217–224, 1985.
10. Nelson, H.C.M., Sauer, R.T. λ repressor mutations that increase the affinity and specificity of operator binding. Cell 42:549–558, 1985.
11. Hecht, M.H., Nelson, H.C.M., Sauer, R.T. Mutations in λ repressor's amino-terminal domain: Implications for protein stability and DNA binding. Proc. Natl. Acad. Sci. U.S.A. 80:2676–2680, 1983.
12. Gutman, G.A., Hatfield, G.W. Nonrandom utilization of codon pairs in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 86:3699–3703, 1989.
13. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. Nature (London) 339:31–36, 1989.
14. Marqusee, S., Baldwin, R.L. Helix stabilization by Glu $\cdots$ Lys^+ salt bridges in short peptides of *de novo* design. Proc. Natl. Acad. Sci. U.S.A. 84:8898–8902, 1987.
15. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold: Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
16. Chou, P.Y., Fasman, G.D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222, 1974.
17. Barlow, D.J., Thornton, J.M. Helix geometry in proteins. J. Mol. Biol. 201:601–619, 1988.
18. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α-helices. Science 240: 1648–1652, 1988.
19. Yockey, H.P. A calculation of the probability of spontaneous biogenesis by information theory. J. Theor. Biol. 67: 377–398, 1977.
20. Bowie, J.U., Sauer, R.T. Identifying determinants of folding and activity for a protein of unknown structure. Proc. Natl. Acad. Sci. U.S.A. 86:2152–2156, 1989.
21. Eisenberg, D., Weiss, R.M., Terwilliger, T.C., Wilcox, W. Hydrophobic moments and protein structure. Faraday Symp. Chem. Soc. 17:109–120, 1982.
22. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.