Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 2146–2150, March 1994
Biochemistry

Folded proteins occur frequently in libraries of random amino acid sequences

ALAN R. DAVIDSON AND ROBERT T. SAUER*

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

Communicated by Howard K. Schachman, December 16, 1993 (received for review September 21, 1993)

ABSTRACT A library of synthetic genes encoding 80- to 100-residue proteins composed mainly of random combinations of glutamine (Q), leucine (L), and arginine (R) has been expressed in *Escherichia coli*. These genes also encode an epitope tag and six carboxyl-terminal histidines. Screening of this library by immunoblotting showed that 5% of these QLR proteins are expressed at readily detectable levels. Three well-expressed QLR proteins were purified and characterized. Each of these proteins has significant α-helical content, is largely resistant to degradation by Pronase, and has a distinct oligomeric structure. In addition, one protein unfolds in a highly cooperative manner. These properties of the QLR proteins demonstrate that they possess folded structures with some native-like properties. The QLR proteins differ from most natural proteins, however, in being remarkably resistant to denaturant-induced and thermal-induced unfolding and in being relatively insoluble in the absence of denaturants.

What sequence features are required for proteins to form stable three-dimensional structures? Of the enormous number of possible amino acid sequences, are those sequences that specify folded proteins common or exceedingly rare? Theoretical studies suggest that many polypeptide sequences containing just two residue types, polar and nonpolar, could form folded structures with native-like properties (1). Moreover, mutagenesis studies (2) and comparisons of evolutionarily related proteins (3) indicate that the simple presence of hydrophobic residues at the correct positions in natural sequences may be one of the major determinants of protein structure. These observations can be interpreted to suggest that relatively little sequence information is required for proteins to adopt folded structures, which in turn supports the possibility that a significant fraction of all sequences may be capable of folding.

Here, we ask whether folded proteins can be isolated from libraries expressing random sequences composed predominantly of three residues: glutamine (Q), leucine (L), and arginine (R). Glutamine and leucine were chosen as representatives of hydrophilic and hydrophobic residues, respectively. Arginine was chosen as a charged residue and was included to increase protein solubility. We refer to the random genes and proteins as "QLR." In constructing the QLR genes, no attempt was made to design the secondary structure, turns, or any tertiary interactions in the encoded proteins (4). This paper describes the construction and screening of a library of random QLR proteins and the purification and characterization of three well-expressed QLR protein variants. Although they differ from natural proteins in several ways, each of the purified QLR proteins is shown to have a folded structure.

MATERIALS AND METHODS

Library Construction and Screening. Fig. 1*A* shows the basic structure of the synthetic genes used in this work. The randomized portion of each gene was constructed from three different 80- to 83-bp oligonucleotides. Each was synthesized on an Applied Biosystems 381A DNA instrument with a 10-bp self-annealing sequence at its 3′ end, allowing enzymatic second-strand synthesis with Sequenase (United States Biochemical) (5). Double-stranded cassettes were produced by restriction cleavage and were then ligated sequentially to produce the major portion of the gene (encoding 70 or 90 amino acids, depending on the presence of one or two central cassettes). This fragment was then ligated to a backbone fragment containing a P_{trc} promoter, a carboxyl-terminal tail containing the DYKDDDDK epitope tag (6) and six histidines, phage f1 and plasmid pBR322 origins of replication, the ampicillin-resistance gene, and the *lacI^q* gene. The backbone fragment of the vector was constructed from plasmid pDW239 (7), a derivative of pTrc99A (Pharmacia).

Constructs were transformed into *Escherichia coli* strain DP700 (8). Ampicillin-resistant colonies were blotted onto nitrocellulose filters and probed with a monoclonal antibody, M2 (IBI), directed against the epitope tag. Horseradish peroxidase-coupled anti-mouse antibody (Amersham) was used as a secondary probe, and the filters were developed with the enhanced chemiluminescence (ECL) system (Amersham). Cell lysis and immunological procedures were performed as described (9). Gene sequences were determined by DNA sequencing using the Sequenase version 2.0 kit (United States Biochemical). In total, 33 isolates that were positive in the primary screen and 64 unselected isolates were sequenced.

Protein Purification and Characterization. The presence of the carboxyl-terminal hexahistidine sequence allowed the QLR proteins to be purified by affinity chromatography on nickel-NTA agarose resin (Qiagen, Chatsworth, CA) (10). By the recommended protocol, the proteins were purified in 6.0 M guanidine hydrochloride (Gdn·HCl) to >95% homogeneity with yields of 1–3 mg per liter of culture. Dialysis of the purified QLR proteins from 6.0 M Gdn·HCl into 10 mM Tris·HCl, pH 8/0.2 mM EDTA resulted in precipitation. The precipitated proteins were suspended at high concentration in 6.0 M Gdn·HCl/50 mM potassium phosphate, pH 5.8, and diluted to lower Gdn·HCl concentrations by addition of phosphate buffer.

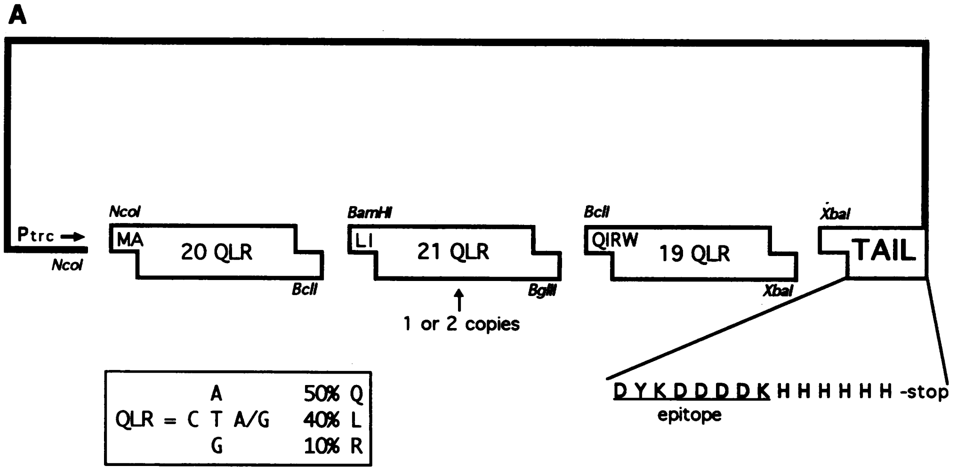
Digestions of the QLR proteins with Pronase (Boehringer Mannheim) were carried out in 4 M urea/50 mM potassium phosphate, pH 5.8, at 37°C for 1 hr. The QLR protein concentration was 370 μg/ml and the final Pronase concentration was 125 μg/ml.

Circular dichroism (CD) spectra were taken using an Aviv 60DS instrument at Gdn·HCl concentrations at which each QLR protein was soluble. Over concentrations ranging from

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

Abbreviation: Gdn, guanidine.

*To whom reprint requests should be addressed.



B

QLR-1:

MAQLLRQLRQLRQQLQQLLRHLLILLQLQQQQLLQLLLLLQRLQLQIRWLLQLLQQLQRQQQQRLQLLRL-tail

QLR-2 :

MAQLLLQLLLRQRQQQQQQQLLLIRQLLQQRQQLQQLLRQLRQLLQILLQLLQQQLRQLLLLLQQLQLQIRWLQQLLLRLQQLLLLQLQQLRL-tail

QLR-3 :

MAQLLLLQQQQQQLQQQQLQQLILQLQLRQLQLLLLLRLLQQLLQIRWLQLLLQLQQRLQQQQQQLQRL-tail

FIG. 1. QLR-protein expression vector and sequences. (*A*) Representation of plasmid backbone and oligonucleotide cassettes used to create the QLR library. In each cassette, 19–21 codons were randomized by using C at the first position; a mixture of A (50%), T (40%), and G (10%) at the second position; and an equal mixture of A and G at the third position. (*B*) Amino acid sequences of the QLR-1, QLR-2, and QLR-3 proteins.

2 μM to 30 μM, the molar ellipticities of the three proteins examined remained unchanged. Fractional helicity in 6.0 M Gdn·HCl was calculated by the formula ([θ_{222}] + 2340/30,300) (11). Gdn·HCl unfolding was monitored by changes in CD ellipticity at 222 nm at denaturant concentrations ranging from 4.0 to 8.0 M. Fluorescence emission spectra were measured in 6.0 M Gdn·HCl/50 mM potassium phosphate, pH 5.8, at protein concentrations from 4 to 20 μM, using an excitation wavelength of 280 nm. The measurements were made on a Perkin–Elmer LS-50 luminescence spectrometer at room temperature.

Analysis of Oligomeric Structure. Gel filtration experiments were performed with a 25-ml Pharmacia Superose 12 column equilibrated in 6.0 M Gdn·HCl/50 mM potassium phosphate, pH 5.8. Samples (200 μl) of QLR-1 (157 μM), QLR-2 (80 μM), and QLR-3 (99 μM) were run at a flow rate of 0.2 ml/min.

Centrifugation of the purified QLR proteins was carried out in a Beckman XL-A analytical ultracentrifuge. In each case, the buffer was 6.0 M Gdn·HCl/50 mM potassium phosphate, pH 5.8. Molecular weight values were calculated by fitting the data to the one-species function, $C(r) = C(a)\exp[\omega^2M(1 - v\rho)(r^2 - a^2)/2RT]$ using the program NONLIN for Macintosh (12, 13), where $C(r)$ is the solute concentration at radius r, $C(a)$ is the solute concentration at a reference distance a, ω is the angular velocity, R is the gas constant, T is the temperature, M is the molecular weight of the protein species, v is the partial specific volume of the protein, and ρ is the density of the solution (14). The partial specific volumes and solution density were calculated from standard formulas (15). Similar values of native molecular weight were obtained in experiments performed at initial protein concentrations of 43, 86, and 143 μM. The average molecular weights shown in Table 1 were calculated by averaging the results obtained at each of the three different protein concentrations and at two different speeds. QLR-1 was centrifuged at 20,000 rpm and 30,000 rpm, whereas QLR-2 and QLR-3 were centrifuged at 15,000 rpm and 20,000 rpm (rotor type An60Ti).

RESULTS

Construction and Screening of the QLR Expression Library. Using the vector and general strategy shown in Fig. 1*A*, we constructed genes encoding random mixtures of a polar residue (Q, 50%), a hydrophobic residue (L, 40%), and a small amount of a charged residue (R, 10%). These genes also encoded a few fixed residues at the cassette junctions, a single tryptophan to allow fluorescence studies, and a carboxyl-terminal tail containing an epitope tag (6) and six histidines to allow affinity purification (10). Genes with one copy of the central cassette encoded 84 amino acid residues; those with two copies encoded 107 residues.

The expression library was transformed into *E. coli*, and colonies expressing detectable levels of the QLR proteins were identified by immunoblotting using a monoclonal antibody directed against the epitope tag. Roughly 1% of the colonies were positive in this assay, but sequencing of randomly selected genes showed that the epitope was out of frame in about 80% of the genes. Hence, about 5% of the in-frame QLR proteins appear to be expressed in *E. coli*. Western blot assays (16) were used to test 88 epitope-positive isolates. Of these, 55 showed crossreacting protein of the expected molecular weight on SDS/polyacrylamide gels. Because unfolded proteins are usually degraded in *E. coli* (17), it seemed possible that the QLR proteins that were expressed at high levels in the cell would be stably folded.

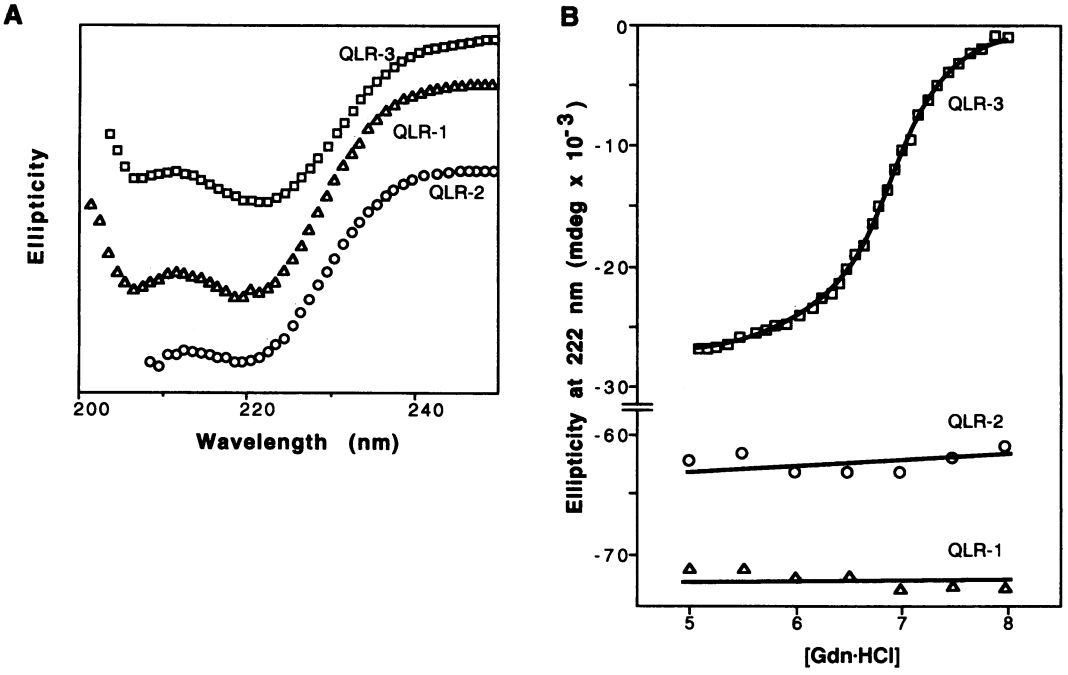


FIG. 2. CD studies of QLR proteins. (*A*) Spectra of QLR-1 (28 μM; 0.4 M Gdn·HCl), QLR-2 (32.5 μM; 6 M Gdn·HCl), and QLR-3 (32 μM; 0.7 M Gdn·HCl). The spectra are offset to allow each to be seen clearly; thus, there are no units on the ellipticity scale. (*B*) Ellipticity of QLR proteins at 222 nm as a function of Gdn·HCl concentration. The protein concentrations were 4.25 μM (QLR-1), 3.1 μM (QLR-2), and 4.0 μM (QLR-3). All CD experiments were performed at 25°C.

CD and Fluorescence Properties of Purified QLR Proteins. Three of the highly expressed QLR proteins (designated QLR-1, QLR-2, and QLR-3; see Fig. 1*B*) were purified. The purified proteins were poorly soluble in the absence of chaotrophic agents. For example, QLR-1 and QLR-3 required 0.4–0.7 M Gdn·HCl for solubility at a concentration of 30 μM, whereas QLR-2 required Gdn·HCl concentrations greater than 4.0 M for solubility at this protein concentration. As a consequence, the biochemical studies described below were performed in buffers containing Gdn·HCl or urea under conditions where the QLR proteins were soluble.

As shown in Fig. 2*A*, the CD spectra of all three QLR proteins indicate the presence of α-helical secondary structure, with minima at 208 and 222 nm (18). The α-helical structure of the QLR proteins is not unexpected, given the high helical propensities of glutamine, leucine, and arginine (19). However, the stability of this structure is surprising. In 6.0 M Gdn·HCl, the fractional α-helix contents calculated from the molar ellipticities at 222 nm are 70% for QLR-1, 60% for QLR-2, and 32% for QLR-3 (Table 1). The α-helical structure of QLR-3 undergoes cooperative unfolding as the Gdn·HCl concentration is raised, while QLR-1 and QLR-2 show no significant unfolding, even at the highest Gdn·HCl concentrations (Fig. 2*B*). Thermal melts of the three proteins were also carried out in 6.0 M Gdn·HCl. None of the proteins showed significant loss of α-helical content up to 90°C, the highest temperature tested.

In fluorescence experiments, the emission maximum of the single tryptophans in QLR-1 and QLR-2 was 348 nm, indicating that the tryptophans in these proteins are solvent exposed. By contrast, the emission maximum for QLR-3 was 337 nm. This indicates that the tryptophan side chain is partially buried in a hydrophobic environment (20).

Resistance to Proteolysis *in Vitro*. Protease resistance is a hallmark of folded proteins (21). QLR-1, QLR-2, and QLR-3 were treated with Pronase, a nonspecific protease. Fig. 3 shows that the QLR proteins were relatively resistant to proteolysis as monitored by CD ellipticity at 222 nm. A control protein, the λ phage repressor's amino-terminal domain, was digested rapidly under the same conditions, providing a positive control that the protease was active under the conditions of the assay. Electrophoresis of the QLR proteins after proteolysis indicated that they had been digested to species of 60–80% of their original size (data not shown). The proteolytic products had lost the carboxyl-terminal epitope, as assayed by Western blot analysis, but retained almost all of their α-helical character. These results suggest that the QLR proteins contain a structured, protease-resistant core together with unstructured, protease-sensitive regions that include the carboxyl-terminal tail.

Oligomeric Structure of QLR Proteins. Several lines of evidence indicate that the QLR proteins possess distinct oligomeric structures. In gel filtration experiments (Fig. 4*A*), the QLR proteins migrated primarily as single species but

Table 1. Properties of purified QLR proteins

	Composition,* %			Monomer	Average	Fractional helicity,‡
Protein	Q	L	R	M_r	M_r†	%
QLR-1	36	43	13	10,585	69,600 (±4%)	70
QLR-2	39	44	11	13,368	40,100 (±7%)	60
QLR-3	46	40	7	10,492	147,000 (±9%)	32

*These percentages do not include the tail.
†These values were determined by analytical equilibrium ultracentrifugation.
‡The fractional helicity was calculated as described in *Materials and Methods*.

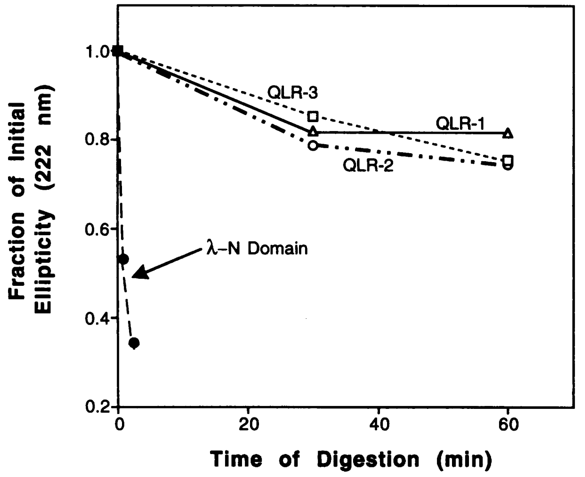


FIG. 3. Proteolysis of QLR proteins. The change in CD ellipticity at 222 nm is plotted as a function of time of digestion with Pronase. The same experiment using λ repressor's N-terminal domain was included as a control to show that Pronase is active under the conditions used. In control experiments not shown, we also determined that peptide bonds formed by glutamine, arginine, and leucine are not inherently resistant to Pronase cleavage.

were eluted at positions expected for species larger than monomers. In polyacrylamide gel electrophoresis (0.9 M acetic acid/8 M urea; the QLR proteins maintained their secondary structure under these conditions as assayed by CD), QLR-2 and QLR-3 migrated as single bands of larger than monomer molecular weight; QLR-1 was somewhat less homogeneous in these experiments but still migrated as one predominant band (data not shown). These experiments indicate that the purified QLR proteins exist as multimers. If these multimers were nonspecific aggregates, broader peaks and poorly defined bands would have been expected in the above experiments. As a result, these data suggest that the QLR multimers have discrete structures.

The quaternary structure of the QLR proteins was further studied by equilibrium centrifugation experiments (Fig. 4 *B–D*). The distribution profiles of the QLR-2 and QLR-3 proteins after centrifugation fit well to a single species model, and their calculated molecular weights (Table 1) suggest that QLR-2 is a trimer and QLR-3 is a tetradecamer. The high average molecular weight of the QLR-1 protein also implies the presence of oligomers, although the relatively poor fit of the experimental data to the theoretical single-species function suggests that this protein may not exist as a single stable oligomeric species. Because QLR-1 was also less homogeneous than the other QLR proteins in both gel filtration and gel electrophoresis experiments, we believe that it probably forms two or more oligomeric species.

DISCUSSION

The purified QLR proteins studied in this work share many properties with natural proteins. These include stable, protease-resistant α-helical secondary structure, discrete quaternary structure, and, in the case of QLR-3, a cooperative

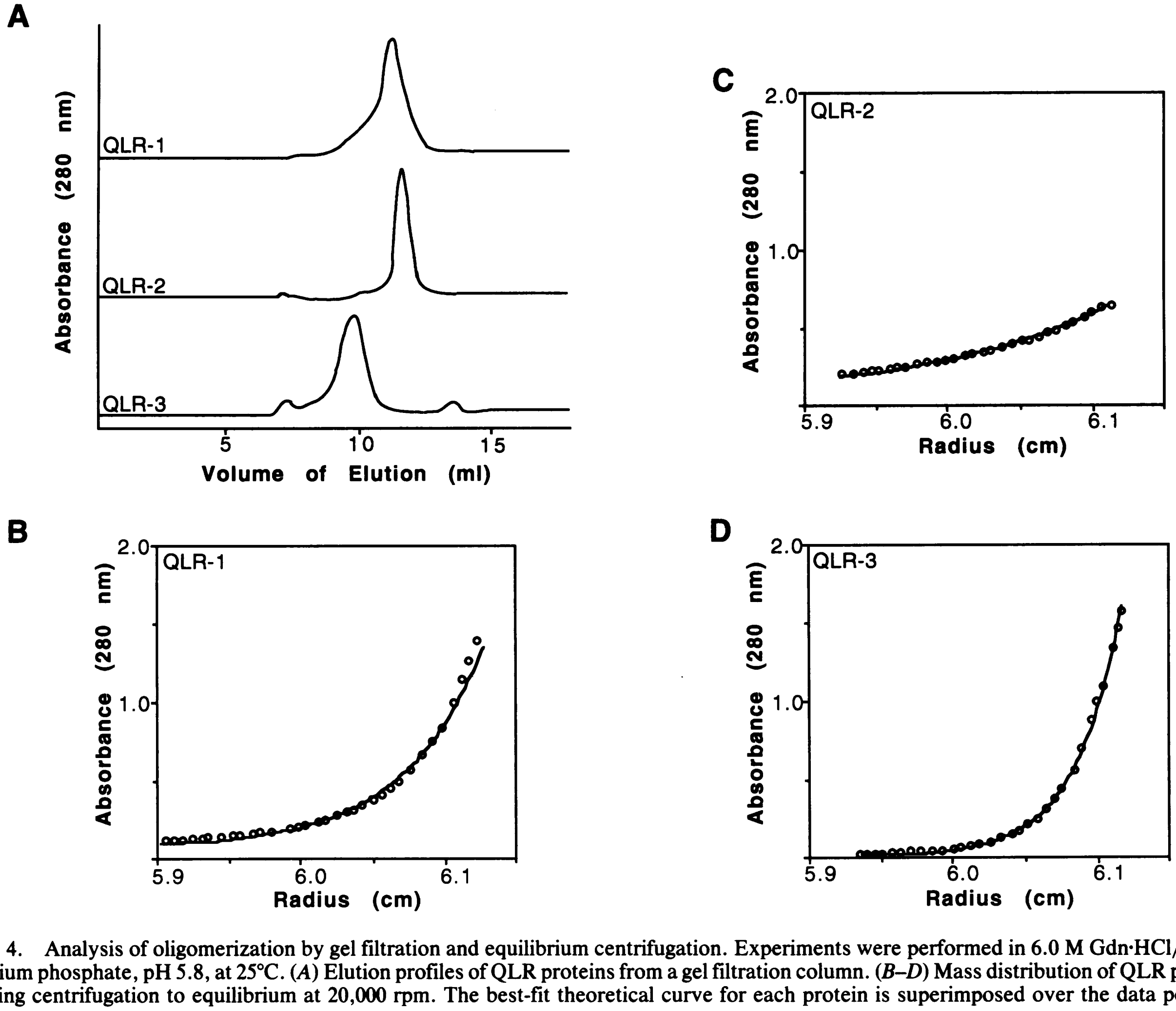


FIG. 4. Analysis of oligomerization by gel filtration and equilibrium centrifugation. Experiments were performed in 6.0 M Gdn·HCl/50 mM potassium phosphate, pH 5.8, at 25°C. (*A*) Elution profiles of QLR proteins from a gel filtration column. (*B–D*) Mass distribution of QLR proteins following centrifugation to equilibrium at 20,000 rpm. The best-fit theoretical curve for each protein is superimposed over the data points.

unfolding transition. Although we have no direct assays for tertiary structure, it is difficult to imagine that the QLR proteins could self-assemble into stable oligomers in the absence of some stable tertiary interactions. Further, the QLR proteins showed significant variations in properties such as helical content, oligomeric structure, tryptophan fluorescence, and stability, which implies that each QLR protein has a different structure in spite of their similar overall compositions.

Despite the many similarities noted above, the QLR proteins that we have studied differ from natural proteins in at least two significant ways. First, the QLR proteins show extraordinary resistance to Gdn·HCl and thermal denaturation. We know of no natural proteins that retain their secondary structure in the presence of 6.0 M Gdn·HCl at 90°C. Second, the QLR proteins require some denaturant for solubility, whereas most natural proteins are soluble in aqueous buffers. It seems likely that the insolubility of the QLR proteins in aqueous buffers arises because these proteins are, in some sense, too hydrophobic, leading to nonspecific aggregation in the absence of agents which reduce the magnitude of the hydrophobic effect. The high hydrophobic content of the QLR proteins may also, at least in part, account for their extreme stability.

Without detailed structural studies, it is not possible to determine whether the QLR proteins have novel folds. Moreover, we have no information concerning possible dynamic aspects of the QLR structures. In terms of certain properties, the QLR proteins resemble α_4, a four-helix bundle protein designed by Regan and DeGrado (22). α_4 is also very stable to denaturation and has been shown to possess native-like features such as helical secondary structure, compactness, and a cooperative unfolding transition. NMR studies indicate that the α-helical backbone of α_4 is relatively well structured but show that the hydrophobic core, which is composed solely of leucines, has significant molten character (23). The unusually high stability of α_4 has been postulated to arise from the molten character of the core, which would reduce the entropy loss associated with core packing (23). Because the cores of the QLR proteins are also likely to be composed almost exclusively of leucines, a similar model could be advanced to explain their unusually high stabilities. In drawing attention to the similarities between α_4 and the QLR proteins, we do not wish to imply that the QLR proteins are likely to have four-helix bundle structures. It is also important to remember that α_4 was the result of a sophisticated design process (22), whereas the QLR proteins were isolated from a library of random sequences.

Although there are many unanswered questions about the properties and structures of QLR-1, QLR-2, and QLR-3, it is important not to lose sight of the broader issues at hand. The major finding of this work is that proteins with folded structures and some native-like properties are remarkably common in libraries of random QLR sequences. Many of the in-frame QLR proteins in the library were expressed at levels as high as those of QLR-1, QLR-2, and QLR-3, and it seems likely that these other QLR proteins will also have folded structures. Overall, we believe that our results are consistent with theoretical studies predicting that a significant fraction of random sequence proteins should fold into unique structures under native conditions (1, 24).

If stably folded molecules with distinct properties can be generated at relatively high frequencies in randomized protein libraries, then it should be possible to screen such libraries for proteins with specific binding activities or even enzymatic activities. This could potentially provide a means for isolating proteins with novel and useful properties. To achieve these goals, it may be advantageous to be able to isolate soluble, monomeric proteins. It seems likely that such proteins would be more native-like and more amenable to structural characterization. Libraries constructed by using different frequencies of the same amino acids used here (e.g., more arginine and less leucine) or with different types or numbers of amino acids might contain a higher frequency of soluble, monomeric proteins and ultimately prove to be useful for the generation of functional molecules.

We thank members of our lab for helpful comments and suggestions and P. Kim, C. Pabo, and B. Tidor for comments on the manuscript. We thank K. Lumb and the P. Kim lab for aid in the use of their analytical ultracentrifuge and D. Waugh for the gift of pDW239. This work was supported by National Institutes of Health Grant AI15706 and by a Canadian Medical Research Council fellowship to A.R.D.

1. Dill, K. A. (1990) *Biochemistry* **29**, 7133–7155.
2. Sauer, R. T. & Lim, W. A. (1992) *Curr. Opin. Struct. Biol.* **2**, 46–51.
3. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196**, 199–216.
4. DeGrado, W. F., Raleigh, D. P. & Handel, T. (1991) *Curr. Opin. Struct. Biol.* **1**, 984–993.
5. Oliphant, A. R., Nussbaum, A. L. & Struhl, K. (1986) *Gene* **44**, 177–183.
6. Hopp, T. P., Prickett, K. S., Price, V. L., Cerretti, D. P., Urdal, D. L. & Conlon, P. J. (1988) *Bio/Technology* **6**, 1204–1210.
7. Waugh, D. S. & Sauer, R. T. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 9596–9600.
8. Parsell, D. A. & Sauer, R. T. (1989) *Genes Dev.* **3**, 1226–1232.
9. Breyer, R. M. & Sauer, R. T. (1989) *J. Biol. Chem.* **264**, 13355–13360.
10. Hochuli, E., Dobeli, H. & Schachter, A. (1987) *J. Chromatogr.* **411**, 177–184.
11. Chen, Y.-H., Yang, J. T. & Martinez, H. M. (1972) *Biochemistry* **11**, 4120–4131.
12. Brenstein, R. J. (1989) NONLIN *for Macintosh* (Robelko Software, Carbondale, IL), Version 0.9.8b4.
13. Johnson, M. L. & Frazier, S. (1985) *Methods Enzymol.* **117**, 301–342.
14. van Holde, K. E. (1971) *Physical Biochemistry* (Prentice–Hall, Englewood Cliffs, NJ), pp. 110–113.
15. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. (1992) in *Analytical Ultracentrifugation in Biochemistry and Polymer Science*, eds. Harding, S., Rowe, A. & Horton, J. (Royal Soc. Chem., Cambridge, U.K.), pp. 90–125.
16. Tsang, V. C., Peralta, J. M. & Simons, A. R. (1983) *Methods Enzymol.* **92**, 377–391.
17. Parsell, D. A. & Sauer, R. T. (1989) *J. Biol. Chem.* **264**, 7590–7595.
18. Greenfield, N. & Fasman, G. D. (1969) *Biochemistry* **8**, 4108–4116.
19. Blaber, M., Zhang, X.-j. & Matthews, B. W. (1993) *Science* **160**, 1637–1640.
20. Teale, F. W. J. (1960) *Biochem. J.* **76**, 381–388.
21. Price, N. C. & Johnson, C. M. (1989) in *Proteolytic Enzymes: A Practical Approach*, eds. Beynon, R. J. & Bond, J. S. (IRL, Oxford, U.K.), pp. 163–180.
22. Regan, L. & DeGrado, W. F. (1988) *Science* **241**, 976–978.
23. Handel, T. M., Williams, S. A. & DeGrado, W. F. (1993) *Science* **261**, 879–885.
24. Shakhnovich, E. I. & Gutin, A. M. (1990) *Nature (London)* **346**, 773–775.