Core-Directed Protein Design. II. Rescue of a Multiply Mutated and Destabilized Variant of Ubiquitin†

Michael D. Finucane and Derek N. Woolfson*

Centre for Biomolecular Design and Drug Development, School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, U.K.

Received April 1, 1999; Revised Manuscript Received June 21, 1999

ABSTRACT: We have applied the method described in the preceding paper [Finucane, M. D., et al. (1999) *Biochemistry 38,* 11604−11612], namely, stability-based selection using phage display, to explore the sequence requirements for packing in the hydrophobic core of ubiquitin. In contrast to the parent protein, which was a structurally compromised mutant, the selected variants could be overexpressed and purified in yields for structural studies. In particular, CD and NMR measurements showed that the selectants folded correctly to stable native-like structures. These points demonstrate the utility of our core-directed method for stabilizing and redesigning proteins. In addition and in contrast to foregoing studies on other proteins, which suggest that hydrophobic cores permit substitutions provided that hydrophobicity and core volumes are generally conserved, we find that the core of ubiquitin is surprisingly intolerant of amino acid substitutions; variants that survived our selection showed a clear consensus for the wild-type sequence. It is probable that our results differed from those from other groups for two reasons. First, ubiquitin may be unusual in that it has strict sequence requirements for its structure and stability. We discuss this result in light of sequence conservation in the eukaryotic ubiquitins and proteins of the ubiquitin structural superfamily. Second, our mutants were selected solely on the basis of stability, in contrast to the other studies that rely on function-based selection. The latter may lead to proteins that are more plastic and tolerant of substitutions.

The reliable design of proteins de novo represents a major challenge to modern-day molecular biology. Approaches taken in this effort can be subdivided into (1) rational design and (2) the selection of proteins from combinatorial libraries. Rational design requires clear rules that link sequence and structure, which are only available at present for simple secondary structures and the more straightforward protein folds such as coiled coils (*1*−*3*). Thus, success in this area has been limited to relatively simple motifs.

To facilitate the design and redesign of more ambitious and complex protein architectures, researchers have turned to combinatorial approaches where proteins with desired properties are selected from libraries of mutant proteins. The selection processes can be either computer-assisted, which make assumptions about protein structure and stability, or experimental, which demand clear strategies for sampling the potentially vast protein sequence space.

To date, experimental selection procedures have largely been constrained by the need to select proteins on the basis of some function, i.e., binding or enzyme activity. Motivated by this, we developed the approach outlined in the preceding paper. In our method, stable, folded proteins are selected from random libraries on the basis of their ability to fold and resist proteolysis. We have tested this using ubiquitin as a model protein, the known, native structure (Figure 1) of which is extremely stable (*4*, *5*). First, we destabilized the native structure by making multiple mutations in the core, and then we restored stability by repacking the core using our method. There were considerable advantages in starting from an unstable parent structure. First, it provided an acid test of the method, as we knew that a more stable variant was accessible; second, during any mutagenesis, a proportion of nonmutant parent clones are expected in the initial library because of inherent inefficiencies in the processes. Phage display is essentially a competitive technique, and thus, occasionally mutants with enhanced properties (in our case stability) may be lost when faced with large numbers of competing mutants which may be equally or even less fit. Thus, by ensuring that the parental, starting ubiquitin sequence was relatively unstable, we aimed to deselect any unmutated parent rapidly and early in panning. This would favor the selection of other mutants on the basis of stability alone.

Ubiquitin was destabilized by replacing seven core residues (positions 3, 4, 13, 15, 17, 26, and 30) with leucine as polyleucine cores are known to increase the conformational heterogeneity of proteins and, thus, lower the conformational specificity and stability (*6*−*8*). In addition, in AL_7 the wild-type (WT)[1] methionine at position 1 was replaced with

† This work was funded by the BBSRC (Grant B04756). The CD instrument used in this study was funded by the Wellcome Trust (Grant 047029).

* Corresponding author. Telephone: +44 1273 678214. Fax: +44 1273 678433. E-mail: dek@biols.sussex.ac.uk.

[1] Abbreviations: AL_7, M1A/I3L/V5L/I13L/V17L/V26L/I30L multiple mutant of ubiquitin; CD, circular dichroism; NMR, nuclear magnetic resonance; NTA, nitrilotriacetic acid; WT, wild-type; UBQ, ubiquitin.

10.1021/bi990766f CCC: $18.00
Published on Web 08/17/1999

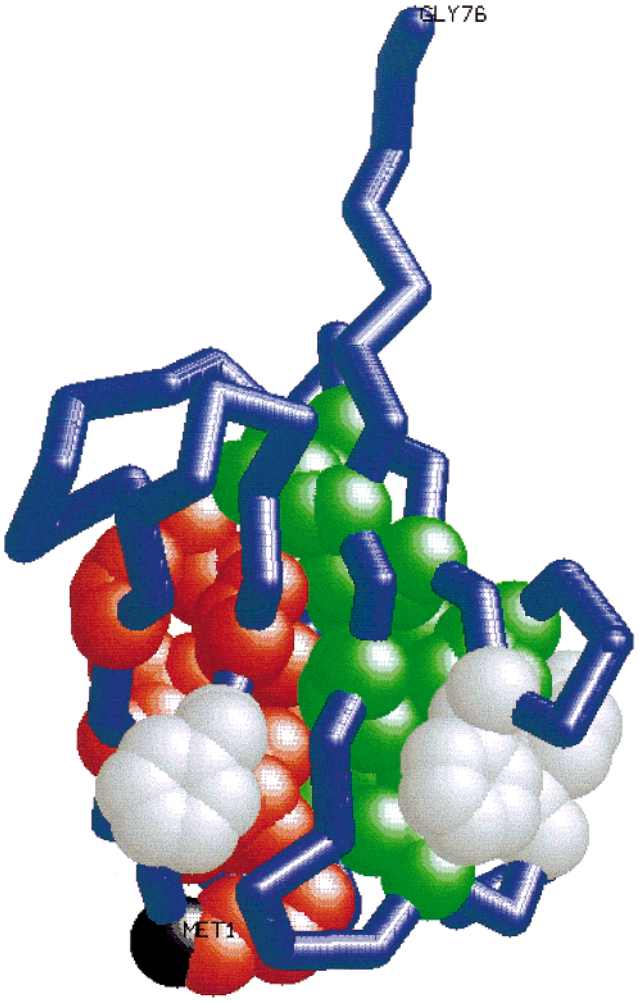

FIGURE 1: Three-dimensional structure of ubiquitin (*51*). The backbone trace is blue with the terminal residues, methionine-1 and glycine-76, in black. The hydrophobic core residues, 3, 5, 13, 15, 17, 23, 26, 30, 41, 43, 50, 56, 61, 67, and 69, are ≥95% shielded from solvent, and methionine-1 is 78% buried. These residues form two clusters shown in red and green. The N-terminal cluster, residues 3, 5, 13, 15, 17, 26, and 30, shown in red cements interactions between the β-hairpin and the α-helix and was the focus of this study. Potential chymotrypsin cleavage sites are gray; these lie on the surface of the protein, but are 66, 77, and 83% removed from solvent, respectively. This figure was created using RASMOL (*52*).

alanine to preserve the WT core volume. Consistent with our notion that AL_7-UBQ would be destablised compared with the WT, the mutant did not express to significant levels. We used the AL_7 construct to generate a library of ubiquitin mutants with partially randomized cores. The initial library contained some parental (AL_7) codons, but no other unexpected bias. Challenging the phage-display library with protease removed the parental sequences completely. Again, this was consistent with AL_7-UBQ being destabilized. Moreover, for the selected sequences, we noted a clear consensus for WT residues; of the eight positions allowed to vary in the initial library, seven preferentially recovered the amino acid chosen in nature. Several of the selected mutants were overexpressed and characterized. In these, we included sequences with the smallest and the largest predicted core volumes. All of the proteins that were analyzed had cooperative thermal melts and NMR spectra indicative of well-folded structures very similar to that of the WT protein. Although none of the selected proteins were as stable as WT ubiquitin, all had melting temperatures within 10 °C of this naturally hyperstable protein. Consistent with foregoing studies on other systems (*9*−*12*), therefore, ubiquitin does accommodate quite large changes within the core and remain folded. However, given the clear near-WT consensus that we recovered and the fact that none of the selected mutants were as stable as ubiquitin itself, we argue that the best solution to the core-packing problem for ubiquitin is the natural WT sequence, or residue combinations extremely close to it.

MATERIALS AND METHODS

Materials. Ubiquitin from bovine red blood cells was purchased from the Sigma Chemical Co. The sequences of all mammalian ubiquitins are identical (Table 3); hence, the sequences of this commercially obtained protein and our recombinant protein are the same apart from the linker regions appended to the termini of the latter. Other materials are described in the preceding paper (*54*).

Construction of the Library. A library of ubiquitin mutants was prepared in which eight core positions were simultaneously randomized to combinations of F, I, L, M, and V. A single-stranded template for mutagenesis was prepared in the *dut1 ung1* strain of *Escherichia coli* CJ236 from the UBQ-AL_7-phage construct described by Finucane et al. (*54*). A 128 bp mutagenic oligonucleotide was constructed using recursive PCR (*13*) from the four overlapping degenerate oligonucleotides: 5′-CTG CAG GGA *DTS* CAG *DTS* TTC *DTS* AAG ACC CTG ACC GGC AAG ACC, 5′-CTC GAT GGT GTC ACT GGG CTC *SAH* CTC *SAH* GGT *SAH* GGT CTT GCC GGT CAG G, 5′-CC AGT GAC ACC ATC GAG AAC *DTS* AAG GCC AAG *DTS* CAG GAT AAA GAG GGC ATC C, and 5′-TGA TCA GGG GGG ATG CCC TCT TTA TCC TG, where D is A, T, or G, H is A, T, or C, and S is G or C. The resulting product was then annealed to the single-stranded template DNA using a modified Kunkel mutagenesis procedure (*14*). We used the annealing/extension protocol described for "sticky-feet" mutagenesis (*15*). The double-stranded primer was denatured at 92 °C in the presence of the single-stranded DNA template and *Taq* polymerase; the temperature was lowered at a rate of 25 °C/min during which time the primer binds to, extends slightly, and increases its complementarity to the template. As a result, binding of the active primer strand to its cognate DNA is favored relative to that to its partner. The resulting double-stranded DNA was electroporated into electrocompetent XL1-blue *E. coli* cells, and untransformed cells were selected against using the antibiotic carbenicillin.

Preparation of Phage. Phage were prepared as described in the preceding paper (*54*).

Panning of the Library. Library phage (0.1 mL, at titers of 4×10^9, 4×10^{10}, 2×10^{12}, and 1×10^{11} cfu/mL for each of four rounds of panning, respectively) were mixed with 0.8 mL of sonication buffer [50 mM Na_2HPO_4 and 300 mM NaCl (pH 7.8)] and 0.25 mL of Ni−NTA agarose beads (Qiagen). Phage were bound to the beads for 10 min. The mixture was then transferred to a microcentrifuge tube filter (cutoff of 10 μm) and washed 14 times with 0.75 mL of sonication buffer, and the supernatant was removed by centrifugation. After washing, the beads were resuspended

in 1.25 mL of sonication buffer, and 40 μL of 1 mM chymotrypsin was added to 1 mL of this suspension. This was left for 100 s at 37 °C, before removing the supernatant by centrifugation. The beads were then washed four times as described before, prior to elution of the remaining bead-bound phage with sonication buffer containing 250 mM imidazole. The recovered phage were then amplified by infection into XL1-blue cells, followed by superinfection with helper phage, overnight growth, and precipitation from the growth supernatant with 20% polyethylene glycol/2.5 M NaCl. These phage were resuspended in 1 mL of sonication buffer, briefly clarified by centrifugation, and used as input for the next round of panning.

Overexpression of Proteins. Plasmid DNA was transformed into calcium-competent *E. coli* HB2151 cells, and grown in 2TY medium (Gibco-BRL). Isopropyl β-D-thiogalactoside was added a final concentration of 1 mM, and cells were harvested 16 h later. Our constructs encoded the leader sequence for pIII prior to the genes for the histidine-tagged ubiquitins. This leader sequence directs expressed proteins to the periplasm where the leader is cleaved. Periplasmic proteins were recovered by rapidly freezing the cells as a suspension in 10 mM Tris buffer (pH 8), followed by rapid thawing. The cells were then recentrifuged, and the supernatant was collected. Proteins were purified from this crude supernatant using Ni-NTA agarose beads following the manufacturer's directions. Proteins were concentrated using an Amicon 8MC microultrafiltration system using YM3 membranes (M_r cutoff of 3000), followed by buffer exchange into 50 mM phosphate buffer at pH 2.2. The protein identities were confirmed by mass spectrometry.

Nuclear Magnetic Resonance. ^{1}H NMR spectra were recorded at Sussex on a Bruker AM-500 and at NIMR Mill Hill on a Varian UNITYplus-500 spectrometer. Spectra were acquired with a sweep width of 5500 Hz, an acquisition time of 0.745 s, and a relaxation delay of 1.9 s. The solvent signal was suppressed by presaturation with a power of 3 dB during the relaxation delay. Spectra were referenced to internal tetrasilylpropionate at 0 ppm. Samples for NMR were prepared as described above, and ^2H_2O was added to a final concentration of 10% (v/v).

Circular Dichroism. CD spectra were recorded on a JASCO J-715 spectropolarimeter calibrated with 10-camphorsulfonic acid. Samples were prepared as described above and diluted to $\leq$50 μM protein with 50 mM sodium phosphate (pH 2.2). Spectra were recorded in 1 or 5 mm quartz cells by averaging the signal for 16 s at 1 nm intervals in the range of 250−195 nm, and a 1 nm bandwidth was used. Buffer baselines were subtracted from all spectra. Thermal unfolding was monitored at 200 nm to measure the development of the negative band characteristic of random coils. In these experiments, the temperature was ramped at a rate of 1 °C/min and the signal recorded every degree, averaging over 16 s. The temperature was controlled by a Peltier device and monitored using an external thermocouple.

Statistical Analysis. We used the standard error of proportion (*16*) to test if wild-type ubiquitin residues were selected from the panned library at statistically significant levels above the expected rate. *Z*-scores were calculated to compare the proportion of WT residues retrieved at a particular position ($P_{WT,pos}$) with the proportions expected had no selection occurred ($P_{E,pos}$). The latter values were calculated assuming a 1:1:1:1:2 ratio of occurrence of amino acids F, I, L, M, and V, which follows from the biased codon used to generate the initial library.

$$Z = \frac{P_{WT,pos} - P_{E,pos}}{\sqrt{\frac{P_{E,pos}(1 - P_{E,pos})}{N}}}$$

N is the number of clones that were sequenced, which was 31. Significant *Z*-scores, those that indicate a selection of WT residues, lie outside the ranges of −1.96 to 1.96 (the 95% confidence interval) and −2.58 to 2.58 (99% confidence interval).

RESULTS

Selection Studies. We used mutagenesis to create a library of hydrophobic-core variants of ubiquitin followed by a proteolysis-based selection to recover protease-resistant mutants from this pool. To avoid selecting a high background of the stable WT sequence, we created a destabilized mutant first by changing eight of ubiquitin's core positions to obtain the AL_7 mutant. Plasmid DNA for this was then used as a template to prepare the library. Using the degenerate codon {GAT}T{GC}, the eight sites targeted in AL_7 were replaced by combinations of the five hydrophobic residues F, I, L, M, and V.

We estimated that the transformed library contained 1.17×10^7 independent transformants. Thirty of these were sequenced (Table 1). These contained 48% parental, AL_7, codons, instead of the 10% expected if mutagenesis had been 100% efficient. This indicated that there was a substantial proportion of nonmutated sites. However, these could be largely ascribed to 40% of the sequenced clones being fully parental and the remainder of the library exhibiting no bias. Thus, the library contained ≈60% individual mutant clones (7×10^6 clones, which represented an ≈20-fold redundancy of the $\approx 4 \times 10^5$ potential mutants). For the sequenced clones, all possible (FILMV) residues were present at each of the targeted positions, except position 3, at which phenylalanine was not found in any of the 30 clones that were sequenced. In addition, two codons were discovered that were not encoded by the degenerate oligonucleotide design. These could have occurred by incorporation of incorrect bases either during the synthesis of the oligonucleotide by PCR or during the mutagenesis itself.

The library was displayed on phage and panned as outlined in the Materials and Methods. Phage from the fourth and final round of panning were characterized first by DNA sequencing. Of the selected clones that were sequenced, ≈40% had identical DNA sequences. This was for a mutant in which residues 6−66 of ubiquitin were deleted, leaving only the sequences for the first and last β-strands, and the C-terminal tail (Figure 1). It was surprising to recover this mutant as it contained a potential cleavage site at phenylalanine-4, and we would have expected it to be unfolded. It is possible, however, that this deletion mutant had a selective advantage over full-length clones because it might have inhibited phage−pilus interactions to a lesser extent than intact protein-phage. Thus, although some deletion mutants may have been cleaved during the selection step, any of them

Table 1: Codon Usage in and Amino Acid Composition of the Ubiquitin Libraries before and after Selection

Sequence Position[a]								Translated sequence[b]	Vol. (Å^3)	Δ[c]	#[d]	T$_M$[f] (°C)
1	3	5	13	15	17	26	30					
							Before Selection					
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	*TTG*	*CTC*	ALLL***L***LLL	740	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	ATC	*CTC*	ALLL***L***LIL	741	7		
GCG	*CTC*	*TTG*	*CTC*	*CTG*	*TTG*	ATG	ATG	ALLL***L***LMM	746	7		
GCG	*CTC*	*TTG*	ATG	GTC	GTG	*TTG*	*CTC*	ALLMV***V***LL	691	7		
GCG	*CTC*	*TTG*	GTC	ATC	ATC	ATC	GTG	ALLVIIIV	691	8		
GCG	TTG	GTG	TTC	GTC	ATC	GTC	ATG	AL***V***FVI***V***M	701	6		
TTC	GTG	ATG	TTC	TTG	TTC	ATG	ATC	FVMF***L***FM***I***	902	6		
ATC	GTG	TTC	ATG	GTC	*TTG*	TTC	GTG	IVFMVLFV	812	8		
ATC	GTC	TTC	GTC	GTC	TTC	ATG	TTG	IVFVVFML	812	8		
ATC	GTG	*TTG*	ATG	TTC	TTC	ATG	GTC	IVLMFFMV	841	8		
TTG	TTG	TTC	ATC	GTC	ATC	ATG	TTC	LLF***I***VIMF	865	7		
TTG	ATG	ATC	ATG	GAG	*TTG*	*TTG*	*CTC*	LMIMELLL	810	8		
ATG	TTG	ATC	ATG	ATC	TTC	*TTG*	*CTC*	***M***LIMIFLL	859	7		
TCG	GTC	*TTG*	TTC	GTC	*TTG*	*TTG*	*CTC*	SVLFVLLL	730	8		
GTG	ATC	*TTG*	TTG	ATG	*TTG*	*TTG*	GTG	V***I***LLMLLV	768	7		
GTC	ATG	ATG	GTG	TTC	ATG	ATG	ATG	VMMVFMMM	814	8		
GTG	GTC	*TTG*	GTG	ATG	ATC	GTG	TTG	VVLVMI***V***L	716	7		
GTG	GTC	ATG	TTG	TTG	GTG	*TTG*	*CTC*	VVML***LV***LL	741	6		
GTG	GTG	GTG	TTG	GTC	GTC	*TTG*	*CTC*	VV***V***LV***V***LL	686	6		
TTC	ATC	*TTG*	ATC	TTG	GTC	*TTG*	GTC	F***I***L***ILV***LV	801	4		
TTC	ATC	GTC	TTG	TTC	GTC	GTG	GTG	F***IV***LF***VV***V	783	4	**1**	54.5
TTC	TTG	GTG	TTC	GTG	GTC	GTC	ATC	FL***V***FV***VVI***	783	4		
TTC	GTC	ATC	TTC	TTG	GTG	*TTG*	*CTC*	FVIF***LV***LL	835	6		
TTC	GTC	GTG	TTC	TTC	ATG	*TTG*	ATC	FV***V***FFML***I***	873	6	**2**	49
TTC	GTC	GTG	TTG	GTG	GTC	*TTG*	ATC	FV***V***LV***V***L***I***	748	5	**3**	54.5
ATC	ATC	GTG	TTG	TTC	*TTG*	*TTG*	ATC	I***IV***LFLL***I***	828	5		
ATC	GTG	GTG	TTG	TTG	GTC	GTC	ATC	IV***V***L***LVVI***	714	3		
TTG	ATC	GTC	TTC	ATC	GTC	GTC	ATC	L***IV***FI***VVI***	776	3		
TTG	ATC	GTC	TTC	GTG	GTC	*TTG*	*CTC*	L***IV***FV***V***LL	774	5		
TTG	TTG	ATC	TTC	GTG	ATG	*TTG*	ATG	LLIFVMLM	832	8	**4**	50
TTG	GTG	GTG	ATC	TTC	ATG	*TTG*	ATC	LV***VI***FML***I***	804	5		
ATG	GTG	ATC	TTG	ATC	GTC	GTG	ATC	***MI***ILI***VVI***	771	3		
ATG	ATC	GTC	TTC	TTC	ATC	*TTG*	ATC	***MIV***FFIL***I***	866	4		
ATG	ATC	GTC	TTC	TTG	GTG	ATC	GTC	***MIV***F***LV***IV	778	3		
ATG	ATC	GTC	TTC	TTG	GTG	GTG	GTC	***MIV***F***LVV***V	751	2		
ATG	ATC	GTC	TTG	TTC	GTG	GTG	ATC	***MIV***LF***VVI***	778	2		
ATG	ATC	GTG	TTG	TTG	GTC	*TTG*	*CTC*	***MIV***L***LV***LL	768	3		
ATG	GTG	GTG	ATC	TTC	GTC	*TTG*	ATC	***M***V***VI***F***V***L***I***	778	3		
ATG	GTG	GTC	ATC	ATC	GTG	ATC	ATC	***M***V***VI***I***V***I***I***	745	3		
ATG	GTG	GTG	ATC	ATC	GTC	*TTG*	ATC	***M***V***VI***I***V***L***I***	744	3		
ATG	GTG	GTC	ATC	TTG	GTG	*TTG*	*CTC*	***M***V***VILV***LL[e]	742	3	**5**	48
ATG	GTG	GTG	ATC	TTG	GTG	*TTG*	*CTC*	***M***V***VILV***LL[e]	742	3		
ATG	GTG	GTC	ATC	*CTG*	GTG	*TTG*	*CTC*	***M***V***VILV***LL[e]	742	3		
ATG	GTC	GTG	ATC	TTG	GTG	*TTG*	*CTC*	***M***V***VILV***LL[e]	742	3		
ATG	GTG	GTG	ATC	GTG	ATC	*TTG*	*CTC*	***M***V***VI***VILL	743	5	**6**	
ATG	GTG	GTG	GTC	TTG	GTG	GTC	ATC	***M***V***V***V***LVVI***	690	2	**7**	
GTC	ATG	*TTG*	TTC	TTC	*TTG*	*TTG*	*CTC*	VMLFFLLL	863	8	**8**	
GTG	GTG	*TTG*	ATC	TTG	TTC	GTC	GTC	VVL***IL***F***V***V	748	5		
GTG	GTC	GTG	TTG	TTC	ATG	*TTG*	ATC	VV***V***LFML***I***	777	6		
GTC	GTC	GTC	TTG	TTC	*TTG*	ATC	ATC	VV***V***LFLI***I***	775	6		

[a] Codons in italic type are identical to those found in the parental, AL$_7$, mutant. [b] Residue types in bold, italic type are identical to wild-type mammalian ubiquitin residues. [c] Δ is the number of residues in the sequence that are different from the WT sequence. [d] Clones chosen for expression and characterization. [e] Identical protein sequences; underscores denote different codon usage in two of the DNA sequences. [f] Midpoint of thermal denaturation curves as determined by CD spectroscopy; error of ±1 °C.

that did survive would have had a competitive advantage during amplification.

We sequenced a total of 31 full-length clones from the selected pool (Table 1). The first, most obvious result was that, in contrast to the starting library, no parental (AL$_7$) sequences were recovered. In addition, only at two of the eight positions were parental codons recovered to a significant degree, namely, at positions 26 and 30. In these positions, there was some propensity for leucine in the retrieved, stable sequences, which is the residue encoded at these sites by the starting parental clone. In some cases, the codons retrieved at these sites could not be distinguished from parental codons. Thus, it is possible that this selection resulted from the bias for parental codons in the starting

Table 2: Counts of the Different Amino Acids Retrieved for Each Position Mutated in the Core of Ubiquitin

amino acid		observed counts for each position							
type	expected count[a]	1	3	5	13	15	17	26	30
F	5.17	6	0	0	10	10	1	0	0
I	5.17	2	**11**[b]	3	**11**	4	2	3	**16**
L	5.17	4	2	3	9	**12**	3	**19**	9
M	5.17	**15**	1	0	0	0	4	0	1
V	10.34	4	17	**25**	1	5	**21**	9	5
Z-score for WT[c]	−	4.74	2.81	5.59	2.81	3.29	4.06	-0.51	5.22
consensus	−	**M**	**I**	**V**	**I**	**L**	**V**	**L**	**I**
WT-UBQ	−	**M**	**I**	**V**	**I**	**L**	**V**	V	**I**

[a] Expected count values were calculated for 31 sequences assuming a distribution of 1:1:1:1:2 for residues F, I, L, M, and V, which follows from the codon usage in the degenerate mutagenic codon {GAT}T{GC}. [b] Consensus residues are denoted in bold type. [c] The *Z*-scores were calculated for each residue of the WT-UBQ sequence (see Materials and Methods). As the *Z*-scores for all but V26 were above the 99% confidence limit (≥2.58), this indicated that WT residues were selected significantly to a greater extent than expected by chance.

library and did not reflect a true selection for leucine residues. We do not believe that this is the case for the following reasons. First, at position 15, where the parental leucine residue was selected, the parental codon, CTG, from the initial library was rarely (1/31) seen in the selected sequences. Second, we compared the numbers of parental and nonparental codons in the starting and selected sequences and found that for positions 15, 26, and 30 (where Leu was selected) the level of usage of parental codons dropped in all cases after selection. Likewise, the remaining positions (1, 3, 4, 13, and 17) showed almost no parental codon usage in the selected sequences. Thus, the intact parental sequence AL_7 and codons from it were effectively expunged by the selection procedure.

Consensus Sequence. In addition to the marked deselection of the parental AL_7 mutant, there was a clear and unambiguous selection for WT residues in the retrieved sequences (Table 1). For example, 130 of the 248 residues (52%) in the selected and characterized sequences were WT, which contrasts with ≈57 WT residues expected on the basis of codon bias used in the library. Accordingly, there was a remarkable selection for aliphatic residues (I, L, and V) over the entire set of core positions. These residues occurred 200 times out of the possible 248 observations (81%, Table 2), and at the expense of the alternative F and M residues. Moreover, the preferences were positional and not just global. For example, at position 1, which is methionine in the WT sequence, but alanine in AL_7, half of the selected clones also had methionine at this site. Similar preferences at the other sites contributed to a strong consensus sequence with just one residue different from WT-UBQ; the consensus retrieved was MIVILVLI compared with the WT sequence MIVILVVI; i.e., only at position 26 was the WT residue not selected. Furthermore, as judged by *Z*-score analysis (Table 2), each of these WT residues occurred significantly more often than was expected on the basis of the doping strategy used to prepare the initial library.

Four identical protein sequences were recovered (Table 1). This set had M, V, V, I, L, V, L, and L at the targeted sites, which is just three steps from the WT sequence. One way for this result to have occurred would be if the same clone was present in many copies in the library. However, we could discard this possibility because in three of the four cases, although the amino acids selected for were identical, the DNA sequences were not. This also eliminated the possibility that diversity in the library was simply being lost during panning, and highlighted the point that the selection being achieved was based on the structural and stability properties of ubiquitin itself.

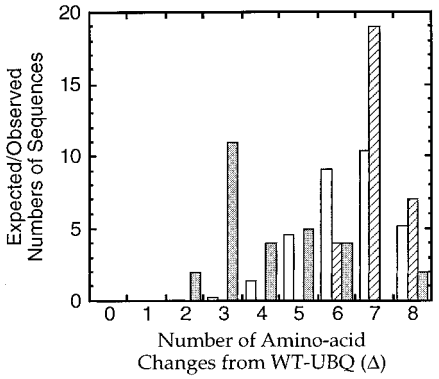

FIGURE 2: Histogram showing the number of changes (Δ) from the WT sequence observed in the characterized clones from the library before (cross-hatched bars) and after selection (shaded bars) compared with those numbers expected in the initial library (white bars). Expected numbers were calculated as $4\Delta\ {}^{8}C_{\Delta}$ for each value of Δ as there are 4Δ different sequences for each of the ${}^{8}C_{\Delta}$ possible combinations. These numbers were normalized before plotting by multiplying by $31/5^8$ to allow direct comparison with the numbers sequenced for each library. Comparison of the selected and expected distributions revealed that sequences with Δ values of 2 and 3 were the most highly selected by factors of ≈80 and 40, respectively.

Given the limited number of clones that could be sequenced, it is difficult to analyze the recovered sequence in any further detail. However, the migration of sequences from the broad composition expected for the starting library and toward more ubiquitin-like sequences is illustrated in Figure 2.

Expression of Selected Core Mutants. The recombinant WT protein and eight of the selected mutants were overexpressed, purified, and characterized. Proteins were chosen for expression studies to cover a wide variety of selected clones (clones **1**−**8**, Table 1). Clone **5** was chosen as it encoded the protein sequence found in four of the 31 selected clones. Clone **2** encoded the protein with the largest potential core volume and clone **7** the smallest. Clones **4** and **8** both had the maximum possible (eight) changes from the natural sequence, while clones **7**, **5**, **1**, **3**, and **2** had two, three, four, five, and six differences from WT ubiquitin, respectively (Table 1).

All of these proteins and the recombinant hexahistidine-tagged WT protein were expressed at sufficiently high levels to give amounts of protein for one-dimensional ^{1}H NMR

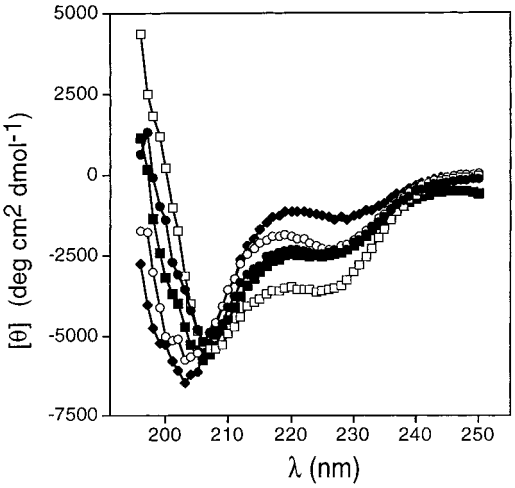

FIGURE 3: CD spectra for: recombinant WT-UBQ, commercial WT-UBQ, and three selected mutants **2**, **5**, and **7** (Table 1). These spectra were recorded at 5 °C in 50 mM phosphate at pH 2.1: (●) recombinant WT-UBQ, (□) commercial WT-UBQ, (○) mutant **2**, (■) mutant **5**, and (◆) mutant **7**.

and other structural studies. By contrast, expression trials for the parent mutant (AL_7) and six mutants chosen at random from the initial library did not yield sufficient protein for NMR, or even for good quality CD spectra. As expression rates correlate with stability (*17*−*19*), these comparative trials indicated that the average stability of mutants had been improved between the starting and selected pools. Accordingly, we envisage that the core-packing, protease-selection method may also be useful for improving expression yields of proteins in general.

Structure and Thermal Stability of the Selected Proteins. Circular dichroism (CD) spectra (Figure 3) were recorded for commercial ubiquitin, recombinant (hexahistidine-tagged) ubiquitin, and the eight selected variants. Recombinant ubiquitin had a spectrum similar to that of the commercial protein (Figure 3); however, the former had a more negative ellipticity at 200 nm, which was expected because of the additional, presumably disordered, residues in the hexahistidine tag and linker regions (see Materials and Methods). The eight selected variants had spectra similar to that of recombinant ubiquitin, although the exact shape differed. It is possible that the differences arose because the mutants had different numbers of aromatic residues (Table 1). Three spectra typical of these mutants are also shown in Figure 3.

Ubiquitin is an extremely stable protein (*4*, *5*) and at pH 7 does not melt below 100 °C (*20*), which precludes measurements of thermal stability at this pH. Therefore, we compared thermal unfolding transitions at pH 2.2, where the molecule is less stable (*4*). At this pH, commercial ubiquitin underwent a cooperative unfolding transition characterized by an increasing negative molecular ellipticity at 200 nm. The midpoint (T_M) of this transition, determined from the first derivative of the signal versus temperature plot, was 65 °C. Similarly, all of the variants that were tested exhibited cooperative thermal denaturation at pH 2.2 with T_M values in the range of 48−55 °C (Figure 4A and Table 1). These compared with a T_M of 62 °C for the recombinant ubiquitin (Figure 4A). To estimate the stability of the mutants under the conditions of the selection assay, we also unfolded one of them at pH 7 and found it to be highly stabilized like the wild-type protein, but with a lower T_M of ≈92 °C (Figure

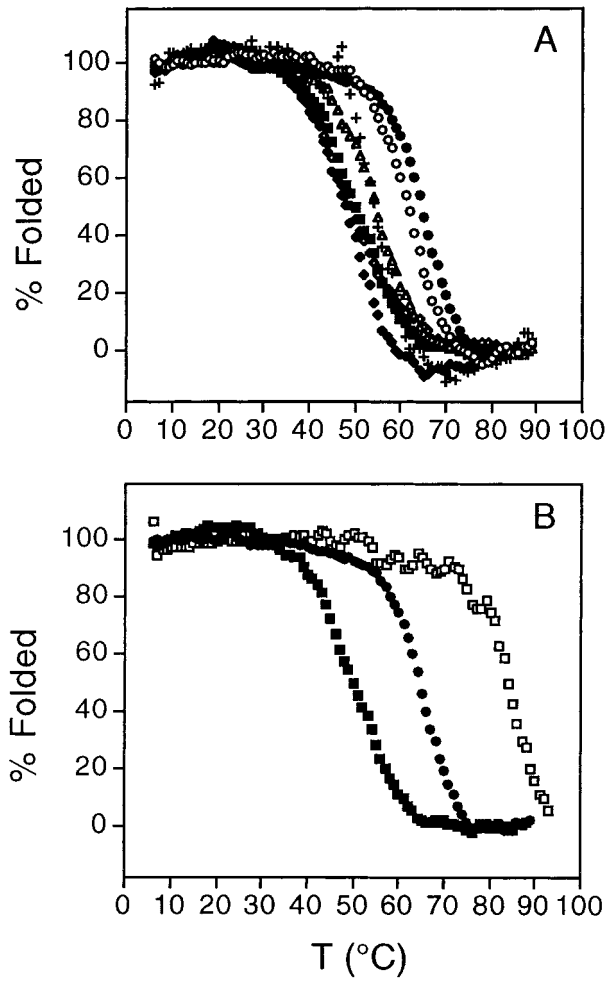

FIGURE 4: Thermal unfolding of WT and mutant ubiquitins (Table 1) followed by changes in the CD signals at 200 nm: (○) recombinant WT-UBQ, (●) commercial WT-UBQ, (+) mutant **1**, (◇) mutant **2**, (△) mutant **3**, (■) mutant **4**, (◆) mutant **5**, and (□) mutant **4** at pH 7. CD spectra were recorded at 5 °C in 50 mM phosphate at pH 2.1 except for the second trace for mutant **4** that is given in panel B (□).

4B), which is approximate because the high-temperature baseline could not be determined. This mutant had a T_M of 50 °C at pH 2.2, and we anticipate that all of our mutants would also have much higher stabilities (>90 °C) at pH 7 where our selection studies were conducted. We could not find any obvious correlation between the T_M values and any features that could be calculated from the selected sequences, e.g., predicted core volume.

One-dimensional ^{1}H NMR spectra were recorded for each of the mutants and the WT proteins (Figure 5). Spectra for the recombinant WT protein (Figure 5B) and commercial ubiquitin (Figure 5A) were closely similar. In native ubiquitin (Figure 1A), two aliphatic signals are shifted upfield and act as reporters of the folded structure; these resonances are from $C_\delta H_3$ of L50 (which is next to Y59 in the three-dimensional structure) and one of the magnetically inequivalent $C_{\gamma 1}H$ values of I61 (next to F45) (*21*, *22*). For recombinant ubiquitin, these two signals were shifted similarly, showing that the native interactions were intact. Likewise, the spread in backbone amide proton signals (6.1−9.5 ppm) was very similar in the two proteins. These data strongly support the fact that the hexahistidine-tagged protein retained the native ubiquitin structure. The only noticeable

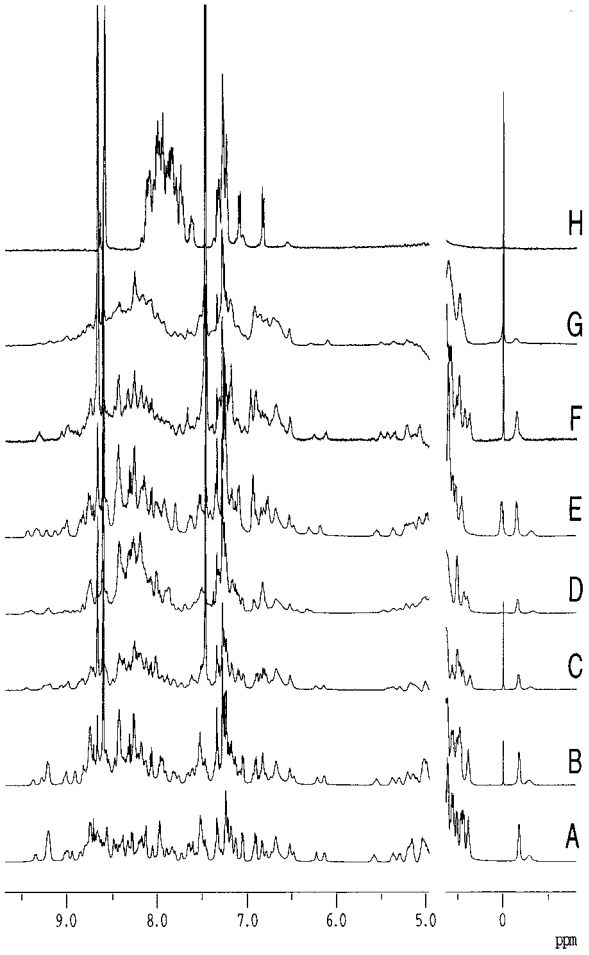

FIGURE 5: ^{1}H NMR spectra of WT and mutant ubiquitins determined with 50 mM phosphate at pH 2.2: (A) recombinant WT-UBQ at 35 °C, (B) commercial WT-UBQ at 35 °C, (C−G) mutants **7**, **5**, **3**, **4**, and **2**, respectively, all at 35 °C, and (H) recombinant WT-UBQ at 70 °C. The mutants are arranged in order of increasing predicted core volume, with the largest (mutant **2**) near the top (G).

differences were the two strong signals at 7.28 and 8.61 ppm in the spectra for the recombinant proteins, which were assigned to the two ring protons of the imidazole side chains in the histidine tag. For comparison, we also recorded the spectrum of recombinant ubiquitin in its unfolded state at 75 °C: the upfield-shifted methyl resonances were not evident, and the amide proton band was reduced to the range of 7.6−8.2 ppm (Figure 5H).

Spectra for each of the selected mutants (presented in order of increasing predicted core volume in parts C−G of Figure 5) also exhibited features more consistent with the folded WT proteins (Figure 5A,B) rather than the unfolded state (Figure 5H). All had the two signals upfield of 0 ppm observed for native ubiquitin. In addition for mutants 1−3, an additional methyl signal was seen at ~0 ppm (Figures 5E−G). In each of these mutants, but in none of the other five, position 1 was occupied by phenylalanine. It is probable, therefore, that the additional upfield shifted signals were due to interactions with Phe-1. In of all the spectra, the amide protons of the mutants covered spectral ranges similar to those observed for the folded WT proteins. Interestingly, we note that the mutant with the largest predicted core volume (mutant **2**, Figure 5G) had broadened signals, indicating that some conformational averaging may have been occurring. Nonetheless, it was clear that each of the mutants had competently folded structures similar to those of the WT proteins.

DISCUSSION

Rescue of a Destabilized Ubiquitin Variant with a Multiply Mutated Core. We have combined mutagenesis, phage-display, and protease selection to recover stable, folded variants of ubiquitin. We focused on residues 1, 3, 5, 13, 15, 17, 26, and 30 of mammalian ubiquitin, which form a clear substructure within the hydrophobic core (Figure 1). To avoid a background of protease-resistant WT protein in our selection, the protein was first destabilized by mutating the eight sites to one alanine and seven leucines, respectively. Using this template, the eight core positions were randomized to combinations of five hydrophobic residues, phenylalanine, isoleucine, leucine, methionine, and valine. This gave 390 625 possible mutants, which included the WT sequence. DNA sequencing of individual clones revealed that $\approx 1/3$ were parental, AL_7-UBQ. However, after protease selection, none of the 31 selected clones had this sequence. The variants that were recovered were found to be stable and folded as judged by CD and NMR measurements. Together, the results demonstrate the utility of our method in repacking and stabilizing poorly packed initial structures.

Volume Constraints in Protein Folding and Stability of Ubiquitin. Others have shown that, alongside hydrophobicity, a key feature preserved in protein cores is volume (*23*, *24*). Consistent with this, Figure 6 shows how the predicted core volumes for our selected mutants lie close to that of native ubiquitin. Figure 6A shows the results of core-volume calculations for 1000 randomly generated sequences expected in the initial library. These volumes do not form a continuous distribution, but are quantized roughly by the volume of one methylene group (≈27 Å^3). Mutants that we chose at random from this library and sequenced experimentally did exhibit a broad distribution of predicted core volumes, which roughly reflected the randomly generated calculated volumes (Figure 6B). In contrast, the predicted volumes of sequences in the pool of stable selected mutants clearly clustered around two values (Figure 6C). The first cluster, which contained ≈33% of the sequences, had roughly the same calculated volume as WT-UBQ (≈750 Å^3). The second cluster of residues, also ≈33% of the sequences, were calculated to have volumes that were greater than that of WT-ubiquitin by a single methylene group (≈780 Å^3). The remaining third of the selected sequences were distributed widely, with a range similar to that found for the starting library.

Up to a point, it is conceivable that increasing the volume of the side chains in a protein core would lead to an increase in thermal stability, because van der Waals interactions increase and potential side chain-packing heterogeneity is likely to be reduced (*25*, *26*). With further increases, however, backbone readjustments may be needed to accommodate the increased volume, or the protein may become over-filled and destabilized (*7*, *27*). For ubiquitin, one mutation is known that introduces an additional methylene to the core and stabilizes the protein (*28*). This is consistent with our finding that $1/3$ of our sequences have predicted core volumes that are one methylene group larger than that of the WT. However, other studies (*20*) and our present work

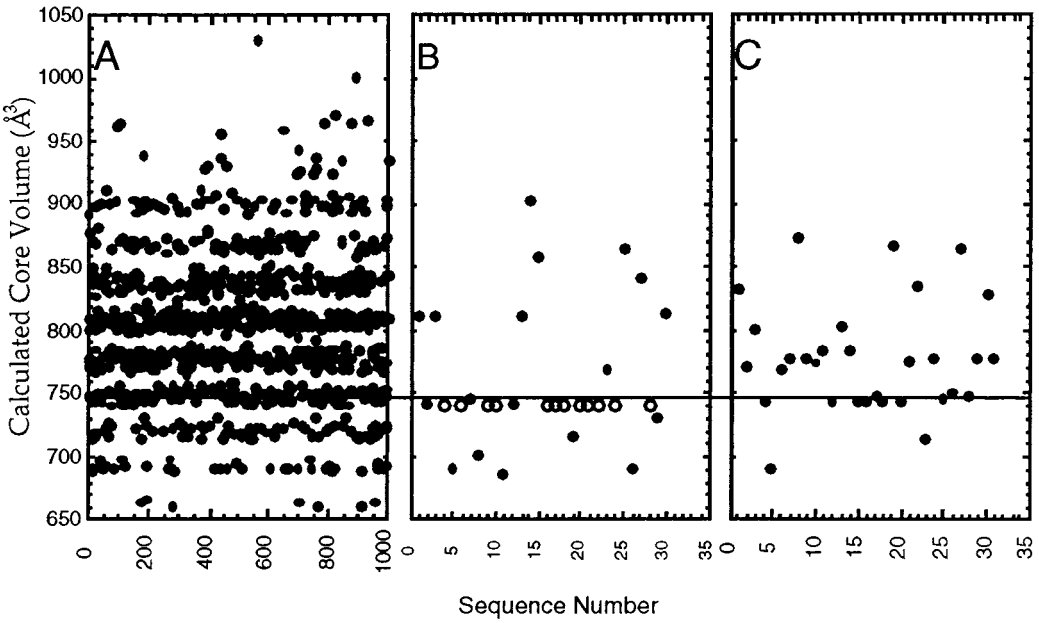

FIGURE 6: Core volumes predicted from the library sequences. (A) Predicted core volumes for mutants from the starting library. One thousand random sets of eight side chains were generated by computer. The proportions of the side chains were biased in the same way as in the oligonucleotides designed to make the library, i.e., one each of F, I, L, and M and two copies of V. The volumes of the side chains were taken from Creighton (*53*). (B) The predicted core volumes of 30 clones sequenced from the starting library. The points for the parental, AL_7 clones (volume of 740 Å^3) are represented by white circles. (C) The predicted core volumes for the 31 clones retrieved after protease selection. Ten of the selectants had approximately the same volume as WT-UBQ (represented by the solid line at 744 Å^3), and 11 had a volume approximately one methylene group larger than this.

suggest that greater volume changes are less well tolerated. Thus, it appears that the optimal core volume for ubiquitin is that given by the wild-type residues, possibly plus one additional methylene.

Sequence Restrictions on the Folding and Stability of Ubiquitin. Ubiquitin is found in all eukaryotes where it is highly conserved. Thus, the protein must be under strong selective pressures. The question is whether these act on protein structure, stability, and function equally or to different degrees.

The proteins recovered in our selection studies displayed a clear consensus for near-WT sequences; the consensus residues recovered at seven of the eight positions targeted were WT. The exception was at position 26, which exhibited a preference for leucine rather than the WT valine. Interestingly, the V26L variant of mammalian ubiquitin has been prepared by others and exhibits an unfolding free energy higher than that of the natural sequence (*28*). In addition, leucine occurs frequently in this position among ubiquitin-like proteins, including yeast Smt3p, rat SUMO-1, and mouse UCRP (*29*, *30*) (Table 3).

To examine if the core residues of ubiquitin were conserved in more-distantly related natural sequences, we analyzed amino acid variation in ubiquitin-like proteins and for proteins with the ubiquitin superfold. Proteins whose sequences are 20−80% identical to that of ubiquitin, but with different functions, have been identified (*29*, *31*). In addition, proteins with the ubiquitin structural superfold are known that are ≤20% identical to ubiquitin (*32*). Using Pfam (*33*) and BLAST (*34*), we gathered a set of ubiquitin-like sequences and removed redundant entries with >90% identity to other members; a representative group of the final set (sequences 5−13) is aligned in Table 3. Consistent with other protein families (*23*, *35*), gross sequence changes correlated with predicted protein surfaces. However, the 16 core sites predicted from the structure of human ubiquitin (Table 3) showed strong levels of conservation; the consensus sequence was identical to mammalian ubiquitin at all positions except 13, 23, and 26 (Table 3). This high degree of core conservation extends to the proteins that share the ubiquitin superfold, but are otherwise unrelated to ubiquitin; theoretical work has located a highly conserved set of nine "nucleus-forming" residues in the hydrophobic core of these structures (*32*), and we note that the consensus for these residues is identical to the ubiquitin-like protein consensus at five of the sites, equivalent to positions 3, 5, 15, 17, and 26 in Table 3. These analyses demonstrate that the conservation of this hydrophobic core is not linked to any functional constraint. On the contrary, and as indicated by our selection studies, the hydrophobic core of ubiquitin is maintained for fundamental structural reasons.

These points raise the following question. Why did we not retrieve the WT and V26L sequences by protease selection if these really are optimal? The best explanation that we can offer is as follows. As our starting library was predicted to have 390 625 mutants, the selection of any one particular sequence was extremely unlikely at ≈1 in 400 000. Thus, put simply, we may not be able to sequence enough selectants to observe specific predetermined sequences. A better way to view the problem is to consider what was the likelihood of retrieving sequences similar to the WT. In essence, our library allowed five types of substitution at eight distinct sites. Thus, for a pool of 390 625 sequences, we can calculate (Figure 2) that 32 of these had one amino acid change from the WT, 448 had two changes, 3584 three changes, and so on. Thus, all other things being equal, we were much more likely to retrieve proteins with two or more changes from the WT than the two specific sequences. In

Table 3: Sequence Alignment of Ubiquitin and Ubiquitin-like Proteins[a,b]

	sequence number
	1 10 20 30 ... 40 50 60 70 76
	| | | | ... | | | | .... |
1	**M**Q**I**F**V**KTLTGKT**I**T**L**E**V**EPSDT**I**EN**V**KAK**I**QDKEG...IPPDQ**Q**R**L**IFAGKQ**L**EDGRT**L**SDYN**I**QKEST**L**H**L**VLR.......LRGG
2	MQIFVKTLTGKTITLEVEASDTIENVKAKIQDKEG...IPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR.......LRGG
3	MQIFVKTLTGKTITLEVESSDTIDNVKAKIQDKEG...IPPDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLR.......LRGG
4	MQIFVKTLTGKTITLEVESSDTIDNVKSKIQDKEG...IPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR.......LRGG
5	MQIFVKTLTGKTITVETEPGDTVGQVKQKIADKEG...VPVDQQRLIYAGKQLEDAQTLADYNIQKESTLHMVLR.......LRGG
6	MLIKVKTLTGKEIEIDIEPTDKVERIKERVEEKEG...IPPQQQRLIYSGKQMNDEKTAADYKILGGSVLHLVLA.......LRGG
7	MIVKVKTLTGKEISVELKESDLVYHIKELLEEKEG...IPPSQQRLIFQGKQIDDKLTVTDAHLVEGMQLHLVLT.......LRGG
8	MQLFVRAQE..LHTFEVTGQETVAQIKAHVASLEG...IAPEDQVVLLAGAPLEDEATLGQCGVEALTTLEVAGR.......MLGG
9	LSILVRNNKGRSSTYEVRLTQTVAHLKQQVSGLEG...VQDDLFWLTFEGKPLEDQLPLGEYGLKPLSTVFMNLR.......LRGG
10	MQVTLKTLQQQTFKIDIDPEETVKALKEKIESEKGKDAFPVAGQKLIYAGKILNDDTALKEYKIDEKNFVVVMVTKP.KAVSTPAP
11	VTITLKTLQQQTFKIRMEPDETVKVLKEKIEAEKGRDAFPVAGQKLIYAGKILSDDVPIRDYRIDEKNFVVVMVT....KTKAGQG
12	LEVLVKTLDSQTRTFIVGAQMNVKEFKEHIRASVS...IPSEKQRLIYQGRVLQDDKKLQEYNVGGK.VIHLVERAP.PQTHLPSG
13	LNIHIKSGQ.DKWEVNVAPESTVLQFKEAINKANG...IPVANQRLIYSGKILKDDQTVESYHIQDGHSVHLVKSQPKPQTASAAG

[a] Rows 1−4, human, *Caenorhabditis elegans*, soybean, and *Saccharomyces cerevisiae* ubiquitin, respectively; row 5, baculoviral vUb; row 6, residues 1−76 of human Nedd-8; row 7, residues 1−76 of *S. cerevisiae* Rub1p; row 8, residues 1−74 of human fau; row 9, residues 30−104 of human UCRP; row 10, residues 1−85 of human Rad23B; row 11, residues 3−84 of human Rad23A; row 12, residues 17−97 of human Bat3; row 13, residues 3−84 of Dsk2 from *S. cerevisiae*. [b] Residues in bold type and underlined are the core residues of mammalian ubiquitin that were mutated in this study; the remaining residues in bold type are the other residues in the hydrophobic core.

this respect, the interesting points are (1) that even though sequences with seven amino acid changes were predicted to account for ≈$^1/_3$ of the sequences of the starting library they were not retrieved at all after selection, but in contrast (2), sequences with two or three substitutions dominate the selected pool even though they were expected to account for only ≈1% of the starting library (Figure 2). This illustrates both the efficiency of our selection procedure and the strict sequence requirements for a stable hydrophobic core in ubiquitin.

Evolutionary Constraints on Protein Sequence. During the course of evolution, nature could not have explored sequence space entirely. One limitation is that the number of possible amino acid sequences, even for small proteins, is very much greater than the number of proteins that could have existed. This has implications for studies that seek sequence−structure relationships in databases of known proteins. For example, the fact that a structural motif is invariably coded by similar sequences does not necessarily mean that the sequence is obligatory for folding, as equally fit alternatives may not have been tested.

A second limitation in natural selection is that once a protein function is "locked in", in the sense that it becomes necessary for the organism's survival, pressure is added that resists or, at least, limits change. For instance, multiple mutations in a protein rarely occur simultaneously; rather, they result from a series of single mutations. If the first mutation is strongly deleterious, such that the protein is unable to function correctly, a compensatory second mutation may not have a chance to occur, as the organism will be selected against (*36*). Even if the first mutation is not sufficiently destabilizing to undermine the survival of the organism, a second mutation usually will not restore WT function completely. Thus, a likely event is reversion to the parental type (*37*). This has been recently confirmed experimentally, with 90% of second-site mutations in T4 lysozyme reverting to the parent (*38*).

It is for these reasons that in nature large areas of sequence space will not be explored (*9*, *35*). Thus, it is of interest to know whether an "optimal sequence" can be obtained for a sequence in the absence of such constraints.

Selection studies of the type described in this paper offer a chance to explore sequence space free from evolutionary constraints, and to attempt to find optimal sequences, if indeed these exist. In our experiments, many possible ubiquitin mutants were generated simultaneously, and we selected those that were folded and stable, thus exploring a large area of the possible sequence space and bypassing the natural step-by-step selection approach. We recovered sequences with a consensus extremely similar to the natural WT sequence, which implies that, for ubiquitin at least, there has been no barrier in nature to finding the optimal sequence as described above. [The fact that a variant of ubiquitin (V26L) has been found which is marginally more stable than the WT does not invalidate this as only a single point mutation would be required to make the change.] One explanation for this is that there are many pathways through sequence space, and at least one could provide a chain of functional proteins to allow the space between ancestral and modern sequences to be traversed. However, our results suggest an intriguing alternative.

The sequences that we selected indicate that the fitness (or stability) landscape for the hydrophobic core of ubiquitin comprises a broad peak with many stable and closely related sequences, including the WT and V26L; a number of sharper peaks also with stable proteins, but with fewer similar

sequences nearby; and valleys of unfit (unstable) sequences separating these peaks. It is intriguing that natural selection has also yielded sequences that lie on the broad peak that our experiments highlighted. This may have occurred by chance, simply because the odds of an ancestral protein arising near or on the broad peak are high. However, we propose that proteins that lie on broad peaks may have a greater evolutionary robustness compared to those that lie on the sharper peaks and have much tighter sequence constraints. Consider a landscape populated by many different sequences. Only a small number of these will match up to any one set of pressures required for a given function. As selection pressures mount (or change), more sequences will be shed from the population. In this scenario, it stands to reason that related sequences on broad peaks will outsurvive those on sharper peaks. This is because many alternative sequences in which the necessary function is maintained will be accessible through single-point mutations. As a result, proteins that lie on broad peaks in the fitness landscape may eventually dominate entirely at the expense of the "less-adaptable" sequences on sharper peaks. It is possible, therefore, that of the potential proteins, those that remain are likely to be more tolerant to small mutations. In other words, it is not by chance that WT-UBQ is a member of broad peak of related sequences, and this may be a feature of proteins in general.

Core-Packing Constraints. Our core-directed design method makes an assumption that the stability of proteins is governed in part by how well the side chains pack together within the hydrophobic core. Others have studied this by examining how permissive the hydrophobic core is to substitution (*24*). In some cases, only a limited number of alternative residues are possible for given positions within the core before functionality is lost (*10*, *39*, *40*). In contrast, other proteins are quite permissive of substitutions (*10*, *26*, *41*−*43*). One way to reconcile these dissimilar findings is to suggest that mutations at some sites have a large effect on stability, while mutations at other sites have little effect. However in most cases, mutations destabilize different proteins by amounts that correlate with the hydrophobicity, the change in the size of the residue, and the degree of burial (*10*, *39*, *44*). (For this argument, we note that the residues being considered are buried within cores and will make roughly the same number of contacts.) Finally, the effects of multiple mutations are often simply additive (*45*).

Another possibility is that mutational tolerance depends on the size of the hydrophobic core being repacked (*39*). For instance, proteins that have small hydrophobic cores may be less stable as they are unable to bury enough hydrophobic surface to achieve sufficient stability to resist unfolding. As a result, side chain interactions may have to be optimized more precisely; for example, the formation of cavities may be precluded. Proteins with relatively large hydrophobic cores, on the other hand, may bury enough hydrophobic surface area to achieve sufficient stability for structural integrity relatively easily. Therefore, these may accommodate destabilizing interactions, such as cavities, more readily.

If this argument held completely, the results presented here for ubiquitin would be even more surprising. Ubiquitin has a reasonably large hydrophobic core, and we randomly mutated a comparatively large number of residues simultaneously; accordingly, we would have expected to recover a wide range of sequences, instead of the narrow range of sequences with a clear consensus actually observed. In addition, there appears to be no simple correlation between core volume and thermal stability of the selected ubiquitin mutants (Table 1). Thus, as we suggested above, it is probable that the ubiquitin sequence is special in the sense that certain, specific core residues are required to specify and consolidate the ubiquitin superfold. In addition, however, our results might to some extent reflect the selection procedure that we used.

For instance, we might have expected a broad range of sequences had we used an assay based on function (i.e., binding or enzymatic activity) that *only* required the protein to be structurally intact and active at temperatures in vivo. However, we selected much more stringently only those proteins that were hyperstable. We propose that this is an additional reason that we recovered a narrow range of sequences. Related to this, the coincidence of the WT and our selected sequences suggests a reason ubiquitin is highly conserved in nature. Ubiquitin is induced at increased levels in the cell under conditions of stress, where levels of denatured proteins are increased (*46*, *47*). It is important, therefore, that ubiquitin itself resists denaturing conditions so that it may continue to function under these circumstances. Accordingly, only a few arrangements of residues within the core may be compatible with the required stability. More generally, we anticipate getting equally strong consensus sequences using stability-based selection procedures on other proteins, even when the natural sequences are not subject to such strong stability pressures in nature and do not show the strong natural conservation of ubiquitin.

Stability-Based versus Function-Based Selection. Above we suggest that stability-based selection procedures are more stringent than function-based ones, as proteins are not in general required to be hyperstable to function optimally. Indeed, in some cases where stability has been deliberately introduced, a decrease in the level of function is observed (*25*, *48*). For this reason, selection procedures based on function may not necessarily yield the most stable proteins.

We have used an assay that tests directly for stability and probes directly the requirements for producing a well-packed, stable core. In contrast to function-based selections, which often show permissive requirements, we find that the number of acceptable residue types at certain core locations in ubiquitin is limited. These results are similar to those obtained by others, for both ubiquitin and other systems, who also avoid function-based approaches (*20*, *49*). Selection methods such as ours could be applied to improving the stabilities of proteins when function-based selection is not possible. For example, the method would be particularly appropriate for stabilizing proteins of de novo design.

Implications for Protein Structure, Prediction, and Design. Our experimental findings combined with the analysis of ubiquitin-like and ubiquitin structural superfamily sequences indicate that the hydrophobic core of ubiquitin is more strictly conserved than those of other proteins (*20*, *23*, *35*). This conservation extends beyond functional boundaries (*32*) and points toward more fundamental underlying structural principles. We propose that the hydrophobic core of ubiquitin has been optimized for stability and plays a key role in the specification of the ubiquitin superfold. Ubiquitin may not

be unique in this respect (*50*), and we anticipate that similar key core-packing arrangements may be found for other repeated protein architectures. Such features would represent clear rules for the prediction and design of protein structures.

ACKNOWLEDGMENT

We thank Mike Gradwell and Tom Frenkiel for their generous assistance with the NMR spectrometer at NIMR, Mill Hill, and members of the Sussex Centre for the Study of Evolution for helpful discussions.

REFERENCES

1. Lupas, A. (1996) *Trends Biochem. Sci. 21*, 375−382.
2. Cohen, C., and Parry, D. A. (1994) *Science 263*, 488−489.
3. Schneider, J. P., Lombardi, A., and DeGrado, W. F. (1998) *Folding Des. 3*, R29−R40.
4. Woolfson, D. N., Cooper, A., Harding, M. M., Williams, D. H., and Evans, P. A. (1993) *J. Mol. Biol. 229*, 502−511.
5. Wintrode, P. L., Makhatadze, G. I., and Privalov, P. L. (1994) *Proteins 18*, 246−253.
6. Handel, T. M., Williams, S. A., and Degrado, W. F. (1993) *Science 261*, 879−885.
7. Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., O'Brien, R., Sturtevant, J. M., and Regan, L. (1996) *Protein Sci. 5*, 1584−1593.
8. Wu, L. C., and Kim, P. S. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 14314−14319.
9. Lim, W. A., and Sauer, R. T. (1989) *Nature 339*, 31−36.
10. Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A., and Matthews, B. W. (1993) *Science 262*, 1715−1718.
11. Desjarlais, J. R., and Handel, T. M. (1995) *Protein Sci. 4*, 2006−2018.
12. Gassner, N. C., Baase, W. A., and Matthews, B. W. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 12155−12158.
13. Prodromou, C., and Pearl, L. H. (1992) *Protein Eng. 5*, 827−829.
14. Kunkel, T. A. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 488−492.
15. Clackson, T., and Winter, G. (1989) *Nucleic Acids Res. 17*, 10163−10170.
16. von-Frauenhofer, J. A., and Murray, J. J. (1976) *Statistics in medical, dental and biological studies*, Tri-Med Books, London.
17. Parsell, D. A., and Sauer, R. T. (1989) *J. Biol. Chem. 264*, 7590−7595.
18. Inoue, I., and Rechsteiner, M. (1994) *J. Biol. Chem. 269*, 29241−29246.
19. Kwon, W. S., Da Silva, N. A., and Kellis, J. T., Jr. (1996) *Protein Eng. 9*, 1197−1202.
20. Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1997) *Protein Sci. 6*, 1167−1178.
21. Di Stefano, D. L., and Wand, A. J. (1987) *Biochemistry 26*, 7272−7281.
22. Weber, P. L., Brown, S. C., and Mueller, L. (1987) *Biochemistry 26*, 7282−7290.
23. Chothia, C., Gelfand, I., and Kister, A. (1998) *J. Mol. Biol. 278*, 457−479.
24. Richards, F. M., and Lim, W. A. (1993) *Q. Rev. Biophys. 26*, 423−498.
25. Ogata, K., Kanei-Ishii, C., Sasaki, M., Hatanaka, H., Nagadoi, A., Enari, M., Nakamura, H., Nishimura, Y., Ishii, S., and Sarai, A. (1996) *Nat. Struct. Biol. 3*, 178−187.
26. Kellis, J. T., Jr., Nyberg, K., and Fersht, A. R. (1989) *Biochemistry 28*, 4914−4922.
27. Dahiyat, B. I., and Mayo, S. L. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 10172−10177.
28. Khorasanizadeh, S., Peters, I. D., and Roder, H. (1996) *Nat. Struct. Biol. 3*, 193−205.
29. Johnson, P. R., and Hochstrasser, M. (1997) *Trends Cell Biol. 7*, 408−413.
30. Lapenta, V., Chiurazzi, P., van der Spek, P., Pizzuti, A., Hanaoka, F., and Brahe, C. (1997) *Genomics 40*, 362−366.
31. Haas, A. L., Katzung, D. J., Reback, P. M., and Guarino, L. A. (1996) *Biochemistry 35*, 5385−5394.
32. Michnick, S. W., and Shakhnovich, E. (1998) *Folding Des. 3*, 239−251.
33. Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998) *Nucleic Acids Res. 26*, 320−322.
34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol. 215*, 403−410.
35. Lesk, A. M., and Chothia, C. (1980) *J. Mol. Biol. 136*, 225−270.
36. Maynard Smith, J. (1970) *Nature 225*, 563−564.
37. Baldwin, E., Xu, J., Hajiseyedjavadi, O., Baase, W. A., and Matthews, B. W. (1996) *J. Mol. Biol. 259*, 542−559.
38. Jucovic, M., and Poteete, A. R. (1998) *Protein Sci. 7*, 2200−2209.
39. Milla, M. E., and Sauer, R. T. (1995) *Biochemistry 34*, 3344−3351.
40. Reidhaar-Olson, J. F., and Sauer, R. T. (1988) *Science 241*, 53−57.
41. Sandberg, W. S., and Terwilliger, T. C. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 1706−1710.
42. Lim, W. A., and Sauer, R. T. (1991) *J. Mol. Biol. 219*, 359−376.
43. Axe, D. D., Foster, N. W., and Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 5590−5594.
44. Main, E. R. G., Fulton, K. F., and Jackson, S. E. (1998) *Biochemistry 37*, 6145−6153.
45. Sandberg, W. S., and Terwilliger, T. C. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 8367−8371.
46. Bond, U., and Schlesinger, M. J. (1985) *Mol. Cell Biol. 5*, 949−956.
47. Finley, D., Ozkaynak, E., and Varshavsky, A. (1987) *Cell 48*, 1035−1046.
48. Gryk, M. R., Jardetzky, O., Klig, L. S., and Yanofsky, C. (1996) *Protein Sci. 5*, 1195−1197.
49. Sieber, V., Plückthun, A., and Schmid, F. X. (1998) *Nat. Biotechnol. 16*, 955−960.
50. Cordes, M. H., Davidson, A. R., and Sauer, R. T. (1996) *Curr. Opin. Struct. Biol. 6*, 3−10.
51. Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987) *J. Mol. Biol. 194*, 531−544.
52. Sayle, R. A., and Milner-White, E. J. (1995) *Trends Biochem. Sci. 20*, 374−376.
53. Creighton, T. E. (1993) *Proteins: Structures and Molecular Properties*, 2nd ed., W. H. Freeman and Co., New York.
54. Finucane, M. D., Tuna, M., Lees, J. H., and Woolfson, D. N. (1999) *Biochemistry 38*, 11604−11612.

BI990766F