Protein Science (1998), 7:2345–2353. Cambridge University Press. Printed in the USA.

Stabilizing the subtilisin BPN′ pro-domain by phage display selection: How restrictive is the amino acid code for maximum protein stability?

BIAO RUAN, JOEL HOSKINS, LAN WANG, AND PHILIP N. BRYAN

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850

(RECEIVED April 22, 1998; ACCEPTED July 9, 1998)

Abstract

We have devised a procedure using monovalent phage display to select for stable mutants in the pro-domain of the serine protease, subtilisin BPN′. In complex with subtilisin, the pro-domain assumes a compact structure with a four-stranded antiparallel β-sheet and two three-turn α-helices. When isolated, however, the pro-domain is 97% unfolded. These experiments use combinatorial mutagenesis to select for stabilizing amino acid combinations at a particular structural locus and determine how many combinations are close to the maximum protein stability. The selection for stability is based on the fact that the independent stability of the pro-domain is very low and that binding to subtilisin is thermodynamically linked to folding. Two libraries of mutant pro-domains were constructed and analyzed to determine how many combinations of amino acids at a particular structural locus result in the maximum stability. A library comprises all combinations of four amino acids at a structural locus. Previous studies using combinatorial genetics have shown that many different combinations of amino acids can be accommodated in a selected locus without destroying function. The present results indicate that the number of sequence combinations at a structural locus, which are close to the maximum stability, is small. The most striking example is a selection at an interior locus of the pro-domain. After two rounds of phagemid selection, one amino acid combination is found in 40% of sequenced mutants. The most frequently selected mutant has a $\Delta G_{\text{unfolding}} = 4$ kcal/mol at 25 °C, an increase of 6 kcal/mol relative to the naturally occurring sequence. Some implications of these results on the amount of sequence information needed to specify a unique tertiary fold are discussed. Apart from possible implications on the folding code, the phage display selection described here should be useful in optimizing the stability of other proteins, which can be displayed on the phage surface.

Keywords: combinatorial genetics; protein folding; site-directed mutagenesis; stopped flow kinetics; thermodynamics

Recent studies have employed combinatorial genetics to investigate how amino acid sequence encodes unique tertiary folds (Sauer, 1996). Many of these studies have shown that a high degree of degeneracy in amino acid sequences is compatible with a particular backbone topology. A simple binary code of polar and apolar amino acids can be enough information to specify secondary structure and in some cases super secondary structure (Kamtekar et al., 1993). What seems to be not as well understood is how much information is required to go from compact secondary structure to stable, unique folds (Betz et al., 1993). The degeneracy of the folding code is an issue with important implications for protein structure prediction and de novo protein design (Laurents et al., 1994; Cordes et al., 1996). If unique states can be encoded with binary information then prediction algorithms relying on very simple energy functions should be able to predict correct tertiary folds. If not, they may become stuck at the level of predicting secondary structure. Analogously, de novo design strategies relying on polar and apolar patterns will succeed in producing unique folds if the correct binary pattern is sufficient information. If it is not, they will continue to generate secondary structure but usually will not generate unique tertiary structure.

Reprint requests to: Philip N. Bryan, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850; e-mail: bryan@umbi.umd.edu.

Abbreviations: [P], pro-domain concentration; Pi, phosphate; pro-wt, wild-type pro-domain of subtilisin BPN′; pro-R1, pro-domain with the following mutations: A23C, K27E, V37L, Q40C; [S], subtilisin concentration; Sbt12, subtilisin BPN′ with the following mutations: Y217K, N218S, and S221C; Sbt15, subtilisin BPN′ with the following mutations: deletion of amino acids 75-83, Y217K, N218S, and S221C; Tris, Tris(hydroxymethyl)amino-methane.

The experiments described here use combinatorial mutagenesis to select for stabilizing amino acid combinations at a particular structural locus and to determine how many amino acid sequences are close to the maximum protein stability. Our experiments use the pro-domain of the bacillus serine protease, subtilisin BPN′, as the target for stabilization. Secretion and folding of subtilisin require two separate processing steps on the initial precursor protein (Wells et al., 1983; Vasantha et al., 1984). A 30 amino acid signal peptide is removed during secretion. The extra-cellular part of the maturation process appears to involve folding of prosubtilisin, self-processing of the 77 amino acid pro-domain to produce a processed complex, and eventually degradation of the pro-domain to create the 275 amino acid mature form of the enzyme (Power et al., 1986; Ikemura et al., 1987). In complex with subtilisin the pro-domain assumes a compact structure with a four-stranded antiparallel β-sheet and two three-turn α-helices (Bryan et al., 1995; Gallagher et al., 1995) (Fig. 1). The folded pro-domain has shape complementary and high affinity to native subtilisin (Wang et al., 1995). When isolated, however, the pro-domain is 97% unfolded even under optimal folding conditions (0.1 M KPi, pH 7.0, 20 °C) (Strausberg et al., 1993).

We have devised a procedure using monovalent phage display to select for stable pro-domain mutants. To perform the selection, the pro-domain was synthesized as a fusion protein with the gene III

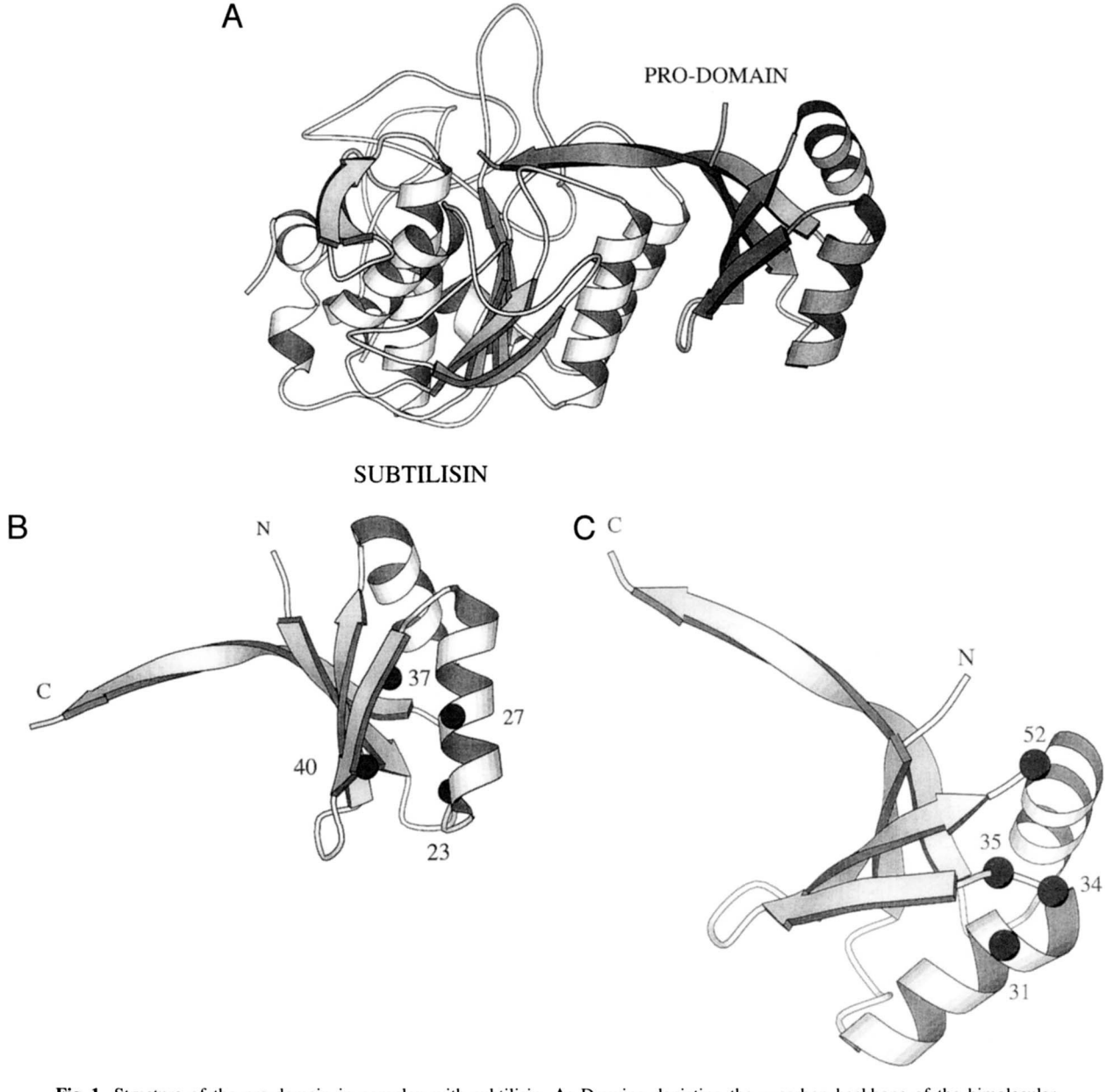

Fig. 1. Structure of the pro-domain in complex with subtilisin. **A:** Drawing depicting the α-carbon backbone of the bimolecular complex of subtilisin (lighter shading) and pro-domain (darker shading). The structure was determined by X-ray diffraction at 2.0 Å resolution (Gallagher et al., 1995). Drawn with MOLSCRIPT (Kraulis, 1991). **B:** Mutagenized positions at locus one. The α-carbon atoms of the side chains targeted for mutagenesis (23, 27, 37, 40) are shown by black spheres. **C:** Mutagenized positions at locus two. The α-carbon atoms of the side chains targeted for mutagenesis (31, 34, 35, 52) are shown by black spheres.

coat protein of the coli phage fd so it is displayed on the surface of phagemid particles (Hoogenboom et al., 1991; Smith & Scott, 1993). The amino acids at a particular structural locus of the pro-domain are randomized and phagemid particles bearing pro-domain mutants with the most stable folds are selected based on their ability to bind to subtilisin. The selection for stability is based on the fact that the independent stability of the pro-domain is very low and that binding to subtilisin is thermodynamically linked to folding. The native tertiary structure of the pro-domain with the four-stranded β-sheet formed correctly is required for it to bind to subtilisin. If mutations are introduced in regions of the pro-domain that do not directly contact subtilisin, their effects on binding to subtilisin will be linked to whether or not they stabilize the native conformation. Therefore, mutations that stabilize independent folding of the pro-domain will increase its binding affinity to subtilisin (Ruvinov et al., 1997). Two libraries of mutant pro-domains were constructed and analyzed. A library comprises all combinations of four amino acids at a structural locus. By selecting for the mutants binding most strongly to subtilisin, we can determine what combinations of amino acids at the locus are the most stabilizing, and how many different combinations are near the maximum stability.

Results

Theory of the selection for stabilized pro-domains

Mutations that stabilize the folding of the pro-domain will increase binding to subtilisin. According to the equilibrium:

$$P_u \overset{K_f}{\Leftrightarrow} P + S \overset{K_P}{\Leftrightarrow} P - S \quad (1)$$

where K_f is the equilibrium constant for folding the pro-domain and K_P is the association constant of folded pro-domain for subtilisin. The observed binding constant is then

$$K_{(P+P_u)} = \{K_f/(1 + K_f)\}K_P. \quad (2)$$

As the fraction of folded pro-domain approaches one, the observed association constant approaches its maximum, K_P. This thermodynamic linkage between stability and binding has been demonstrated previously using a series of designed mutations, which increase the stability of the pro-domain (Ruvinov et al., 1997).

About 97% of the wild-type pro-domain is unfolded at 25 °C, corresponding to a $\Delta G_{\text{unfolding}}$ of -2.1 kcal/mol (Bryan et al., 1995). The observed binding constant will be fairly sensitive to increases in stability of up to 4 kcal/mol. At that point the $\Delta G_{\text{unfolding}}$ of the pro-domain would be 2 kcal/mol, corresponding to 97% folded pro-domain, and the observed binding constant to subtilisin would be equal to 97% of its maximum value (K_P). The observed binding constant for the wild-type pro-domain (pro-wt) and Sbt12 subtilisin is 2×10^8 M^{-1} in 0.1 M KPi pH 7.0 at 25 °C (Strausberg et al., 1993). Sbt-12 is an inactive mutant (S221C) mutant of subtilisin used to avoid proteolysis of the pro-domain. K_P is predicted from equation 1 to be $\sim 7 \times 10^9$ M^{-1}.

Control selection

To determine conditions for phagemid binding, we first performed a selection in which phagemid particles containing the pro-wt-g3p fusion protein were added to 10 pmoles of biotinylated Sbt12 (1 mL of a 10 nM solution). The biotinylated Sbt12 is in turn bound to streptavidin-coated magnetic beads, which are collected on a magnetic particle concentrator. The input of phagemids was 2×10^{10} colony forming units (cfu), which corresponds to 0.03 pmoles of fusion protein. The number of phagemid bound, as determined by cfu, was compared to a control experiment in which no biotinylated Sbt12 was present. In the presence of biotinylated Sbt12, 0.01% of the input phagemids were eluted after the washing procedure. In the absence of biotinylated Sbt12 ~300-fold fewer phagemids were eluted after washing. The binding and washing conditions therefore were stringent enough that only a small percentage of pro-wt phagemids remained bound throughout.

These conditions were shown to selectively bind phagemids relative to M13K07 phage by performing the binding experiment as described above except that 2×10^{12} plaque forming units (pfu) M13K07 phage were added to the beads along with the 2×10^{10} cfu of phagemids. After washing, the eluent contained a ratio of 2.5 cfu:1 pfu, a 250-fold enrichment of the phagemids.

Mutant libraries

Quartets of amino acids in two regions of the pro-domain were chosen for randomization. Each randomized quartet encodes 20^4 (160,000) different sequences, thus most of the possible sequences are likely to be represented in the libraries of 1–2 million clones routinely produced in the PCR mutagenesis procedure. The selected quartets of amino acids do not have potential contacts with subtilisin (Fig. 1).

Selection for stabilizing mutations in the protein interior

The first mutagenesis and selection were carried out at a locus in the interior of the pro-domain. The side chains selected for mutagenesis were A23 and K27 in α-helix 1 and V37 and Q40 in β-strand 2 (Fig. 1B). Each amino acid side chain points toward the interior of the protein and contacts at least one of the other side chains. In round one of the selection, phagemids containing mutagenized pro-g3p fusion protein were added to biotinylated Sbt12 as described above. About 0.001% of the input phagemids were eluted after binding and washing. The phagemids eluted from the biotinylated Sbt12 were grown up and subjected to a second round of binding. In the second round 0.4% in the input phagemids bound. After three rounds of selection, 1% of the input phage bound. This is about 100-fold higher than the amount of pro-wt phagemids, which bound in the control experiment suggesting that tighter binding mutants of pro had been selected.

The DNA sequences of 12 pro-domain mutant phagemid were determined after no selection and after two and three rounds of selection (Table 1). The DNA sequences of the mutagenized sites after no selection reflect a random distribution of amino acids. After two rounds of selection, however, a strong consensus sequence emerges. Almost 75% of the sequences contain cysteine at both positions 23 and 40. The exact sequence C23,E27,L37,C40 was found five times. At least three of these were independent clones because three different codons for L37 were obtained. In the NNS (S = G or C) genetic code used in generating the random mutants, there are three codons for leucine but only one for cysteine and one for glutamate. After the third round of selection, C23 and C40 are found together in 21 of the 24 clones sequenced, and the exact sequence C23,E27,L37,C40 is observed 14 times. The genotypic diversity appears to decrease in the third round, how-

Table 1. *Sequences of phagemid mutants at locus 1*

Position	23	27	37	40	Number of
pro-wt	A	K	V	Q	occurrences
			Unselected		
	T	R	D	T	1
	E	C	L	P	1
	Q	L	V	A	1
	L	L	N	K	1
	M	Q	E	A	1
	V	I	R	A	1
	S	S	H	L	1
	Q	T	E	P	1
	A	N	P	E	1
	L	L	P	K	1
	K	M	F	Q	1
			2nd round		
	C	E	L (TTG)	C	2
	C	E	L (CTG)	C	1
	C	E	L (CTC)	C	2
	C	L (TTG)	P	C	2
	C	E	P	C	1
	C	M	L	C	1
	V	T	T	V	1
	D	E	V	M	1
	F	W	V	E	1
			3rd round		
	C	E	L (TTG)	C	10
	C	E	L (CTG)	C	4
	C	L	P	C	2
	C	E	T	C	1
	C	L	L	C	1
	C	L	M	C	1
	C	S	L	C	1
	C	G	I	C	1
	L	S	L	H	1
	S	E	V	V	1
	E	E	T	V	1

ever, so selection results must be weighted in light of the possibility that some clones may become over-represented from "founder effects" in the growth of the phagemid particles.

Computer modeling of the consensus amino acids shows that the sulfurs of C23 and C40 can come within 2.1 Å to form a disulfide cross-link (c1 = −180°, c2 = 142°, c(SS) = −104°, c2′ = −160°, c1′ = 72°). L37 makes good van der Waals contacts with I30 and A47. The carboxylate of E27 would appear from modeling to protrude well into solvent and make no protein–protein hydrogen bonds or salt bridges. It is not evident why E27 would be strongly preferred over K, Q, D, or N.

Characterization of the consensus sequence

We synthesized the mutant C23,E27,L37,C40 (denoted pro-R1) to assess its conformation stability. Circular dichroic spectra were used to determine the apparent equilibrium constant for folding. Analysis of the stability was carried out in 100 mM NaOAc, pH 5.0 at [P] = 50 μM. Previous studies have shown that stabilized pro-domains have a tendency to dimerize at pH 7.0 (Ruvinov et al., 1997). Based on the equilibrium sedimentation data, pro-R1 is 99% monomeric under these conditions at pH 5.0. The equilibrium constant for folding in 100 mM NaOAc, pH 5.0 is the same as the intrinsic equilibrium constant for folding (independent of dimerization) in 100 mM KPI, pH 7.0 (Ruvinov et al., 1997). The CD spectrum of pro-wt is typical of a largely random coil structure with a minimum ellipticity at 198 nM (Fig. 2). The mean residue ellipticity at 222 nM is −2,000 deg cm^2 dmol^{-1} at 25 °C. This value is characteristic of a largely random coil structure (Goodman & Kim, 1989; Merutka et al., 1991). The mean residue ellipticity of pro-R1 at 222 nM is −5,100 deg cm^2 dmol^{-1} (Fig. 2). The difference in ellipticity at 222 nM of the pro-domain-subtilisin complex minus subtilisin is also equal to −5,100 deg cm^2 dmol^{-1}. This value is taken to be the ellipticity of the native pro-domain at 25 °C. In the presence of 1mM DTT, the CD spectrum of pro-R1 is similar to the spectrum of pro-wt, indicating that pro-R1 is largely unfolded when the disulfide cross-link is broken.

Analysis of the temperature unfolding profile

Circular dichroism

Temperature unfolding profiles of pro-wt and pro-R1 are presented in Figure 3. The ellipticity of pro-wt at 222 nM decreases linearly with temperature with Δmre = −27 deg cm^2 dmol^{-1} °C^{-1}. This temperature dependence is typical of random polypeptides and small proteins in high Gu-HCl (Goodman & Kim, 1989; Merutka et al., 1991), indicating that the pro-wt is unfolded at all temperatures. The profile of pro-R1 shows that unfolding occurs over the temperature range of 40 to 80 °C. To calculate the mid-point of thermal denaturation, pro-wt is used as the baseline for the unfolded state. The fraction native then is determined by subtracting the signal minus unfolded baseline and dividing by the total CD difference between 100% folded and 0% folded at that temperature. The mid-point of the melting transition is ~65 °C. The unfolding reaction of pro-R1 is 90% reversible after scanning to 90 °C.

Differential scanning calorimetry

The amount of excess heat absorbed by a protein as the temperature is increased through a transition from the folded to un-

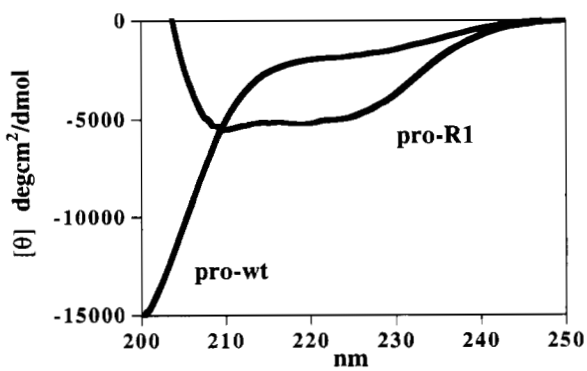

Fig. 2. CD spectra of pro-wt and Pro-R1. Mean residue ellipticity (deg cm^2/dmol) is plotted vs. wavelength. Spectra were measured in 0.1 M NaOAc, pH 5.0 using a 1 mm cylindrical cuvette at 25 °C with [P] = 50 μM.

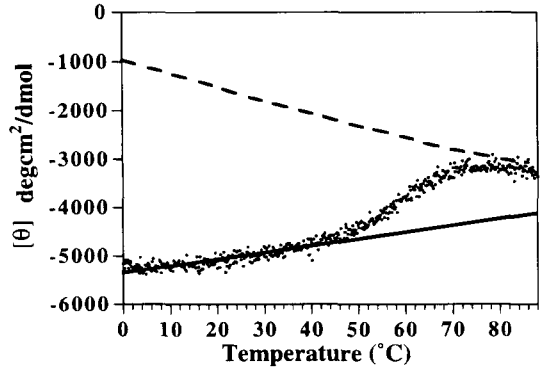

Fig. 3. Temperature unfolding profile for Pro-R1 at pH 5. Mean residue ellipticity (deg cm^2 $dmol^{-1}$) at 222 nM is plotted vs. temperature in the range from 0 to 85 °C. The temperature profile was recorded using a 1 mm cylindrical cuvette with protein concentration 50 μM in 100 mM NaOAc, pH 5.0. The dashed line is the temperature profile for pro-wt. The solid line is an extrapolated baseline for the folded state.

folded state at constant pressure provides a direct measurement of the ΔH of unfolding (Privalov & Khechinashvili, 1974). Calorimetric data were obtained for pro-R1 in 100 mM NaOAc pH 5.0 (Fig. 4). The midpoint of denaturation under these conditions is 65 °C. Both the calorimetric and v'ant Hoff enthalpies are 48 kcal/mol at 65 °C.

If a protein unfolds in a two-state manner, then the temperature dependence of the unfolding reaction will be determined by the thermodynamic state functions ΔH, ΔS, and ΔCp according to the Gibbs-Helmholtz equation:

$$\Delta G = \Delta H_o - T\Delta S_o + \Delta Cp(T - T_o - T \ln T/T_o), \quad (3)$$

where ΔH_o and ΔS_o are the enthalpy and entropy of unfolding evaluated at a reference temperature T_o, and ΔG is the free energy of unfolding at a temperature T (Brandts, 1964; Privalov, 1979; Becktel & Schellman, 1987).

The DSC data were converted to $\Delta G_{\text{unfolding}}$. A stability curve ($\Delta G_{\text{unfolding}}$ vs. T) was calculated using Equation 3 and the enthalpy and entropy of unfolding obtained by DSC (ΔH_{339} = 48.8 kcal/mol, ΔS_{339} = 0.144 kcal/mol) (Fig. 5). The change in heat capacity ($\Delta Cp_{\text{unfolding}}$ = 0.83 kcal/mol) was determined from the curvature of the free energy profile (Privalov 1979; Becktel & Schellman, 1987). The $\Delta G_{\text{unfolding}}$ at 20 °C is 4 kcal/mol. The selected mutant is 6 kcal/mol more stable than pro-wt, which has a $\Delta G_{\text{unfolding}}$ of −2.1 kcal/mol at 20 °C. Most of the increase in stability is attributable to the disulfide cross-link since reduction of the disulfide with 1 mM DTT unfolds the protein almost completely (see below).

Analysis of pro-domain-facilitated subtilisin folding

The isolated pro-domain has been shown in vitro to accelerate subtilisin folding in a bimolecular reaction. Since none of the mutations in the pro-domain directly contact subtilisin in the complex, their effects on accelerating the folding of subtilisin are linked to whether or not they stabilize the native conformation of the pro-domain (Wang et al., 1998). This linkage allows us to use subtilisin folding as a functional assay to detect even small changes in pro-domain stability. We have previously engineered an inactive mutant of subtilisin, denoted Sbt15, for use in vitro folding studies (Bryan et al., 1992). Sbt15 is identical to the Sbt12 subtilisin used in phagemid selection except amino acids 75 to 83 are deleted from Sbt15. This deletion accelerates in the vitro folding of subtilisin and facilitates characterization of pro-domain mutations (Strausberg et al., 1993; Wang et al., 1998).

The effect of the selected mutations was determined by mixing unfolded Sbt15 with an excess of pro-domain and measuring the rates and amplitudes of fluorescence changes during a single turnover of subtilisin folding (Strausberg et al., 1993; Bryan et al., 1995; Wang et al., 1995). The folding reaction of subtilisin, in the presence of pro-domain, can be followed by an increase in tryptophan fluorescence of 1.5-fold due to changes in the environments of the three tryptophans in subtilisin upon its folding and binding of the pro-domain. The pro-domain does not contain tryptophan residues and thus has no intrinsic fluorescence at 345 nM upon excitation at 300 nM. Therefore fluorescence increases observed at 345 nM are due solely to the conversion of unfolded Sbt15 to a folded complex with pro-domain. The folding reaction was followed using $[S_u]$ = 1 μM and $[P]$ = 5 μM in 30 mM Tris, 5 mM

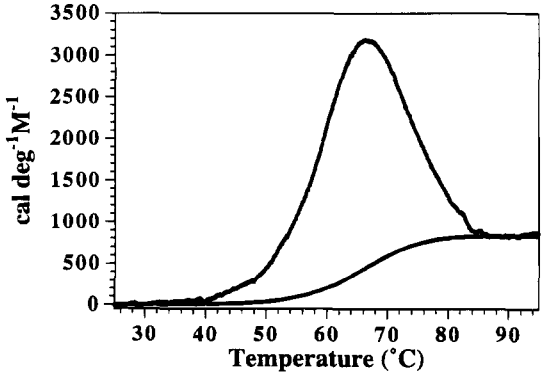

Fig. 4. DSC of pro-R1. Calorimetric data were obtained for a solution of 0.7 mL 100 mM NaOAc pH 5.0 with [pro-R1] = 8.5 mg/mL. Total amount of protein was 700 nmol.

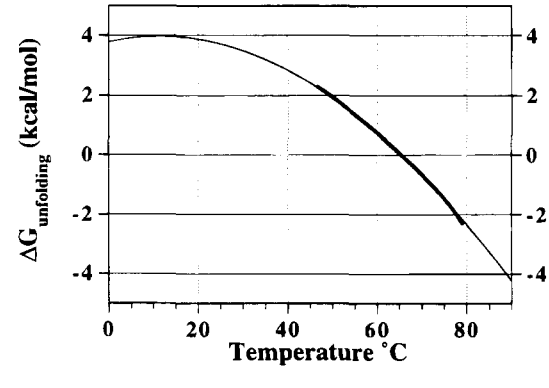

Fig. 5. $\Delta G_{\text{unfolding}}$ as a function of temperature for pro-R1. The temperature unfolding profile measured by DSC for pro-R1 (Fig. 4) was converted to an apparent $\Delta G_{\text{unfolding}}$ (thick line). The thin line is a theoretical curve calculated using the Gibbs-Helmholtz equation: $\Delta G_{\text{unfolding}} = \Delta H_o - T\Delta S_o + \Delta Cp(T - T_o - T \ln T/T_o)$, where T_o = 339 K, ΔH_o = 48.8 kcal/mol, ΔS_o = 0.144 cal/mol, and ΔCp = 0.83 kcal/mol.

KPi, pH 7.5 at 25 °C. Under these conditions a single cycle of Sbt15 folding is measured. The product is a one to one complex of Sbt15 and pro-domain. The folding curve for Sbt15 with 5 μM pro-wt can be fit with a single exponential equation with a rate constant of 0.0034 s^{-1} (Fig. 6A).

The folding reactions of Sbt15 with pro-R1 was much faster than with pro-wt but showed biphasic changes in fluorescence (Fig. 6B). The kinetic behavior is consistent with pro-R1 being fully folded and will be described in detail elsewhere (B. Ruan, J. Hoskins, & P.N. Bryan, in prep.). In the presence of 1 mM DTT, which breaks the disulfide cross-link, the reaction is much slower and follows single exponential kinetics with a rate of 0.0033 s^{-1}. This indicates that the stability of pro-R1 without the disulfide cross-link is similar to pro-wt.

Select of a turn sequence

A second selection was carried out at a locus near the surface of the pro-domain. The side chains selected for mutagenesis were S31, G34, G35, N52. G34 and G35 make a tight turn between α-helix 1 and the β3 strand (Fig. 2C). The main-chain bond angles of G34 (ϕ = 95° and ψ = 16°) are energetically unfavorable for amino acids other than glycine. S31 is in α-helix 1 and was chosen because of the potential of its side chain to interact with a side chain at 34 and N52 is the helix capping amino acid for α-helix 2 and was chosen because of the potential of its side chain to interact with a side chain at 35.

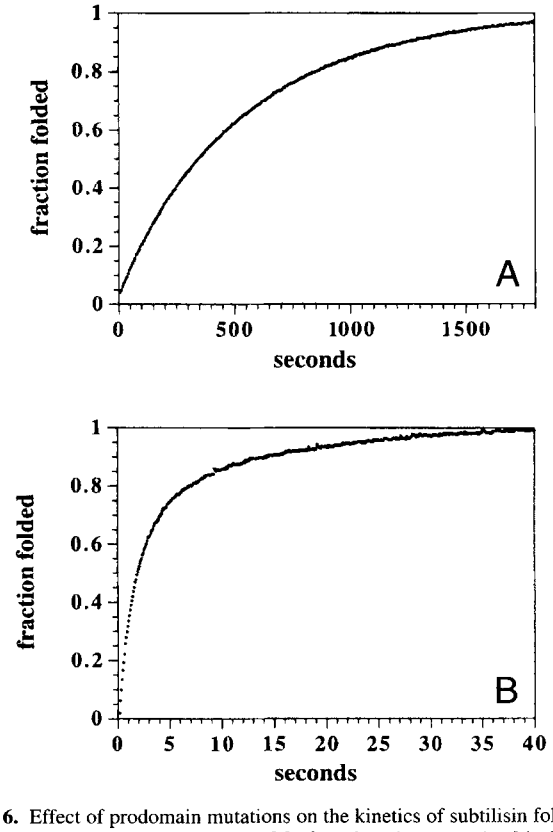

Fig. 6. Effect of prodomain mutations on the kinetics of subtilisin folding. The 1 μM denatured Sbt15 and 5 μM of prodomain were mixed in 5 mM KPi, 30 mM Tris-HCl, pH 7.5 at 25 °C. The reaction was followed by the increase in tryptophan fluorescence at 345 nM. **A:** Folding curve for Sbt15 with pro-wt. **B:** Folding curve for Sbt15 with pro-R1.

In round one of the selection, about 0.001% of the input phagemids were eluted after binding and washing. This was the same percentage bound as in the selection at the interior locus, however, after three and four rounds of selection only 0.005% of the input phage bound.

After four rounds of selection, 17 phagemid clones were sequenced. No consensus sequence was strongly preferred (Table 2). Phenylalanine occurred 15% of the time (about three times as often as it would occur in a random sequence). This over-representation may result from binding of subtilisin to the unfolded state of the pro-domain. Phenylalanine is a preferred amino acid in the P1 substrate binding pocket of subtilisin. The wild-type sequence G34,G35,N52 was found five times, four times with I31, and once with L31. All of the I31,G34,G35,N52 phagemids had the same DNA sequence, so they may not be independent isolates.

We synthesized the I31 and L31 mutants to assess their stability. Both of the mutants appeared mostly unfolded by CD. To assess small changes in the folding equilibrium, we measured the rate of subtilisin folding in the presence of 5 μM of each mutant. The subtilisin folding rates observed with S31I and S31L were 0.0014 and 0.001 s^{-1}, respectively. The better of the two, S31I, accelerated subtilisin folding 2.5 times slower than pro-wt. We have shown previously for mutants of low stability that the rate of subtilisin folding is a function of the fraction of folded pro-domain (Wang et al., 1998). Therefore we can calculate free energy of unfolding for each mutant by comparing its ability to accelerate Sbt15 folding with that of pro-wt, which has $\Delta G_{\text{unfolding}} = -2.1$ kcal/mol. The change in $\Delta G_{\text{unfolding}}$ is equal to $RT \ln(k_{mut}/k_{wt})$, where k_{mut} is the subtilisin folding rate with the mutant and k_{wt} is subtilisin folding rate with pro-wt. From this we calculate that S31I is 0.5 kcal/mol less stable than pro-wt and S31L is 0.7 kcal/mol less stable than pro-wt.

We suspect that the pro-wt sequence is the most stable combination for this quartet. Our failure to identify the pro-wt sequence may be the result of the stringency of our selection, which was designed to identify pro-domains that bind more strongly than pro-wt. Since only 0.01% of pro-wt phagemids are retained throughout the washing procedure, discrimination between sequences bind-

Table 2. *Sequences of phagemid mutants at locus 2*

Position	31	34	35	52	Number of
pro-wt	S	G	G	N	occurrences
			4th round		
	I	G	G	N	4
	L	G	G	N	1
	D	F	F	Y	1
	L	F	S	R	1
	G	F	Y	F	1
	F	V	W	F	1
	E	Q	R	T	1
	S	S	Q	F	1
	P	R	G	L	1
	L	W	N	H	1
	F	S	E	F	1
	C	A	E	S	1
	T	G	E	T	1
	S	A	N	V	1

ing weaker than pro-wt and adventitious binding is difficult. Nevertheless, we were able to detect two sequences that were within 1 kcal/mol of pro-wt above background binding and both of these sequences contained G34, G35, and N52. This suggests that these are the preferred amino acids.

Discussion

Studies using combinatorial mutagenesis have been very useful in studying the relationship between amino acid sequence and protein structure. One combinatorial approach is to create random libraries with a common pattern of polar and apolar amino acids and determine what fraction of the library attains a degree of foldedness. Several of these studies suggest that a binary code of polar and nonpolar amino acids sometimes is sufficient to encode secondary structure and some level of tertiary structure (Kamtekar et al., 1993; West & Hecht, 1995). In at least one case a protein designed by binary patterning is very native-like (Roy et al., 1997). Another type of combinatorial experiment assesses how many different sequences can be tolerated in a defined region of a protein without destroying function (Reidhaar-Olson & Sauer, 1988; Lim & Sauer, 1991; Gu et al., 1995; Cordes et al., 1996; Sauer, 1996). These experiments show that many different combinations of amino acids can be accommodated in a selected region of a protein without destroying function. It would be erroneous to conclude from these results, however, that all combinations of tolerated amino acids are roughly equivalent in stability. In fact, these experiments usually select for mutants that are less stable than the wild-type protein. For example, Gu et al. (1995) have used phage display to select for folded variants of an IgG binding protein. They found that most of their folded variants denature just above room temperature. This is the minimum stability that would not affect binding. The fact that most selected mutants have this stability indicates that they occur more frequently than more stable sequences.

Our study indicates that the number of amino acid combinations close to the maximum stability is small. The most striking example is the selection at the interior locus 23, 27, 37, 40. Sequences with C23 and C40 occur much more frequently than any other combination. The most frequently selected sequence has a $\Delta G_{\text{unfolding}}$ = 4 kcal/mol at 25 °C, an increase of 6 kcal/mol relative to the naturally occurring sequence. The presence of a disulfide is the dominant energetic contribution because the mutant unfolds in the presence of a reducing agent that breaks the cross-link. A disulfide cross-link requires that two side chains interact in a very limited number of arrangements, which creates a favorable entropic contribution unique to disulfides. On the other hand, the stereochemical requirements for disulfide bonds are very restrictive making it difficult to find places in proteins where the full effect of entropic stabilization due to a disulfide will be manifest. On the basis of the selection at only one locus, one might conclude that a stabilizing disulfide is a special case and the selection of a small number of sequences is biased by overriding energetic advantages of a disulfide.

Perhaps more surprising was the selection of a conserved sequence at the surface locus. The wild-type G34,G35,N52 sequence was the only sequence found more than once among the 17 phagemids sampled, even though there is no obvious topological reason why the expanded backbone flexibility of two glycines would be required to make this turn.

The overall conclusion is that the combinations of sequences at a structural locus that are close to the maximum stability is fairly restricted. Further, the information content of the most stable combinations toward specifying a particular fold is higher than for merely tolerable combinations. Given that energetic contributions of independent loci to protein stability are additive, tolerable combinations at a locus are defined in terms of the information content at all other independent loci. A good example of this was an experiment in which apolar amino acids within the core of T4 lysozyme were replaced progressively with methionine (Gassner et al., 1996). Each progressive substitution destabilized the protein, but up to 10 methionine substitutions were tolerated without unfolding the molecule. One interesting conclusion of this experiment emphasized by the authors is that a simplified hydrophobic core is sufficient to retain native properties of T4 lysozyme. We would also note that the stability of wild-type T4 lysozyme is 16 kcal/mol at 25 °C (Kitamura & Sturtevant, 1989). Therefore, in the context of all other loci, T4 lysozyme can tolerate a loss in stability of 7 kcal/mol as a result of the 10 methionine substitutions in its core and maintain its native structure. Other studies have shown that computer-designed hydrophobic cores of equal volume are not equal in stability but rather result in a range of stabilities (Desjarlais & Handel, 1995; Lazar et al., 1997).

Several factors confound deciphering folding codes. First, folding information is diffuse because of the additive nature of individual energetic contributions of independent loci (Lattman & Rose, 1993). Secondly, natural proteins do not have the optimum combinations at all loci. (Optimum is used here to mean the most stable combinations.) Most natural proteins generally have $\Delta G_{\text{unfolding}}$ between 5–15 kcal/mol. It is possible that many natural proteins do not have optimal sequences at any loci that would explain why proteins with nearly identical backbone topologies have no discernible sequence identity.

Available evidence suggests that the amount of sequence information needed to specify a unique backbone topology lies between the tolerable (e.g., the binary code) and the optimal. Accordingly, one approach to protein design is to construct the correct binary pattern of polar and apolar amino acids and then superimpose on that pattern several specific interactions that would promote the correct tertiary fold and uncode competing folds. This may not be the best pathway to a unique fold, however. The pro-wt has essentially the same binary pattern of amino acids as the stable pro-R1 mutant, yet has no detectable secondary structure by CD or two-dimensional NMR (J. Orban, unpubl. results). The sequence is only a few kcal/mol away from being a stable fold, however. The pro-R1 mutant illustrates that optimizing one set of local interactions results in a stable unique fold. We have shown previously that introduction of three mutations in different regions of the pro-domain stabilizes the native fold by 2.5 kcal/mol (Ruvinov et al., 1997). An overemphasis on the binary pattern may result in designed proteins in which secondary structural elements have high independent stability at the expense of the stability of specific tertiary interactions (Cordes et al., 1996). Such proteins would be expected to exhibit the molten-like behavior observed for most proteins of binary design.

Apart from possible implications on the folding code, the phage display selection described here should be useful in optimizing the stability of other proteins, provided that the protein can be displayed on the phage surface and that a binding property can be used to select for the native fold. The number of proteins that have been displayed in active form on the phage surface has grown rapidly (Kay et al., 1996). In the case of an enzyme, the binding target could be to a substrate or transition state analog. The pro-

cedure would be first to introduce a destabilizing mutation (remote from the binding site) to unfold the protein and thus indirectly weaken binding to the target. Next, amino acids at an independent locus would then be randomized. Finally, rehabilitated proteins would be selected by phage display based on improved binding to the target. Given the additive contributions of individual interactions to proteins stability and the fact that most proteins are not optimally stable, this procedure should allow selection of the most stabilizing combinations of amino acids at the desired locus.

Material and methods

Vector construction

The fd gene III fusion phagemid pHEN1 (Hoogenboom et al., 1991) was used to produce fusion phagemid particles displaying the pro-domain g3p fusion protein on their surfaces. In pHEN1 the g3p fusion is under the transcriptional control of the lacZ promoter, and the fusion protein contains a pelB signal sequence to direct the fusion protein to the periplasm. The feasibility of the method was first demonstrated by constructing and producing wild-type pro-domain g3p fusion phage. The coding region of the wild-type pro-domain (pro-wt) was amplified by PCR using 5′ and 3′ flanking oligonucleotides containing PstI and NotI sites respectively, digested with PstI and NotI, ligated into pHEN1 to yield pHEN1-Pro-wt, and transformed by electroporation into *Escherichia coli* TG-1. PCR analysis was used to identify ampicillin resistant transformants containing the pro-domain coding region and the fusion construct was verified by DNA sequencing.

Phagemid growth and rescue

A fresh transformant colony of TG-1 containing pHEN1-pro-wt was used to inoculate 2 mL of LB, 100 μg/mL ampicillin, 1% glucose and grown to mid-log phase at 37 °C. One milliliter of log phase culture was used to inoculate 25 mL of LB, 100 μg/mL ampicillin, and phagemid particles rescued by the addition of M13K07 helper phage to 2×10^8 plaque forming units (pfu)/mL. Following 1.5 h, kanamycin was added to 75 μg/mL and the culture allowed to grow overnight at 37 °C. Phagemid particles were purified and concentrated from the culture supernatant by two polyethylene glycol precipitations and resuspended in 10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0 to 2×10^{12} colony forming units (cfu)/mL.

Construction of mutagenic library

To illustrate to mutagenesis techniques, the construction of the random library at positions 23, 27, 37, 40 is described in detail. The random library of pro-domain genes was constructed synthetically by the sequential annealing and extension of five oligonucleotides. Two of the oligonucleotides contain the random codon NNS where N represents any of the four nucleotides and S represents either G or C to mutagenize A23, K27, V37, and Q40 of the pro-domain;

JHPP.01, 5′-CAATGTCGACGATGAGCNNSGCTAAGAAGNNS GATGTCATTTCTGAAAAAGGC-3′
JHPP.02, 5′-TGCGTCTACATATTTGAASNNCTTTTGSNNTTT CCCGCCTTTTTCAGAAATG-3′
JHPP.03, 5′-GCAGGGAAATCAAACGGGGAAAAGAAATATA TTGTCGGGTTTAAACAGACAATGTCGACGATG AGC-3′
JHPP.04, 5′-TTTCAATTCTTTTACAGCTTTTTCGTTTAATGT AGCGCTAGCTGCGTCTACATATTTGAA-3′
JHPP.05, 5′-GTACGCATGTGCTACGTGATCTTCTTCAACGT AAGCGACGCTCGGATCCTTTTTCAATTCTTTT ACAGC-3′.

The final extension product was PCR amplified using 5′ and 3′ flanking oligonucleotides containing PstI and NotI sites, respectively, digested with PstI and NotI, agarose gel purified, and small molecular weight impurities removed using a 100,000 MWCO ultrafiltration concentrator. The purified pro-domain DNA was ligated into 2 μg of PstI and NotI digested pHEN1, split into ten fractions and used to transform *E. coli* TG-1 by electroporation yielding 1×10^6 independent transformants. Transformants were amplified by overnight growth in 1 L of LB, 100 μg/mL ampicillin, 1% w/v glucose at 35 °C. Phagemid particles were prepared essentially as described above except that 0.5 mL of the overnight amplified culture was used to inoculate 50 mL of LB, 100 μg/mL ampicillin, 1% w/v glucose, and grown at 37 °C to establish the midlog phase culture. DNA sequence analysis was used to verify the randomized codons.

Affinity selection

For selection of phagemid particles displaying pro-domain-g3p fusion protein, a selection procedure similar to that described by Hawkins et al. (1992) was used. The 2×10^{10} cfu of phage stock was diluted into 1 mL of MPBT (3% w/v skimmed milk powder, 0.1 M KPi, pH 7.0, 0.5% v/v Tween 20) containing 10 nM biotinylated subtilisin Sbt12 and incubated at room temperature for 3 h on an inclined rotating wheel. Streptavidin-coated magnetic beads (6×10^7 beads, Dynabeads M-280, Dynal) were added and incubated for 15 min. Magnetic beads were collected on a magnetic particle concentrator (Dynal MPC) and washed seven times with 1 mL of MPBT. Phagemid particles were eluted by a 5 min incubation in 133 μL of 50 mM triethylamine, pH 11.1 and then neutralized with 50 μL of 1.0 M KPi, pH 7.0. The eluted phage were used to infect 0.5 mL of stationary TG-1, amplified by overnight growth in 50 mL of LB, 100 μg/mL ampicillin, 1% w/v glucose at 37°, and phagemid particles prepared as described above. The affinity selection was repeated for three to four rounds of selection. Selected mutants of the pro-domain were subcloned, expressed, and purified as described (Strausberg et al., 1993).

Kinetics of catalyzed subtilisin folding

Refolding of subtilisin was followed by monitoring changes in tryptophan fluorescence (excitation λ = 300, emission λ = 345) using a SPEX FluoroMax spectrofluorimeter for manual mixing experiments and a KinTek Stopped-Flow Model SF2001 for rapid kinetic measurements as described (Strausberg et al., 1993). The subtilisin, Sbt15, used in this study was an engineered version which was catalytically impaired by mutating the active-site serine 221 to cysteine and was modified to eliminate the high-affinity calcium site A. Details of the construction and characterization of this mutant are described in Bryan et al. (1992).

Circular dichroism

Circular dichroism (CD) measurements were performed with a Jasco spectropolarimeter, model J-720 as described (Ruvinov et al., 1997).

Differential scanning calorimetry

Differential scanning calorimetry (DSC) measurements were performed with a Hart 7707 DSC heat conduction scanning microcalorimeter interfaced with an IBM personal computer as described (Alexander et al., 1992). The temperature was increased from 0 to 95 °C at a scan rate of 30 °C/h. The solution mass of all protein and control solutions was near 0.70 g per ampoule. The scans were done in 100 mM NaOAc, pH 5.0. Samples to be scanned were prepared by rehydrating lyophilized protein in the appropriate buffer and dialyzing against the same buffer. The concentration of the dialyzed protein was determined by UV absorbance using the 1 mg/mL = A_{275nM} of 0.67. The number of nanomoles of protein ranged from 400 to 800, corresponding to 5–10 mg/mL.

Acknowledgments

The authors wish to thank Richard Prescott for synthesizing the oligonucleotides used in site-directed mutagenesis and DNA sequencing, and Patrick Alexander and Susan Strausberg and Michael Hecht for useful discussion. This work was supported by NIH grant GM42560.

References

Alexander P, Fahnestock S, Lee T, Orban J, Bryan P. 1992. Thermodynamic analysis of the folding of the Streptococcal Protein G IgG-binding domains B1 and B2: Why small proteins tend to have high denaturation temperatures. *Biochemistry 31*:3597–3603.

Becktel WJ, Schellman JA. 1987. Protein stability curves. *Biopolymers 26*:1859–1877.

Betz SF, Raleigh DP, DeGrado WF. 1993. De novo protein design: From molten globules to native-like states. *Curr Opin Struct Biol 3*:601–610.

Brandts JF. 1964. Thermodynamics of protein denaturation. II. A model of reversible denaturation and interpretations regarding the stability of chymotrypsinogen. *J Am Chem Soc 86*:4302–4314.

Bryan P, Alexander P, Strausberg S, Schwarz F, Wang L, Gilliland G, Gallagher DT. 1992. Energetics of folding subtilisin BPN′. *Biochemistry 31*:4937–4945.

Bryan P, Wang L, Hoskins J, Ruvinov S, Strausberg S, Alexander P, Almog O, Gilliland G, Gallagher TD. 1995. Catalysis of a protein folding reaction: Mechanistic implications of the 2.0 Å structure of the subtilisin-prodomain complex. *Biochemistry 34*:10310–10318.

Cordes MHJ, Davidson AR, Sauer RT. 1996. Sequence space folding and protein design. *Curr Opin Struct Biol 6*:3–10.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

Gallagher TD, Gilliland G, Wang L, Bryan P. 1995. The prosegment-subtilisin BPN′ complex: Crystal structure of a specific foldase. *Structure 3*:907–914.

Gassner NC, Baase WA, Matthews BW. 1996. A test of the jigsaw puzzle model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc Natl Acad Sci USA 93*:12155–12158.

Goodman EM, Kim PS. 1989. Folding of a peptide corresponding to the alpha-helix in bovine pancreatic trypsin inhibitor. *Biochemistry 28*:4343–4347.

Gu H, Qian Y, Bray ST, Riddle DS, Shiau AK, Baker D. 1995. A phage display system for studying the sequence determinants of protein folding. *Protein Sci 4*:1108–1117.

Hawkins RE, Russell SJ, Winter G. 1992. Selection of phage antibodies by binding affinity: Mimicking affinity maturation. *J Mol Biol 226*:889–896.

Hoogenboom HR, Griffiths AD, Johnson KS, Chiswell DJ, Hudson P, Winter G. 1991. Multi-subunit proteins on the surface of filamentous phage: Methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res 19*:4133–4137.

Ikemura H, Takagi H, Inouye M. 1987. Requirement of pro sequence for the production of active subtilisin in *Escherichia coli*. *J Biol Chem 262*:7859–7864.

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary pattern of polar and nonpolar amino acids. *Science 262*:1680–1685.

Kay BK, Winter J, McCafferty J. 1996. *Phage display of peptides and proteins*. San Diego, CA: Academic Press Inc.

Kitamura S, Sturtevant JM. 1989. A scanning calorimetric study of the thermal denaturation of the lysozyme of phage T4 and the Arg 96 → His mutant form thereof. *Biochemistry 28*:3788–3792.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Lattman EE, Rose GD. 1993. Protein folding: What's the question? *Proc Natl Acad Sci USA 90*:439–441.

Laurents DV, Subbiah S, Levitt M. 1994. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci 3*:1938–1944.

Lazar GA, Desjarlais JR, Handel TM. 1997. De novo design of the hydrophobic core of ubiquitin. *Protein Sci 6*:1167–1178.

Lim WA, Sauer RT. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol 219*:359–376.

Merutka G, Shalongo W, Stellwagen E. 1991. A model peptide with enhanced helicity. *Biochemistry 30*:4225–4248.

Power SD, Adams RM, Wells JA. 1986. Secretion and autoproteolytic maturation of subtilisin. *Proc Natl Acad Sci USA 83*:3096–3100.

Privalov PL. 1979. Stability of proteins small globular proteins. *Adv Protein Chem 33*:167–241.

Privalov PL, Khechinashvili NN. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *J Mol Biol 86*:665–684.

Reidhaar-Olson JF, Sauer RT. 1988. Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science 241*:53–57.

Roy S, Ratnaswamy G, Boice JA, Fairman R, McLendon G, Hecht MH. 1997. A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties. *J Am Chem Soc 119*:5302–5306.

Ruvinov S, Wang L, Ruan B, Almog O, Gilliland G, Eisenstein E, Bryan P. 1997. Engineering the independent folding of the subtilisin BPN′ prodomain: Analysis of two-state folding vs. protein stability. *Biochemistry 36*:10414–10421.

Sauer RT. 1996. Protein folding from a combinatorial perspective. *Fold Design 1*:27–30.

Smith GP, Scott JK. 1993. Libraries of peptides displayed on filamentous phage. *Methods Enzym 217*:228–257.

Strausberg S, Alexander P, Wang L, Schwarz F, Bryan P. 1993. Catalysis of a protein folding reaction: Thermodynamic and kinetic analysis of subtilisin BPN′ interactions with its propeptide fragment. *Biochemistry 32*:8112–8119.

Vasantha N, Thompson LD, Rhodes C, Banner C, Nagle J, Filpula D. 1984. Genes for alkaline and neutral protease from Bacillus amyloliquifaciens contain a large open-reading frame between the regions coding for signal sequence and mature protein. *J Bacteriol 159*:811–819.

Wang L, Ruan B, Ruvinov S, Bryan PN. 1998. Engineering the independent folding of the subtilisin BPN′ pro-domain: Correlation of pro-domain stability with the rate of subtilisin folding. *Biochemistry 37*:3165–3171.

Wang L, Ruvinov S, Strausberg S, Gallagher TD, Gilliland G, Bryan P. 1995. Prodomain mutations at the subtilisin interface: Correlation of binding energy and the rate of catalyzed folding. *Biochemistry 34*:15415–15420.

Wells JA, Ferrari E, Henner DJ, Estell DA, Chen EY. 1983. Cloning sequencing and secretion of Bacillus amyloliquifaciens subtilisin in *Bacillus subtilis*. *Nucleic Acids Res 11*:7911–7925.

West MW, Hecht MH. 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci 4*:2032–2039.