Structure 14, 1499–1510, October 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.str.2006.08.006

Assessing Computational Amino Acid β-Turn Propensities with a Phage-Displayed Combinatorial Library and Directed Evolution

Hung-Ju Hsu,[1,2] Hong-Ju Chang,[1,2] Hung-Pin Peng,[1] Shan-Sheng Huang,[1] Ming-Yen Lin,[1] and An-Suei Yang[1,*]
[1] Genomics Research Center
Academia Sinica
128 Academia Road, Section 2
Taipei, 115
Taiwan

Summary

Structure propensities of amino acids are important determinants in guiding proteins' local and global structure formation. We constructed a phage display library—a hexa-HIS tag upstream of a CXXC (X stands for any of the 20 natural amino acids) motif appending N-terminal to the minor capsid protein pIII of M13KE filamentous phage—and developed a novel directed-evolution procedure to select for amino acid sequences forming increasingly stable β-turns in the disulfide-bridged CXXC motif. The sequences that emerged from the directed-evolution cycles were in good agreement with type II β-turn propensities derived from surveys of known protein structures, in particular, Pro-Gly forming a type II β-turn. The agreement strongly supported the notion that β-turn formation plays an active role in initiating local structure folding in proteins.

Introduction

β-turn is the most abundant nonrepetitive secondary structure among protein structures. A recent survey of known protein structures shows that 44% of total protein structures belong to coil regions, which adopt neither α helix nor β strand (Fitzkee et al., 2005). A closer examination of these coil regions reveals that 53% of these coil residues participate in forming β-turns (Panasik et al., 2005). The findings reemphasize that β-turns are critical determinants in protein stability and folding (Richardson, 1981; Rose et al., 1985).

Statistical surveys of protein structural database have indicated that sequence preferences for each of the different types of β-turn (mainly, type I, II, I′, and II′) are strongly type dependent (Hutchinson and Thornton, 1994), although the intrinsic origin of the sequence preferences is largely unknown. As in any statistical analysis of protein sequence-structure relationships, the amino acid preferences can be rationalized in two mutually exclusive interpretations: either the amino acid residues are actively involved in guiding substructure formation (the active model for sequence preference), or these amino acid residues are the ones best tolerated in the passively formed substructures (the passive model for sequence preference). Statistical analyses alone frequently do not distinguish one model from the other. Hence, model systems that isolate context-dependent factors from the intrinsic sequence preferences of the substructures are required to verify the statistical sequence preferences and to elucidate the underlying physical principles.

In contrast to the appealing active model for the sequence preferences in α helix (Chakrabartty and Baldwin, 1995) and β sheet (Minor and Kim, 1994), it has been controversial in interpreting the statistical sequence preferences in β-turns observed in protein structures whether to use the active or the passive model. On the one hand, β-turns are thought to be involved in protein-folding initiation (see, for example, Falcomer et al. [1992], Rose et al. [1985], Searle and Ciani [2004], Yang et al. [1996], and references therein); as shown in the zip-up folding of the β-hairpin, the folding process initially involves the formation of a β-turn in the model peptide (Bonvin and van Gunsteren, 2000; Munoz et al., 1997). However, evidence from molecular-folding simulations supports the hydrophobic collapse folding model in a different model peptide system (Dinner et al., 1999; Pande and Rokhsar, 1999), where turn formation occurs only after the stabilization of the flanking strands. On the other hand, turns are thought to be quite tolerant of sequence variations without significantly compromising protein structure and stability (Castagnoli et al., 1994). However, experimental evidence has also suggested that sequence preferences in β-turns do contribute to protein stability and folding, and the folding free energy contribution correlates to the occurrence frequency of the amino acids (Searle, 2004; Simpson et al., 2005).

In this work, a biological combinatorial host-guest system was developed to address the following questions: (1) Are there sequence propensities for the second and third turn positions in isolated model type I and type II β-turns? (2) If the sequence propensities do exist, how do these sequence propensities compare with the frequencies of amino acid occurrence in β-turns of known protein structures? And (3) does the active model or passive model account for the β-turn propensities? Type I and type II β-turns were chosen for two reasons: first, these two types of turns are the most prevalent nonrepetitive secondary structures in proteins (type I turn is the most abundant β-turn [46%] in protein structures, followed by type II turn [19%]; Panasik et al., 2005); second, while type I′ and II′ β-turns have been studied extensively with antiparallel β-hairpins as the host systems (Searle, 2004; Simpson et al., 2005), type I and II β-turns are not compatible with antiparallel β-hairpin's right-handed twist and have thus been neglected for lacking a well-established host system.

In principle, the host system needs to be as small as possible so that turn stability factors resulting from long-range interactions can be minimized. In addition, the turn conformation needs to be confined by the host system so that the modulation of the turn conformation due to sequence replacement can be limited. From these perspectives, the CPXC (X: guest amino

*Correspondence: yangas@gate.sinica.edu.tw
[2] These authors contributed equally to this work.

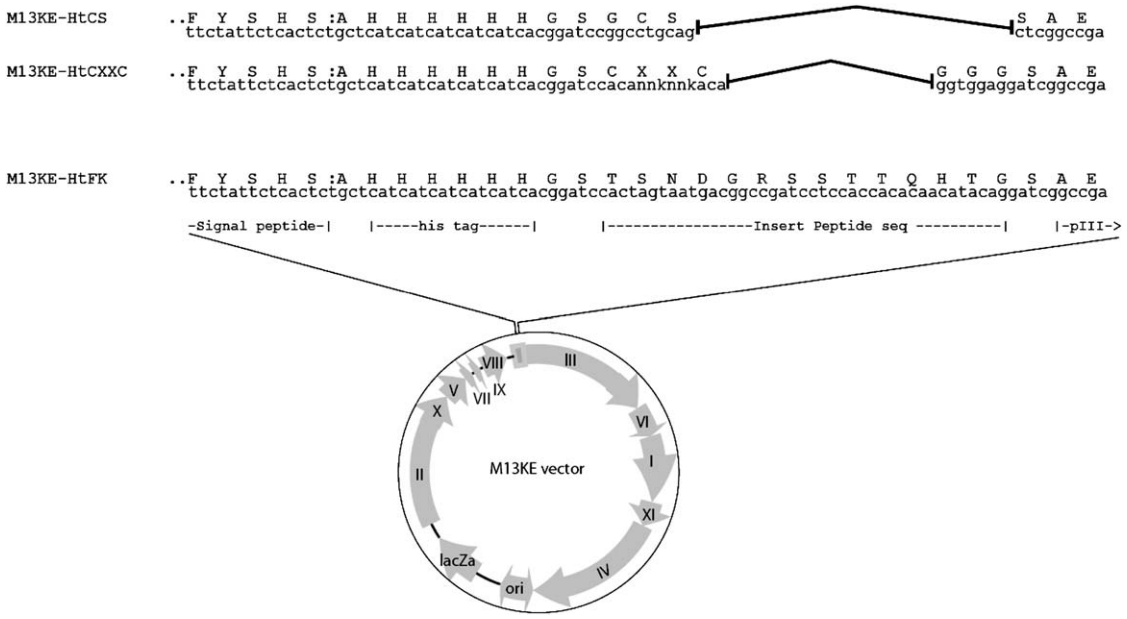

Figure 1. The M13KE Phage Vector and Phage-Displayed Peptide Constructs

The sizes of M13KE-HtCS, M13KE-HtCXXC (X stands for any one of the 20 natural amino acids), and M13KE-HtFK are 7261, 7273, and 7291 bp, respectively.

acid residue) cyclic tetrapeptide system pioneered by Scheraga and coworkers (Falcomer et al., 1992) is particularly attractive in that the tetrapeptides form type I or type II β-turns when the peptide cyclizes through a disulfide bond bridging the first and the fourth Cys—the disulfide-bonded structure is known to be compatible with the left-handed twist of type I and type II β-turn but not compatible with type I′ and II′ β-turns, for which the conformations are twisted in the right-handed direction (Richardson, 1981). The intrinsic free energy for forming the disulfide bond is considered a constant, and the observed equilibrium constant between the oxidized and the reduced species in the presence of oxidizing/reducing agents has been used to determine the free energy of forming a β-turn by the two residues between the two Cys (see Falcomer et al., 1992 and references therein).

Extending the basic idea of the CPXC host-guest system (Falcomer et al., 1992), we made use of filamentous phage display technology (Lowman, 1997; Smith, 1985) to populate the sequence space of the CXXC motif, where X stands for any of the 20 natural amino acids. Two sequence fragments—a hexa-HIS tag N-terminal to the CXXC motif—were inserted side by side between the leader sequence and the pIII gene of M13KE phage vector. After the phage library was constructed, cycles of directed evolution were applied to the phage particle population—evolutionary pressure was applied so that the survivability of the progeny phage particles was dependent on the ability of the CXXC motif to form an intramolecular disulfide bond between the two cysteines. We developed a novel evolutionary procedure designed specifically for selecting peptides with stable intramolecular disulfide bridges. The results that emerged indicated that the sequences forming stable disulfide-bridged CXXC motifs were in good agreement with the sequences with prominent type II β-turn propensities observed in protein structures. In particular, CPGC forming a type II β-turn emerged as the most stable combination of sequence and conformation out of a total of 800 possibilities encoded in the combinatorial phage library. This finding suggested that β-turn formation in proteins follows the active model, as in β sheet and α helix formation.

Results

Feasibility Test of the TCEP/NTCB-Based Directed-Evolution Procedure with the M13KE-HtCS Phage

The TCEP/NTCB-based directed-evolution procedure uses immobilized TCEP [Tris(2-carboxyethyl)phosphine on agarose beads] as the reducing agent to form free thiol groups in the CXXC motif. The reduced thiol groups are susceptible to the peptide bond cleavage reaction at the N terminus of the reduced cysteines with NTCB (2-nitro-5-thiocyanobenzoate). The cleaved species also loses the hexa-HIS tag and hence its ability to bind to Ni-NTA beads, which are used to select for CXXC sequences forming a stable intramolecular disulfide bond.

Before carrying out the TCEP/NTCB-based directed-evolution experiments, we constructed M13KE-HtCS (see Figure 1) phage to demonstrate the feasibility of the directed-evolution procedure. (1) The TCEP/NTCB has no deleterious effect on the vitality of the M13KE phage particles. We compared the titer of the TCEP/NTCB-treated M13KE-HtCS phage particles and the titer of the control M13KE-HtCS phage particles, and no detectable loss of phage titer due to TCEP/NTCB treatment was observed. (2) The displayed hexa-HIS tags bind to Ni-NTA beads, and phage particles without the hexa-HIS tag can be removed from the phage population through washing the Ni-NTA beads with binding

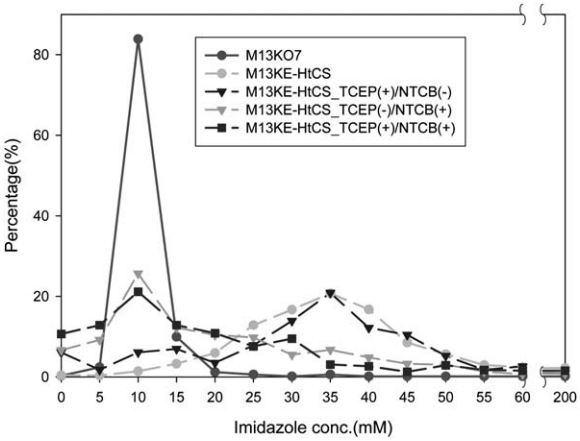

Figure 2. Ni-NTA Agarose Gel Elution Curves of Recombinant Phage Particles M13KE-HtCS and M13K07 against Increasing Concentrations of Imidazole Buffer

The M13KE-HtCS sample was treated with (+) or without (−) TCEP and/or NTCB.

buffer. The M13KE-HtCS phage particles were first mixed with Ni-NTA beads and then eluted with increasing concentrations of imidazole. As shown in Figure 2, the elution curve for M13KE-HtCS peaked at 35 mM imidazole, which is different from the peak at 10 mM imidazole in the elution curve of the M13K07 helper phage. Unlike M13KE-HtCS, M13K07 does not display the hexa-HIS tag. The contrast between the two elution curves indicated that the M13KE-HtCS phage particles bound to Ni-NTA beads through the expressed hexa-HIS tag, and the phage particles with the hexa-HIS tag can be separated from nondisplaying phages through the elution process. (3) NTCB can remove the hexa-HIS tag from the phage particle by cleaving the peptide bond N-terminal to a reduced cysteine—indeed, the elution curve for the NTCB-treated M13KE-HtCS phage particles peaked at 10 mM imidazole, as shown in Figure 2. This elution curve is in good agreement with the elution curve of the M13K07 helper phage, indicating that the cleavage of the hexa-HIS tags had been completed for the majority of the M13KE-HtCS phage particles.

Figure 2 also compares the elution curves of the untreated, NTCB-treated, TCEP-treated, and TCEP/NTCB-treated M13KE-HtCS. As expected, the TCEP-treated curve agreed with the untreated curve, while the TCEP/NTCB-treated curve agreed with the NTCB-treated curve. The results indicated that TCEP treatment alone had no observable effect on the Ni-NTA binding of the phage particles.

Sequence Distribution of the M13KE-HtCXXC Library and the Effect of Propagation on the Sequence Distribution Bias

Figure 1 depicts the DNA construct of the phage-displayed CXXC motif library (M13KE-HtCXXC). By titering the transformants after electroporation of the DNA construct and by sequencing the single colonies from the transformants, we estimated that the library's complexity was 5.3×10^4; i.e., the M13KE-HtCXXC construct

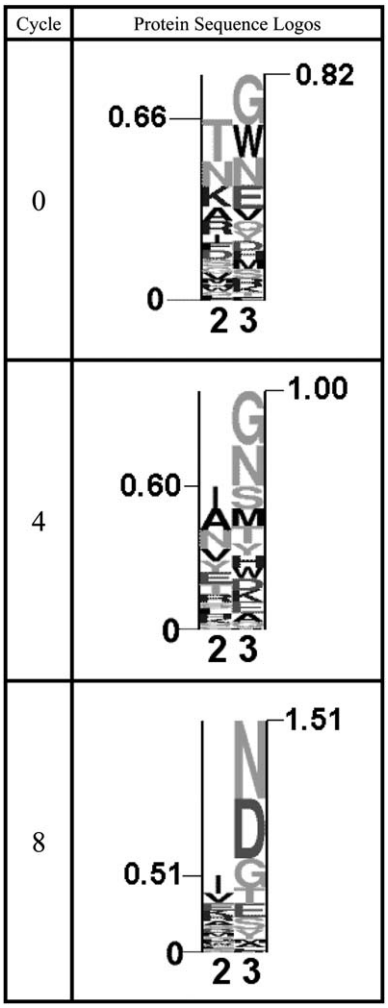

Figure 3. Protein Sequence Logos for the Second and Third Amino Acid Residues of Randomly Sampled Phage Populations from the Zeroth, Fourth, and Eighth Generations of Propagation

The background amino acid probabilities used in the PSL calculation were derived based on the amino acid codon frequencies of the degenerate code NNK.

covered the possible phage species (20 × 20 total) by about 100-fold.

It is well known that a phage library population could be biased due to preferential expression of the phage species after a few rounds of propagation (Rodi et al., 2002; Scholle et al., 2005). We carried out sequence analysis of phage single colonies selected from each generation of propagation, which was amplified by ~1000-fold by infecting the host with the previous generation of phage population of about 10^8 pfu of phage particles. The progeny phage particles were titered, and 70 single colonies from the titer plates were randomly selected for sequencing (more than 50 sequences are considered adequate to produce significant statistics for each generation [Rodi et al., 2002]). The process was repeated for eight rounds. The sequencing results for the zeroth, fourth, and eighth generation are shown in Figure 3.

The protein sequence logo (PSL) histograms (Schneider and Stephens, 1990) in Figure 3, summarizing

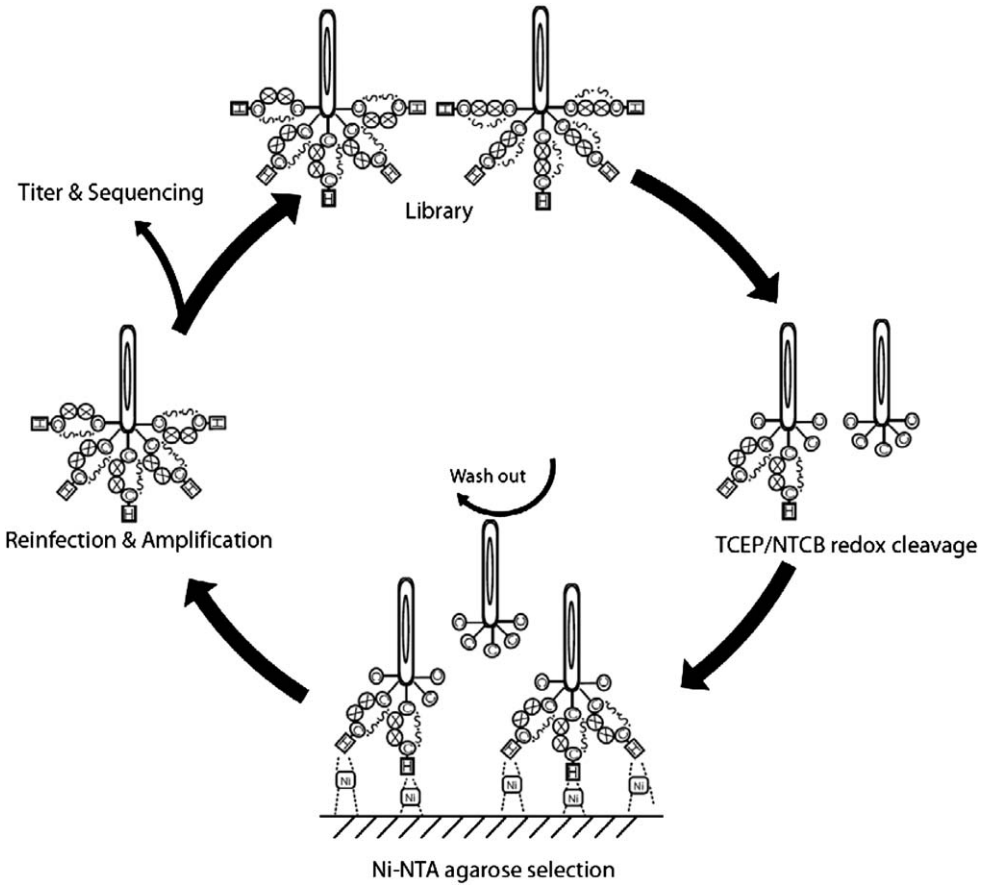

Figure. 4. TCEP/NTCB-Based Directed-Evolution Cycle

The M13KE-HtCXXC library was constructed and treated with TCEP/NTCB to cleave hexa-HIS tag from the phage particles. Some species in the library formed less stable disulfide-bridged tetracyclic CXXC motifs than the others and thus were more susceptible to the cleavage reaction. These species were eliminated from propagating to the next cycle of directed evolution.

the sequence preferences at the second and the third residues of the CXXC motif, suggested that the sequence preferences of the M13KE-HtCXXC library were not seriously biased in the zeroth generation; obvious sequence preference bias appeared only after the fourth generation. Clear sequence preferences emerged after eight rounds of propagation, indicating that the sequence preferences reflected the competitiveness in phage particle morphogenesis and infectivity.

The sequence preferences after the eighth propagation indicated that the second residue in the CXXC motif was comparatively less selective in amino acid type; in contrast, the third residue became increasingly selective toward Asn, Asp, Gly, Glu, Thr, Ser, and Tyr. The trends were in agreement with the type II and, to a lesser extend, type I β-turn sequence preferences derived from protein structures (Hutchinson and Thornton, 1994). The emerging sequence preferences after the eighth generation of propagation agreed with the expectation that sequences forming stable turn structures have stable disulfide bridges and thus are fitter to propagate. Figure 3 also indicated that the TCEP/NTCB-based directed-evolution procedure should not exceed more than four rounds; otherwise, the preferential expression bias would perplex the sequence preferences due to TCEP/NTCB-based directed evolution.

TCEP/NTCB-Based Directed Evolution of the Cyclic CXXC Motif Sequence Preferences

Figure 4 depicts the TCEP/NTCB-based directed-evolution cycle. The phage species emerging from the directed-evolution cycle should be the species for which the CXXC motif is most resistant to the TCEP/NTCB treatment.

Figure 5 compares the imidazole-Ni-NTA elution curves for the M13KE-HtCXXC library from the first and the fourth generation of TCEP/NTCB-based directed evolution. NTCB treatment cleaved the majority of the displayed CXXC motifs in the phage particles from the first generation of TCEP/NTCB-based directed evolution (see the elution curve M13KE-HtCXXC[Evo1]_TCEP[−]/NTCB[+] in Figure 5), indicating that a majority of the population of this phage generation did not form stable enough intramolecular disulfide bonds resistant to the NTCB-mediated peptide cleavage reaction. After four rounds of directed evolution, a stable disulfide-bonded species was evident from the elution curve of the NTCB-treated M13KE-HtCXXC phage particles (see the elution curve M13KE-HtCXXC[Evo4]_TCEP[−]/NTCB[+] in Figure 5). In contrast to the elution curve of the NTCB-treated first-generation phage (which peaked at an imidazole concentration of 10 mM, see the elution curve M13KE-HtCXXC[Evo1]_TCEP[−]/NTCB[+]

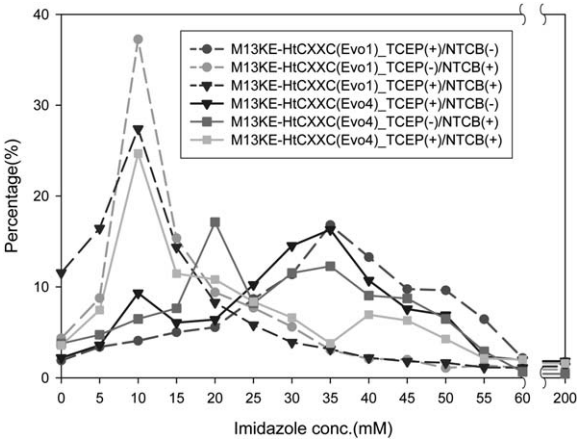

Figure 5. Ni-NTA Agarose Gel Elution Curves of Recombinant Phage Particles Treated with or without TCEP and/or NTCB against Increasing Concentrations of Imidazole Buffer

The figure plots the elution curves for the M13KE-HtCXXC phage library from the first generation (Evo1) and the fourth generation (Evo4) of TCEP/NTCB-based directed evolution.

in Figure 5), the elution curve of the NTCB-treated M13KE-HtCXXC fourth-generation library peaked at an imidazole concentration between 20 and 35 mM (see the elution curve M13KE-HtCXXC[Evo4]_TCEP[−]/NTCB[+] in Figure 5), indicating that the fourth-generation phage particles were protected from NTCB-mediated cleavage reaction by forming an intramolecular disulfide bond between the two cysteines in the CXXC motif. These intramolecular disulfide bonds can be reduced by TCEP-mediated reduction reaction. As shown in Figure 5, the TCEP/NTCB-treated M13KE-HtCXXC phage population had almost identical elution curves (see Figure 5 for TCEP[+]/NTCB[+] elution curves) as that for M13K07 (see Figure 2), suggesting that the disulfide bond in the CXXC motif was susceptible to the reduction reaction mediated by TCEP and that the hexa-HIS tags were removed from the phage particles through the NTCB-mediated cleavage reaction. As expected, species resistant to the TCEP/NTCB treatment (see the peak at 40 mM imidazole of the elution curve M13KE-HtCXXC[Evo4]_TCEP[+]/NTCB[+] in Figure 5) were visibly enriched after four rounds of TCEP/NTCB-based directed evolution.

Table 1 shows the sequencing results for randomly selected phage species from the first three rounds of TCEP/NTCB-based directed evolution. Again, more than 50 sequences from each round of directed evolution were randomly determined for statistical analysis. The most abundant species that emerged from the directed-evolution cycles (see the bold sequences in Table 1) was M13KE-HtCPGC, which appeared 33 times out of a total 190 randomly selected colonies in the first three rounds of TCEP/NTCB-based directed evolution. In contrast, the NNK coding background probability would predict the frequency of appearance for the M13KE-HtCPGC species to be only 4/1000. The dominance of the CPGC population is in good agreement with the prominent propensity for Pro-Gly in the second and the third position of the type II β-turn, as observed both in isolated model peptide systems (Dyson et al., 1988; Falcomer et al., 1992) and in proteins (Hutchinson and Thornton, 1994). Moreover, the third residue position was strongly selective toward Asn, Asp, Gly, Glu, and His, which are also compatible with the middle residues for type II β-turn conformation (Dyson et al., 1988; Falcomer et al., 1992; Hutchinson and Thornton, 1994).

Sequences from the fourth round of the directed evolution were dominated by the M13KE-HtFK species (see Figure 1 for DNA sequence), which did not appear in the first two rounds of directed evolution and only occasionally appeared in the third round of evolution. This species was assumed to be a rare side product of constructing the phage library but finally emerged as the predominant species of the directed evolution. This is

Table 1. Random Samples of Sequence Population from the First through the Third TCEP/NTCB-Based Directed-Evolution Cycles

First Cycle			Second Cycle			Third Cycle		
CAEC	CKLC	CRSC	CADC	CMGC	CRSC	CADC	CPDC	CSDC
CAEC	CLDC	CRSC	CDSC	CMGC	CRSC	CDAC	CPDC	CSHC
CAGC	CLNC	CSKC	CEHC	CMGC	CSKC	CDAC	CPEC	CTSC
CAGC	CLNC	CSNC	CENC	CNNC	CSNC	CENC	**CPGC**	CVDC
CANC	CLSC	CSNC	CENC	CPDC	CSNC	CENC	**CPGC**	CVDC
CANC	CLSC	CSSC	CENC	**CPGC**	CSNC	CESC	**CPGC**	CVDC
CDIC	CMFC	CSTC	CENC	**CPGC**	CSQC	CGHC	**CPGC**	CVEC
CENC	CNNC	CSVC	CESC	**CPGC**	CSYC	CGTC	**CPGC**	CVMC
CENC	CNRC	CTGC	CEYC	**CPGC**	CTQC	CHEC	**CPGC**	CVSC
CGDC	**CPGC**	CTIC	CGTC	**CPGC**	CTSC	CIGC	**CPGC**	CYGC
CHDC	**CPGC**	CTSC	CHEC	**CPGC**	CTSC	CIGC	**CPGC**	
CHEC	**CPGC**	CTSC	CIDC	**CPGC**	CVDC	CINC	**CPGC**	
CHGC	**CPGC**	CTSC	CIGC	**CPGC**	CVDC	CINC	**CPGC**	
CHGC	**CPGC**	CVDC	CIGC	**CPGC**	CVDC	CIRC	**CPGC**	
CHGC	**CPGC**	CVDC	CIGC	CPSC	CVDC	CKGC	**CPGC**	
CIDC	**CPGC**	CVDC	CKSC	CQFC	CVDC	CLEC	**CPGC**	
CIDC	**CPGC**	CVDC	CKSC	CQQC	CVDC	CLGC	**CPGC**	
CIGC	**CPGC**	CVDC	CLEC	CQSC	CVGC	CLNC	CQNC	
CINC	**CPGC**	CVDC	CLGC	CQSC	CVGC	CMEC	CRDC	
CISC	CQEC	CVDC	CLNC	CRHC		CMEC	CRHC	
CKGC	CQSC	CVEC	CLSC	CRNC		CNDC	CRHC	
CKGC	CRNC	CVGC	CLTC	CRNC		CNDC	CRNC	
CKGC	CRNC	CVGC	CMEC	CRNC		CNDC	CRSC	

not surprising, given that M13KE-HtFK does not have any cysteine in the displayed peptide.

Comparison of the Sequence Preferences Derived from the CXXC Host System with Amino Acid Occurrence Frequencies from Known Protein Structures

Figure 6 shows the PSL histograms for the sequence data in Table 1. In contrast to the control PSL histograms shown in Figure 3, the selection power due to TCEP/NTCB-based directed evolution was evident even in the first round of the directed evolution. As shown in Figure 6, the second residue of the CXXC motif was less selective than the third residue. This feature is known to be specific to type II β-turn (see Figure 7), where the second residue is in the more tolerant poly-Pro (P_{II}) conformation area (ϕ_2, $\psi_2 \approx -60°$, 120°), and the third residue in the more restricted left hand 3_{10} conformation area (ϕ_3, $\psi_3 \approx 90°$, 0°). In good agreement with the torsion angle requirements of type II β-turns (see Figure 7), Pro was the most favorable residue at the second position of the CXXC motif after three rounds of directed evolution, while the third position of the CXXC motif was strongly favorable for Asp, Asn, Gly, Glu, and His. We asked how the sequence preferences favorable for the β-turns in the model system compare with the sequence preferences for the β-turns observed in protein structures.

The sequence preferences for the major turn types obtained from protein structures are shown as standard PSL histograms in Figure 7. The β-turns were flanked by two well-formed secondary structure elements, and the connecting coil region between the two secondary structure elements were the second and the third residues of the β-turns (see Experimental Procedures). The subset of β-turns was chosen because two possible mechanisms are likely for turn formation between two secondary structure elements: (1) the β-turn forms first to initiate the folding and packing of the two flanking secondary structure elements; or (2) the secondary structure elements fold and pack first, forcing the turn to form regardless of the sequence preferences of turn formation. If sequence preferences for isolated model β-turns (as shown in Figure 6 and Table 1) would indeed be in agreement with the sequence preferences derived from the subset of structural data (as shown in Figure 7), the significance would be that β-turns must play a more active role in initiating folding, even when β-turn formation was closely connected to the context of the flanking secondary structure elements.

Although the β-turn sequence preferences shown in Figure 7 were derived from a subset of β-turns in the PDB, a comparison of the results shown in Figure 7 with previously published β-turn propensities (Hutchinson and Thornton, 1994) indicated that the two sets of sequence preferences were in general agreement. We calculated the PSSMs (position-specific scoring matrix) for the sequence profiles derived from Hutchinson and Thornton (1994) and from sequence profiles shown in Figure 7 by using Equation 2. The five sets of PSSM elements corresponding to the sequence profiles shown in Figure 7 were respectively plotted (data not shown) against the corresponding PSSM elements calculated based on the statistics from Hutchinson and Thornton (1994). The correlation coefficients are 0.861, 0.667, 0.853, 0.684, and 0.926 for type I, I′, II, II′, and VIII β-turn, respectively, indicating that the PSSMs derived from the two sets of statistics were in good quantitative agreement.

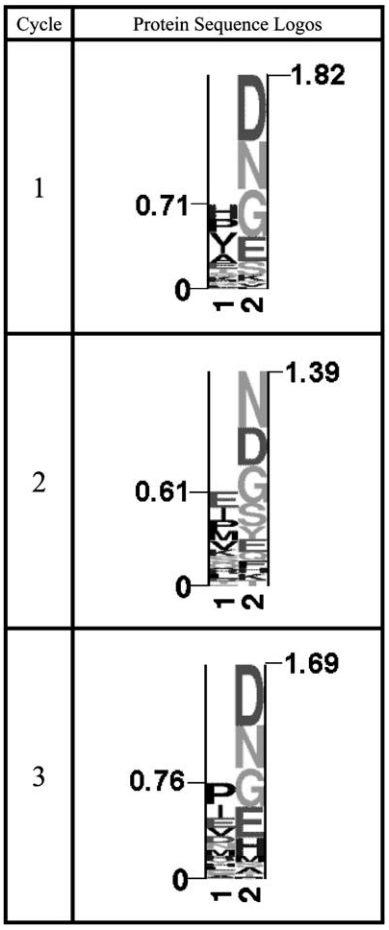

Figure 6. Protein Sequence Logos for the Second and Third Amino Acid Residues of Randomly Sampled Phage Populations from the First, Second, and Third Rounds of TCEP/NTCB-Based Directed Evolution

See Table 1 for detailed sequence information. The background amino acid probabilities used in the PSL calculation were derived based on the amino acid codon frequencies of the degenerate code NNK.

The β-turn sequence preferences from isolated model peptide systems shown in Figure 6 were quantitatively compared with the β-turn sequence preferences derived from structural database analyses. The quantitative comparisons are shown in Figure 8. The y axes of Figure 8 are *Δave_score* (see Equation 3 in the Experimental Procedures) for the randomly sampled sequences (see Table 1 and Figure 6) to score against the PSSMs calculated from the sequence profiles shown in Figure 7 (results shown in Figures 8A and 8B) or to score against the PSSMs derived from the statistics of Hutchinson and Thornton (1994) (results shown in Figures 8C and 8D). The x axes of the figures are the five major groups of β-turns. The first, second, and third generations are shown side by side for comparison in the figure. The increasingly positive *Δave_score* indicates an increasing match between the two sequence profiles (shown in

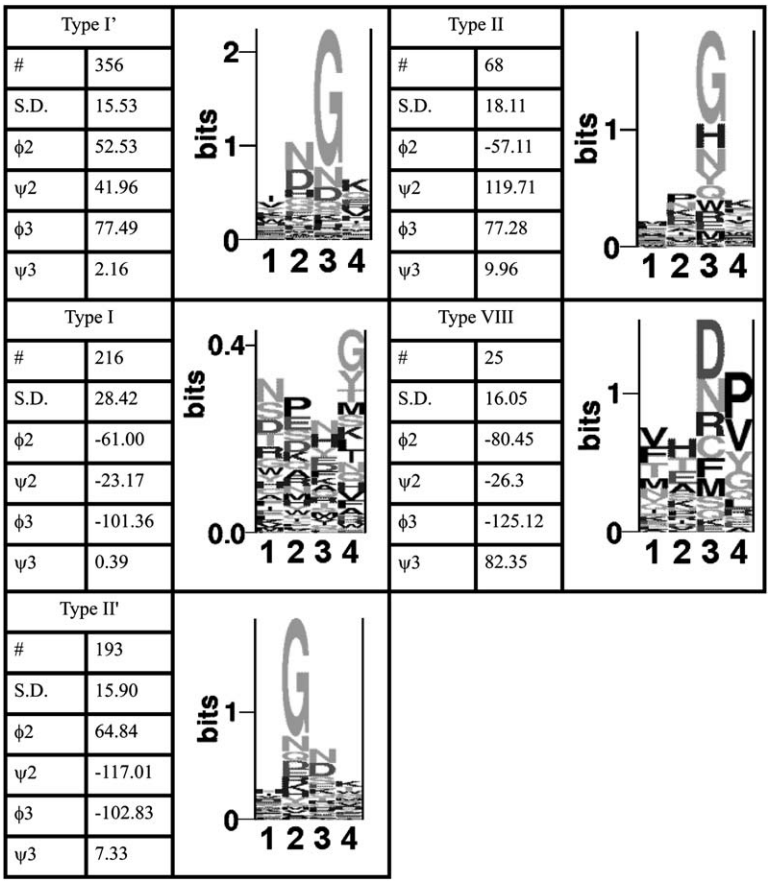

Figure 7. Protein Sequence Logos for β-Turns Flanked by Well-Structured α Helix or β Strand in Protein Structures

The geometrical centers of the β-turn clusters are the average torsion angles of the second and the third residues ($\phi 2$, $\psi 2$, $\phi 3$, $\psi 3$) of the β-turns. The standard deviation (SD) and number (#) of the β-turns in each of the clusters are also shown. The SD is defined by the following equation: $SD = \left(\sqrt{\sum_{i=1}^{n} \left[(\Phi 2_i - \overline{\Phi 2})^2 + (\Psi 2_i - \overline{\Psi 2})^2 + (\Phi 3_i - \overline{\Phi 3})^2 + (\Psi 3_i - \overline{\Psi 3})^2 \right]} \right) / n$, where n is the number (#) of the β-turns in each of the clusters, and $(\overline{\Phi 2}, \overline{\Psi 2}, \overline{\Phi 3}, \overline{\Psi 3})$ is the geometrical center of the β-turn cluster. Default background amino acid probabilities derived from protein structures were used for the PSL calculations.

Figure 6 and Figure 7). A completely random match would produce a zero *Δave_score*; a negative *Δave_score* would indicate a reverse correlation between the two sequence profiles.

The results shown in Figures 8A and 8B indicated that the statistical preferences of type II β-turns from protein structures were strongly correlated with the sequence preferences of the two middle residues in the cyclic CXXC motif. Type I and VIII β-turns also correlated with the statistical preferences, but to a lesser degree. In particular, M13KE-HtCPGC emerged as the predominant species in Table 1, in good agreement with the finding that Pro-Gly is the most favorable sequence for type II β-turns (see Figure 7 and also Hutchinson and Thornton, 1994). The sequences favorable for the CXXC motif are not consistent with the sequence preferences of the type I′ and type II′ β-turns at the second position of the four turn residues (see Figure 8A), in agreement with the expectation that the CXXC motif is compatible with neither of the inverse turns.

Figures 8C and 8D indicated that, again, the sequence preferences for stable CXXC motifs were strongly correlated with the sequence preferences of type II β-turns in protein structures, and to a lesser extent with type I β-turns, but not with type VIII β-turns. Again, the sequences favorable for the CXXC motif are not consistent with the sequence preferences of type I′ and type II′ β-turns at the second position of the four turn residues (see Figure 8C).

The scale of Figures 8A and 8B is smaller than the scale shown in Figures 8C and 8D. This is because the nature of the pseudocount term in the PSSM calculation as shown in Equation 2, where larger statistical sample volume (as in Hutchinson and Thornton, 1994) results in diminishing the effect of the pseudocount term and increasing the absolute magnitude of the PSSM elements. Despite the difference in the magnitude of the two set of figures, the similarity in the trends and patterns of the two sets of figures produced with two sets of PSSMs further supported the notion that the statistics of β-turn propensities from the two surveys (on all β-turns and on a subset of β-turns flanked by secondary structure elements [Hutchinson and Thornton, 1994]) reflected similar β-turn propensities.

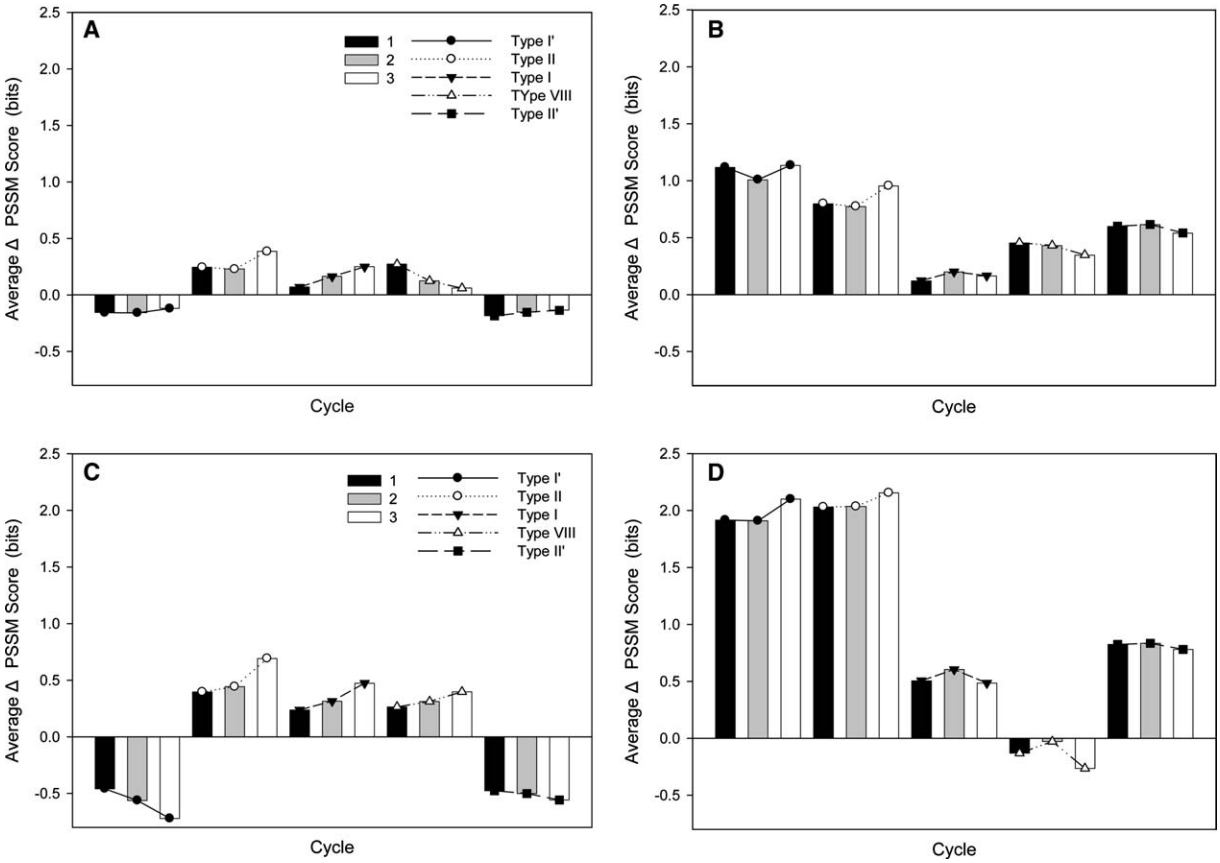

Figure 8. *Δave_score* for the Second and Third β-Turn Residue

(A and B) The *Δave_score* for the second and third β-turn residues, respectively, calculated with the sequences shown in Table 1 against the PSSM (see Equation 2) derived from the sequence profiles shown in Figure 7. Trends of the TCEP/NTCB-based directed evolution for the first through third cycles are shown with histograms in black, gray, and white, respectively. Symbols in the figure keys depict the five types of β-turns as in Figure 7.

(C and D) The *Δave_score* for the second and third β-turn residues, respectively, calculated with the sequences shown in Table 1 against the PSSM (see Equation 2) derived from the statistics of Hutchinson and Thornton (1994).

Discussion

The results that emerged from TCEP/NTCB-based directed-evolution experiments shown in Table 1 and Figure 6 provided sequence preferences for type II and I β-turn formation in a model system. In particular, Pro-Gly forming a type II β-turn (Falcomer et al., 1992) appeared as the most stable conformation among the total of 800 possible combinations. This result is in agreement with the statistical observation that Pro-Gly has the highest occurrence frequency in type II β-turns from protein structures (Hutchinson and Thornton, 1994). The significance of the agreement is that sequence preferences of local protein structures are correlated with conformational stabilities in the absence of context dependence. Even in the situation where the β-turns are flanked by two well-formed secondary structure elements, the sequence preferences of the β-turns were not significantly perturbed by the structural context (see Figures 6–8). Although the model host-guest system of CXXC used in this work forms constrained turn structures in a covalently bonded cyclic tetrapeptide, the results nevertheless suggested that the rigidity of the host system did not significantly compromise the correlation of the intrinsic turn propensities and the statistical surveys from β-turns in proteins—supporting the notion that β-turn residues play an active role (i.e., the active model) in forming the turn structure in proteins.

β-turn-forming propensities of Pro-X amino acid pairs have been compared with host-guest systems. Dyson et al. (1988) measured the chemical shift of the amide-proton resonance of Asp in the pentapeptide sequence Tyr-Pro-X-Asp-Val (X = Gly, Asn, Phe, Ser, or Val) to determine the β-turn population, which in turn determines the β-turn propensities of the amino acid type in the X position of the pentapeptide. The result shows that the pentapeptide with X = Gly has the highest β-turn population, in good agreement with our observation shown in Table 1. Falcomer et al. (1992) measured the thiol/disulfide equilibrium of a series of tetrapeptides: Ac-Cys-Pro-X-Cys-NHMe (X = Gly, Asn, Ser, Phe, Val, or Aib) under oxidation conditions to determine quantitatively the β-turn-forming propensities of the Pro-X pairs. Although Asn and Gly in the X position are the two residues with the highest β-turn-forming potential, Asn has slightly superior β-turn-forming potential compared to Gly: 0.07 kcal/mol in standard conformational free-energy difference (Falcomer et al., 1992). Direct structural

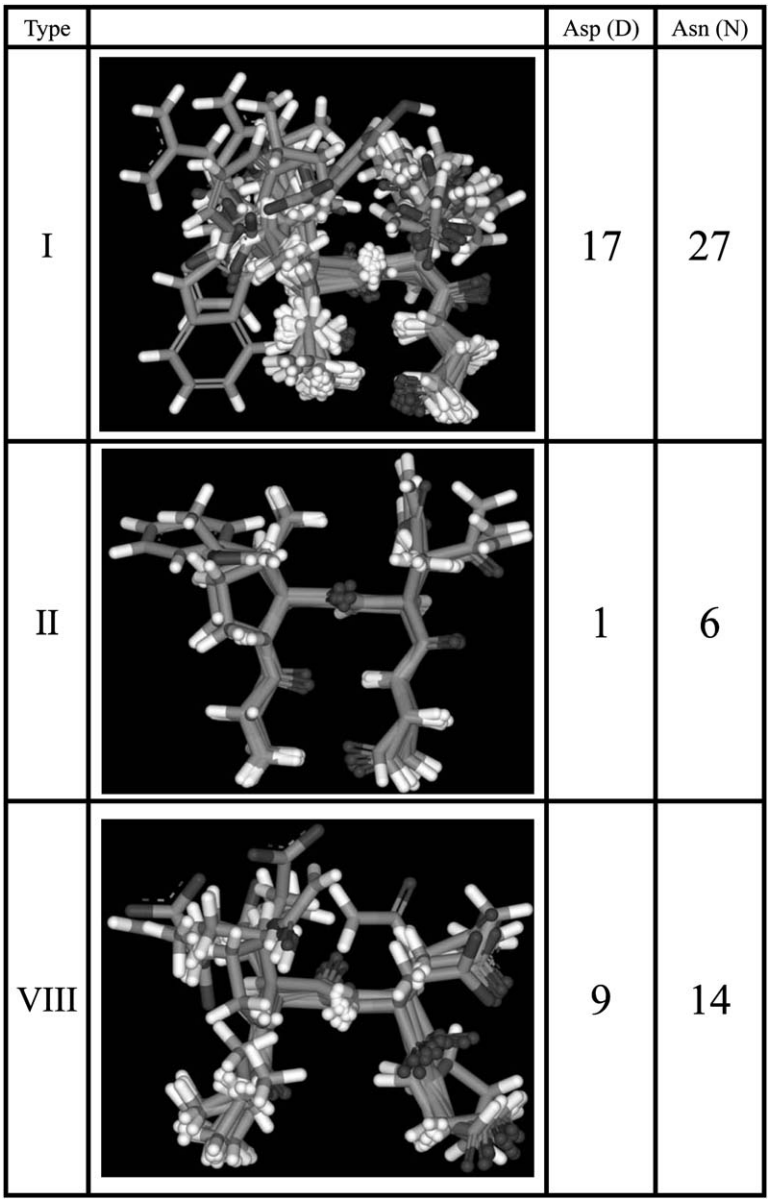

Figure 9. Superimposed Conformations of the Type I, Type II, and Type VIII β-Turns

The conformations were derived as a subset of the clusters shown in Figure 7, where only the β-turns with Asn/Asp at the third position (upper right corner of the β-turns shown in the panels) were shown. The multiple superimpositions were generated with the PrISM program (Yang and Honig, 2000).

evidence based on two-dimensional NMR in aqueous solution indicates that the disulfide-bonded cyclic CPGC forms type II β-turn (Falcomer et al., 1992). Yang and Honig (Yang et al., 1996) used an empirical force-field in conjunction with a finite difference Poisson-Boltzmann (FDPB)-based continuum solvation model to calculate the conformational energy for Ala-Ala, Ala-Gly, Gly-Ala, Gly-Gly, Pro-Ala, and Pro-Gly forming the four major types of β-turn. Among these combinations of turn types and sequences, Pro-Gly forming a type II β-turn is the most stable of all, in agreement with the most predominant species that emerged from TCEP/NTCB-based directed evolution (see Table 1).

Residues with short and polar side chains (in particular, Asp and Asn) at the third residue position are expected to stabilize all types of β-turn (Hutchinson and Thornton, 1994). The physical principles underlying the favorable factors in stabilizing β-turn structures by the short polar residues have not been clearly understood. Figure 9 shows the conformations of type I, II, and VIII β-turns with Asp and Asn at the third position of the β-turns. Close examination of the side-chain conformations of the Asp and Asn residues did not reveal significant preference for a particular side-chain conformation; it is likely that water interactions between the short polar side chain and the peptide backbone help to stabilize the turn conformations (Griffiths-Jones et al., 1999).

Protein engineering for stable protein structures through directed evolution of recombinant phage populated with the biological combinatorial method has been a promising approach (Finucane et al., 1999; Finucane and Woolfson, 1999; Kristensen and Winter, 1998; Sieber et al., 1998). The TCEP/NTCB-based directed-evolution method was developed in this work as a complementary protein engineering platform to evolve stable peptide structures forming intramolecular disulfide bonds. As local structure prediction algorithms are

increasingly accurate in predicting sequence segments in forming protein local structures as protein structure building blocks (Kuang et al., 2004; Yang and Wang, 2002; 2003), experimental confirmation of the local structure propensities in isolated model systems would further clarify the intrinsic structural determinants from the factors that are more relevant to folding interactions involving amino acid residues distant in sequence.

Experimental Procedures

Construction of the M13KE-HtCS Phage Vector

The propagation of *E. coli* ER2738 and M13KE phage was carried out following the instruction manual of the NEB PHD system (New England BioLabs, product #E8120S, version 2.9). M13KE phage was purchased from NEB. Phage was propagated in *E. coli* ER2738, also purchased from NEB. Double-stranded DNA of M13KE phage vector was purified from the infected *E. coli* with NucleoSpin Plasmid QuickPure Kit (BD Bioscience, product #636963) following the manufacturer's manual. Digest 1 μg M13KE vector with Acc651 and EagI (in a 25 μl reaction solution prepared in 1 × NEB #3 buffer, 1 × BSA, 10 units of Acc651, and 10 units of EagI; Acc651 and EagI were purchase from NEB) at 37°C for 4 hr and purify the digested vector with gel extraction kit (Montage, product #LSKGEL050, Millipore).

We constructed M13KE-HtCS vector by inserting a DNA fragment into the restriction-enzyme-digested M13KE vector. Two oligonucleotides were synthesized by Integrated DNA Technologies: HT, 5′-CCCGGGTACCTTTCTATTCTCACTCTGCTCATCATCATCATCAT CACGGATCCTCTTG TTCTTCGGCCGAAACATG-3′ and HTRR, 5′-CATGT TTCGGCCGAAGAACAAGAGGATCCGTGATGATGATGATG ATGAGCAGAG TGAGAATAGAAAGGTACCCG GG-3′. After annealing (annealing conditions: 5 μg of each of the single-stranded DNA in 50 μl of buffer containing 100 mM NaCl, 50 mM HEPES [pH 7.5]; heat to 95°C and cool slowly—decreasing in a step of 0.5°C per 30 s—to room temperature in a thermal cycler), the DNA duplex was digested with Acc65I and EagI: 1 μg of annealed duplex was digested in 25 μl reaction solution containing 1× NEB #3 buffer, 1× BSA, 10 units of Acc651, 10 units of EagI at 37°C for 4 hr. Gel purification of the digested duplex: the digest reaction solution was loaded on a 10% TAE PAGE; band slice was then ground and extracted with TE buffer. The extracted DNA duplex was further purified by phenol/chloroform extraction, chloroform extraction, and ethanol precipitation. The cut duplex was quantitated by PAGE before inserting into the previously digested M13KE vector—in a 20 μl reaction solution containing 25 ng cut M13KE vector, 2.5 ng insert, 1× NEB ligase buffer, and 400 units of T4 ligase from NEB at 16°C overnight.

The ligation solution was electroporated into electrocompetent *E. coli* ER2738 in several separate shocks. Each electroporation was carried out with 80 μl competent cells in a 2 mm cuvette shocked with Ec3 parameter (3 kV, 4.7 ms) in a GenePulser (BioRad). The electrocompetent cells were prepared following the BioRad manual accompanying the instrument. Immediately after the electroporation shock, 1 ml SOC (20 g/l tryptone, 5 g/l yeast extract, 0.5 g/l NaCl, 10 mM $MgCl_2$, 2.5 mM KCl, 20 mM glucose) was added to each cuvette. After 30 min recovery at 37°C, the transfected *E. coli* was plated onto LB/IPTG/X-gal plates for overnight incubation at 37°C. Single blue plaques were picked and cultured.

Phage ssDNA was isolated from each culture for sequencing. The sequencing procedure was carried out following the instruction manual of the NEB PHD system. The only modification of the NEB sequence template purification protocol was to suspend the final ethanol-precipitated ssDNA pellet in pure water instead of TE buffer before sending it out for sequencing. As suggested in the NEB protocol, the −96 primer (5′-CCC TCA TAG TTA GCG TAA CG-3′) was used for automated sequencing. DNA sequencing was carried out in the core laboratory of the Institute of BioMedical Science, Academic Sinica, Taiwan.

Construction of the M13KE-HtCXXC Phage Library

We constructed the M13KE-HtCXXC library by inserting a DNA library into the M13KE-HtCS vector, following the protocol of Noren and Noren (2001) with minor modifications. A single-stranded DNA library oligonucleotide, 5′-GTCATTACTAGTGGATCCTGTNNKNNKT GTGGTG GAGGATCGGCCGGG-3′ (K: G or T, and N: A, C, G, or T), and the extension primer, 5′-TGACCCGGCCGATCCTCC-3′, were synthesized by Integrated DNA Technologies. The library oligonucleotide (4.5 μg) was annealed with the extension primer (4.5 μg) in a total volume of 50 μl reaction containing 100 mM NaCl. The oligonucleotide solution was heated to 95°C and slowly cooled 1°C/minute to 37°C. The annealed duplex was extended with Klenow fill-in reaction: 200 μl reaction solution containing 50 μl annealing mixture of duplex, 1× NEB #2 buffer, 400 uM dNTP, 30 units of Klenow fragment of DNA polymerase I (purchased from NEB) at 37°C for 10 min. The extended product was heat inactivated, restriction digested, purified, and ligated into M13KE-HtCS via BamHI and EagI sites (molar ratio, vector:insert = 1:3). The ligation product was electroporated into *E. coli* ER2738. Except for the ligation sites, the reaction conditions, purification procedures, and electroporation conditions were identical to those described above in the construction of the M13KE-HtCS vector. The transfected cells were cultured overnight for titering. The titer procedure followed the protocol in the instruction manual of the NEB PHD system. Single plaques of the M13KE-HtCXXC library were isolated and sequenced with the same procedures as described in the previous paragraph. Figure 1 depicts the sequence details of the M13KE-HtCS and M13KE-HtCXXC constructs.

Binding of the Phage Particles to Ni-NTA Beads

Transfected *E. coli* ER2738 cells were amplified in LB medium (containing 20 ug/ml tetracycline) with early-log (OD_{600} = 0.3–0.5) *E. coli* ER2738 for 4.5 hr at 37°C with vigorous shaking (200 rpm). The supernatant containing phage particles was separated from the cells by centrifugation at 13,000 × g for 15 min at 4°C. Thirty milliliters of phage solution was precipitated by adding 1/6 volume of PEG (20%)/NaCl (2.5 M) at 4°C overnight. The phage pellet was resuspended in 10 ml Ni-NTA binding buffer (50 mM NaH_2PO_4, 300 mM NaCl [pH 8.0]). We diluted the phage solution to around 10^9 pfu/ml and added 1 ml diluted phage solution to 50 μl of the Ni-NTA gel (purchased from QIAGEN). The mixture was shaken gently at 25°C for 60 min. We loaded the mixture into an Eppendorf filter (UltraFree-MC 5.0 μm, Amicon) and centrifuged at 50 rpm at room temperature for 1 min to filter out the binding buffer. The phage particles were eluted each time from the Ni-NTA beads by mixing the Ni-NTA beads with 500 μl of elution solution with varying concentration of imidazole (from 0 to 200 mM imidazole, 50 mM NaH_2PO_4, 300 mM NaCl [pH 8.0]). Immediately after the mixing, the elution solution was spun out of the Ni-NTA beads and the phage concentrations were titered.

Directed Evolution toward Selecting Stable Disulfide-Bridged Conformations Using TCEP/NTCB-Based Evolutionary Pressure

PEG-precipitated phage particles were suspended in 200 mM Tris-acetate (pH 8) to a final phage concentration of ~10^{11} pfu/ml. 500 μl of the phage solution was added to equal volume of immobilized TCEP (Tris(2-carboxyethyl)phosphine on argarose bead, Pierce product #77712). TCEP gel reduces disulfide bridges exposed to solvent while leaving buried disulfide bridges intact. The mixture was incubated for 60 min in a container filled with nitrogen gas at 37°C with mild shaking and then was centrifuged at 1000 rpm for 1 min. The TCEP gel spin-down was discarded, and NTCB (2-nitro-5-thiocyanobenzoic acid, purchased from Sigma) in 200 mM Tris-acetate (pH 8.0) was added immediately to the supernatant to the final concentration of 4 mM. NTCB cyanylates cysteine thiols. The solution was incubated for 60 min at 40°C in containers filled with nitrogen gas. After the incubation, phage particles were precipitated from the NTCB reaction solution by adding 2/5 volume of PEG (20%)/NaCl (2.5 M) at 4°C for 1 hr. The N-terminal peptide bond of the modified cysteinyl residue can then be cleaved under alkaline conditions. We suspended the phage pellet in 100 μl sodium borate buffer (50 mM [pH 9]) and incubated the solution at 50°C for 60 min. Peptide bound N-terminal to the free thiol group—and thus the hexa-HIS tag—was cleaved from the phage particles (Wu et al., 1996).

For directed evolution, the TCEP/NTCB-treated phage library particles were added to Ni-NTA beads for binding as described in the previous section. The flow through and the 3 × 500 μl of washing

buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole [pH 8.0]) were discarded. 500 μl of elution buffer (500 mM imidazole, 50 mM NaH_2PO_4, 300 mM NaCl [pH 8.0]) was added to the Ni-NTA beads, and the eluent was collected for amplification (see above). The amplified phage solution was treated with TCEP/NTCB as described above to start the next cycle of directed evolution.

Structure/Sequence Survey and Structural Clustering of β-Turns

β-turn structures/sequences were extracted from nonredundant protein structures (PDB_SELECT version 2003 April, 1999 chains) in the PDB (Protein Data Bank). Since β-turn analysis in proteins has been well established (Hutchinson and Thornton, 1994), we focused on a subset of β-turns flanked by two secondary structure elements (β strand or α helix). These β-turns have four residues; the first and the fourth residues were at the termini of either β strand or α helix, and the distance between the first and the fourth Cα atom was less than 6 Å. The turn, β strand, and α helix assignment for each residue in the protein structures was derived based on the DSSP algorithm (Kabsch and Sander, 1983). The search yielded 914 local structures with the PrISM program (Kuang et al., 2004; Yang, 2002; Yang and Wang, 2002; 2003).

Following the conventional definition of β-turns on the basis of the backbone torsion angles, we defined the structural distance function *d(x,y)* based on the two sets of phi(ϕ)-psi(ψ) angles of the turn residues (the second and the third residue) of structure *x* and *y*:

$$d(x,y) = \sqrt{\frac{\sum_{i=1}^{L}(\Phi xi - \Phi yi)^2 + \sum_{i=1}^{L}(\Psi xi - \Psi yi)^2}{2L}}, \quad (1)$$

where *L* equals 2 in β-turn structures. The phi(ϕ)-psi(ψ) angles were calculated with the DSSP program (Kabsch and Sander, 1983).

The 914 β-turn structures were clustered according to their structural similarities by an automated procedure. We used Equation 1 to calculate the all-against-all (914 × 914) structural distance (*d*[*x,y*]) matrix. The UPGMA (unweighted pair group method with arithmetic mean) algorithm (Sokal and Michener, 1958) was then used to merge the structural distance matrix elements into distinguished clusters by increasing stepwise the partition cutoff value. The merging cycle was terminated as the increases of the partition cutoff value began to result in only incremental decreases in cluster numbers, and the clusters geometric centers and distributions (see Figure 7) follow the conventional definitions for various β-turn types as shown in Hutchinson and Thornton (1994). The algorithm yielded 17 clusters. We selected the largest five clusters, which contained 858 β-turns of the total 914 β-turns (94%). The clusters and the geometrically averaged centers of the clusters are shown in Figure 7.

Local Structure-Based Sequence Profile and PSSM Calculation

For a cluster of structurally aligned β-turn sequences, the PSSM (position specific scoring matrix) (Altschul et al., 1997) of the turn residues can be calculated based on the amino acid sequence profile of the local structures (Yang and Wang, 2002; 2003). Equation 2 (Altschul et al., 1997; Tatusov et al., 1994; Yang and Wang, 2003) shows the derivation of the elements in a PSSM:

$$W(J)_i = \log_2\left[\frac{1}{P_i} \times \left(\frac{n(J)_i + \sqrt{N(J)} \times P_i}{N(J) + \sqrt{N(J)}}\right)\right]. \quad (2)$$

$W(J)_i$ is the position-specific scoring matrix element in bit unit for amino acid type *i* at position *J* of the sequence profile. *N(J)* is the total number of sequences in the sequence profile at position *J*. $n(J)_i$ is the count of the amino acid type *i* at position *J* of the sequence profile. P_i is the background probability for amino acid type *i* in protein structures. $\sqrt{N(J)} \times P_i$ is the pseudocount; the purpose of including this term in the equation is to avoid a logarithmic error when $n(J)_i$ equals to zero (Tatusov et al., 1994; Yang and Wang, 2003). The β-turn survey shown in Table 1 of Hutchinson and Thornton (1994) has provided all the statistics needed in Equation 2, enabling us to calculate the PSSM for all β-turn types surveyed in Hutchinson and Thornton. We also derived PSSM for the sequence profiles from the five clusters derived in the previous section with the same formulation.

Comparison of the Sequence Preferences that Emerged with Amino Acid Occurrence Frequencies Derived from Protein Data Bank Survey

To evaluate the sequence match between a set of multiple-aligned sequences and a PSSM at specific positions *J*, we calculated the average score of the query set of sequences against the PSSM by using the mean score of random sequences against the same PSSM as baseline; i.e.,

$$\Delta ave\_score(J) = \frac{\sum_{i=1}^{20} W(J)_i n(J)_i}{\sum_{i=1}^{20} n(J)_i} - \frac{\sum_{i=1}^{20} W(J)_i nr_i}{32}. \quad (3)$$

$W(J)_i$ and $n(J)_i$ have been defined in Equation 2. nr_i is the number of redundancy of the codon encoding amino acid type *i* from the degenerate codon NNK (N: A, T, C, or G; K: G or T), which encodes 20 natural amino acids and an amber stop codon (TAG) in 32 combinations. If the query sequences were randomly selected from the pool of the phage library without any biases, the sequence population should reflect the number of codons used in encoding each amino acid type (Rodi et al., 2002); that is, the first term in the right hand side of Equation 3 should approach the second term of the equation, and *Δave_score(J)* should approach zero. The more positive the *Δave_score(J)* is, the closer match the set of query sequences is to the PSSM. Inversely, a negative *Δave_score(J)* indicates reverse correlation of the set of sequences to the PSSM.

Acknowledgments

This work was supported by funds from Genomics Research Center at Academia Sinica and by National Health Research Institute Innovative Research Grant NHRI-EX95-9525EI.

Received: June 16, 2006
Revised: July 28, 2006
Accepted: August 1, 2006
Published: October 10, 2006

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bonvin, A.M., and van Gunsteren, W.F. (2000). beta-hairpin stability and folding: molecular dynamics studies of the first beta-hairpin of tendamistat. J. Mol. Biol. *296*, 255–268.

Castagnoli, L., Vetriani, C., and Cesareni, G. (1994). Linking an easily detectable phenotype to the folding of a common structural motif. Selection of rare turn mutations that prevent the folding of Rop. J. Mol. Biol. *237*, 378–387.

Chakrabartty, A., and Baldwin, R.L. (1995). Stability of alpha-helices. Adv. Protein Chem. *46*, 141–176.

Dinner, A.R., Lazaridis, T., and Karplus, M. (1999). Understanding beta-hairpin formation. Proc. Natl. Acad. Sci. USA *96*, 9068–9073.

Dyson, H.J., Rance, M., Houghten, R.A., Lerner, R.A., and Wright, P.E. (1988). Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. J. Mol. Biol. *201*, 161–200.

Falcomer, C., Meinwald, Y., Choudhary, I., Talluri, S., Milburn, P., Clady, J., and Scheraga, H. (1992). Chain reversals in model peptides: studies of cystine-containing cyclic peptide. 3. Conformational free energies of cyclization of tetrapeptides of sequence Ac-Cys-Pro-X-Cys-NHMe. J. Am. Chem. Soc. *114*, 4036–4042.

Finucane, M.D., and Woolfson, D.N. (1999). Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin. Biochemistry *38*, 11613–11623.

Finucane, M.D., Tuna, M., Lees, J.H., and Woolfson, D.N. (1999). Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. Biochemistry *38*, 11604–11612.

Fitzkee, N.C., Fleming, P.J., and Rose, G.D. (2005). The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. Proteins *58*, 852–854.

Griffiths-Jones, S.R., Maynard, A.J., and Searle, M.S. (1999). Dissecting the stability of a beta-hairpin peptide that folds in water: NMR and molecular dynamics analysis of the beta-turn and beta-strand contributions to folding. J. Mol. Biol. *292*, 1051–1069.

Hutchinson, E.G., and Thornton, J.M. (1994). A revised set of potentials for beta-turn formation in proteins. Protein Sci. *3*, 2207–2216.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kristensen, P., and Winter, G. (1998). Proteolytic selection for protein folding using filamentous bacteriophages. Fold. Des. *3*, 321–328.

Kuang, R., Leslie, C.S., and Yang, A.S. (2004). Protein backbone angle prediction with machine learning approaches. Bioinformatics *20*, 1612–1621. Published online February 26, 2004. 10.1093/bioinformatics/bth136.

Lowman, H.B. (1997). Bacteriophage display and discovery of peptide leads for drug development. Annu. Rev. Biophys. Biomol. Struct. *26*, 401–424.

Minor, D.L., Jr., and Kim, P.S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. Nature *367*, 660–663.

Munoz, V., Thompson, P.A., Hofrichter, J., and Eaton, W.A. (1997). Folding dynamics and mechanism of beta-hairpin formation. Nature *390*, 196–199.

Noren, K.A., and Noren, C.J. (2001). Construction of high-complexity combinatorial phage display peptide libraries. Methods *23*, 169–178.

Panasik, N., Jr., Fleming, P.J., and Rose, G.D. (2005). Hydrogen-bonded turns in proteins: the case for a recount. Protein Sci. *14*, 2910–2914.

Pande, V.S., and Rokhsar, D.S. (1999). Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. Proc. Natl. Acad. Sci. USA *96*, 9062–9067.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. Adv. Protein Chem. *34*, 167–339.

Rodi, D.J., Soares, A.S., and Makowski, L. (2002). Quantitative assessment of peptide sequence diversity in M13 combinatorial peptide phage display libraries. J. Mol. Biol. *322*, 1039–1052.

Rose, G.D., Gierasch, L.M., and Smith, J.A. (1985). Turns in peptides and proteins. Adv. Protein Chem. *37*, 1–109.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Scholle, M.D., Kehoe, J.W., and Kay, B.K. (2005). Efficient construction of a large collection of phage-displayed combinatorial peptide libraries. Comb. Chem. High Throughput Screen. *8*, 545–551.

Searle, M.S. (2004). Insights into stabilizing weak interactions in designed peptide beta-hairpins. Biopolymers *76*, 185–195.

Searle, M.S., and Ciani, B. (2004). Design of beta-sheet systems for understanding the thermodynamics and kinetics of protein folding. Curr. Opin. Struct. Biol. *14*, 458–464.

Sieber, V., Pluckthun, A., and Schmid, F.X. (1998). Selecting proteins with improved stability by a phage-based method. Nat. Biotechnol. *16*, 955–960.

Simpson, E.R., Meldrum, J.K., Bofill, R., Crespo, M.D., Holmes, E., and Searle, M.S. (2005). Engineering enhanced protein stability through beta-turn optimization: insights for the design of stable peptide beta-hairpin systems. Angew. Chem. Int. Ed. Engl. *44*, 4939–4944.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315–1317.

Sokal, R.R., and Michener, C.D. (1958). A statistical method for evaluation systematic relationships. Univ. Kans. Sci. Bull. *28*, 1409–1438.

Tatusov, R.L., Altschul, S.F., and Koonin, E.V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA *91*, 12091–12095.

Wu, J., Gage, D.A., and Watson, J.T. (1996). A strategy to locate cysteine residues in proteins by specific chemical cleavage followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Anal. Biochem. *235*, 161–174.

Yang, A.S. (2002). Structure-dependent sequence alignment for remotely related proteins. Bioinformatics *18*, 1658–1665.

Yang, A.S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. *301*, 691–712.

Yang, A.S., and Wang, L. (2002). Local structure-based sequence profile database for local and global protein structure predictions. Bioinformatics *18*, 1650–1657.

Yang, A.S., and Wang, L. (2003). Local structure prediction with local structure-based sequence profiles. Bioinformatics *19*, 1267–1274.

Yang, A.S., Hitz, B., and Honig, B. (1996). Free energy determinants of secondary structure formation: III. beta-turns and their role in protein folding. J. Mol. Biol. *259*, 873–882.