DESIGN OF HYDROPHOBIC CORE OF *E.coli* MALATE DEHYDROGENASE BASED ON THE SIDE-CHAIN PACKING

H. KONO
Tsukuba Life Science Center, The Institute of Physical & Chemical Research (RIKEN), Tsukuba, Ibaraki 305, Japan

M. NISHIYAMA, M. TANOKURA
Biotechnology Research Center, The University of Tokyo, Tokyo 113, Japan

J. DOI
Bioinformation Engineering Laboratory Department of Biotechnology, The University of Tokyo 1-1-1, Yayoi, Bunkyo-ku, Tokyo 113, JAPAN

We have developed computational programs for the de novo design of hydrophobic cores of proteins. The first program optimizes side-chain conformations using an updated rotamer library for potential hydrophobic residues, based on the backbone structure of the protein of interest. The second program selects candidates to be engineered among the sequences by estimating changes in Gibbs free energy between the folded and unfolded structure of the proteins with new sequence. Using these programs, we constructed several variants of *E. coli* malate dehydrogenase (eMDH) which could have increased stability at 25°C, compared to the wild type enzyme. To quantitate stability change between variants and the wild type, circular dichroism spectra were measured as a function of guanidine hydrochloride concentration at 25°C, pH 7.0. This analysis showed that three variants constructed in this study were stabilized more than or equal to the wild type. This demonstrated that our programs may be powerful tools to design new proteins with high stability.

1 Introduction

The de novo design of protein is one of the ultimate purposes in protein engineering. Many efforts have been made in the creation of a new amino acid sequence that folds into a predetermined three dimensional topology [1,2,3]. Empirical knowledge obtained so far has been tested in the design of a sequence from a given structure.

Protein folding problem is not only essential for understanding of the nature of protein structure, but also has immediate importance to many aspects of biology and biotechnology. There has been increasing interest in design of more stable variants of existing proteins. Godzik [4] tried to find out ideal sequences for a plastcyanin from 20^{99} sequence space using a threading algorithm. However, the obtained sequences showed a poor similarity to the native

sequence. Also, using a genetic algorithm as well as the threading potential, Jones *et al*[5] tried to design a new α-helical protein from the sequence giving all β-structure. The designed protein, however, exhibited little helical property in water containing ethylene glycol or methanol. Thus, no success have been achieved in design of a whole protein, although the threading method has been claimed to be potentially capable of sequence design.

In addition, Hellinga and Richards [6] designed a hydrophobic core of λ repressor using a simulated evolution algorithm. They succeeded in obtaining the sequences with biological activity in the higher rank of the calculation. But, the relationship of the energetics they used and biological activity of these sequences was not clear. Desjarlais and Handel [7] also successfully engineered a hydrophobic core of phage 434 cro protein. The sequences for the core were designed to achieve only a tight packing using a genetic algorithm. Although the order of protein stability obtained by their calculation did not always agree with that by the experiments, the results showed that tight packing is one of the important factors in de novo design of stable protein. Based on the side chain packing, Lee & Levitt [8] and Lee [9] predicted the energetics of protein thermostability and Wilson *et al*[10] carried out the energetic analysis of substrate binding. The packing contribution to the protein stability was well reviewed by Lim & Richards [11].

In this paper, we describe two computational programs, which produce protein sequences with increased thermostability by optimizing side-chain packing of hydrophobic core. And the calculated thermostability using these programs was compared with that of biologically engineered proteins.

2 Materials and Methods

2.1 Design of sequences

Two programs were used in the prediction of amino acid sequence for a core and the first one was based on the automata network method [12] as previously described. Using this program, a set of amino acid sequences and their side chain conformations which could achieve a tight side chain packing were predicted by random jumps in the sequence and conformational space of hydrophobic amino acids such as Ala, Val, Leu, Ile, Met and Phe. As for side-chain conformation of each residue site, an updated rotamer library [13] was used together with the rotamer derived from the crystal structure of *E. coli* malate dehydrogenase. Consequently, 123 sets of sequences were obtained after 1,000 runs at various initial states of the network.

Then the second program was used to estimate Gibbs free energy changes

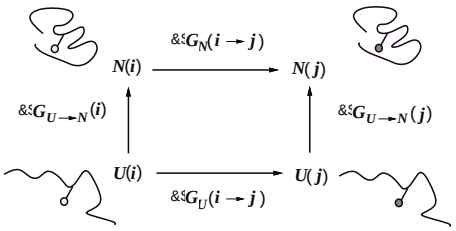

Figure 1: The thermodynamic cycle between wild type (i) and its variant (j).

of the sequence produced by the first program. In estimation of Gibbs free energy changes, hydration, side-chain entropy, bond energy and non-bond energy were considered. The stability of a sequence i is given by changes in free energy ΔG^{H_2O} between native state, N, and unfold state, U,

$$\Delta G_{U \Rightarrow N}(i) = \Delta E^{solvation}_{U \Rightarrow N}(i) + \Delta E^{entropy}_{U \Rightarrow N}(i) + \Delta E^{nb}_{U \Rightarrow N}(i) + \Delta E^{bond}_{U \Rightarrow N}(i) \quad (1)$$

where $\Delta E^{solvation}_{U \Rightarrow N}(i), \Delta E^{entropy}_{U \Rightarrow N}(i), \Delta E^{nb}_{U \Rightarrow N}(i), \Delta E^{bond}_{U \Rightarrow N}(i)$ are terms due to changes in the solvation, entropy, non-bonded interactions and covalently bonded structure between the unfolded and native state of the protein, respectively. The energy difference between two sequences i and j is given by the thermodynamic cycle in Figure 1. This thermodynamic linkage allows the difference in the stabilities of two sequences to be described by two equivalent forms: the natural difference of the stabilities

$$\Delta\Delta G_{U \Rightarrow N}(i \Rightarrow j) = \Delta E_{U \Rightarrow N}(j) - \Delta E_{U \Rightarrow N}(i) \quad (2)$$

and a second form that does not reflect chemical reality but is accessible by considering the cycle,

$$\Delta\Delta G_{U \Rightarrow N}(i \Rightarrow j) = \Delta E_N(i \Rightarrow j) - \Delta E_U(i \Rightarrow j). \quad (3)$$

The cycle can be written as the sum of the differences of each component

$$\begin{aligned} \Delta\Delta G_{U \Rightarrow N}(i \Rightarrow j) \quad = \quad & \Delta\Delta E^{solvation}_{U \Rightarrow N}(i \Rightarrow j) + \Delta\Delta E^{entropy}_{U \Rightarrow N}(i \Rightarrow j) \\ & + \Delta\Delta E^{nb}_{U \Rightarrow N}(i \Rightarrow j) + \Delta\Delta E^{bond}_{U \Rightarrow N}(i \Rightarrow j). \end{aligned} \quad (4)$$

Each of these terms is subject to the thermodynamic linkage relationship and can therefore be calculated by equations analogous to either Eqs. 2 or 3.

The contribution of the solvation term to the folding energy of a sequence i with n residues long, can be given as follows,

$$\Delta E^{solvation}_{U \Rightarrow N}(i) = \sum_{k=0}^{n} g_k \frac{A^0_k(i) - A_k(i)}{A^0_k(i)} \quad (5)$$

where g_i is the free energy of transfer between gas and water [14] and A^0 and A are solvent accessible surface areas in the unfolded state and in the folded state, respectively. The free energy of solvation, to be added to the AMBER energy, must correspond only to the additional interactions of the atoms of the solute with water. Consequently, we used the parameters derived from observed free energies of transfer from gas to aqueous solution. The area was calculated using the program developed based on the MSEED algorithm [15]. Since the unfolded protein is assumed to be in an extended form, the contribution of the solvation term to the energy differences of the sequences becomes

$$\Delta\Delta E_{U\Rightarrow N}^{solvation}(i \Rightarrow j) = \sum_{k=0}^{n} g_k \frac{A_k^0(j) - A_k(j)}{A_k^0(j)} - \sum_{k=0}^{n} g_k \frac{A_k^0(i) - A_k(i)}{A_k^0(i)}. \quad (6)$$

The loss of entropy on protein folding causes an important contribution to the stability of a protein. Since neither glycine nor proline was being considered in this study, the main chain entropy was assumed to be unaffected by the sequence changes. Loss of side-chain entropy on the folding is reflected by the degrees of rotational freedom in the native and unfolded state of all side chain. Because all side chains were fully buried in this simulation, the side-chain entropies in the native were assumed to be 0. The entropies in the unfolded were taken from the literature [16]. Then, the change in entropy is

$$\Delta\Delta E_{U\Rightarrow N}^{entropy}(i \Rightarrow j) = \sum_{k=0}^{n} (S_k(j) - S_k(i)). \quad (7)$$

As for non-bonded interaction, only the van der Waals interactions were considered. The coulombic interaction term was ignored, due to the fact that only uncharged residues were used in this calculation. The differences in stabilities of the sequences due to the non-bonded interactions were determined using

$$\Delta\Delta E_{U\Rightarrow N}^{nb}(i \Rightarrow j) = \Delta E_N^{vdw}(i \Rightarrow j) - \Delta E_U^{vdw}(i \Rightarrow j). \quad (8)$$

$\Delta E_N^{vdw}(i \Rightarrow j)$ is obtained directly from the calculation by evaluating a 6-12 potential function with parameters obtained from the AMBER 3A force field [17]. $\Delta E_U^{vdw}(i \Rightarrow j)$ could be ignored due to the same reason described for the solvation term. Since the energy of covalent bonding should be unchanged in respect of the the folding state, $\Delta\Delta E_{U\Rightarrow N}^{bond}(i \Rightarrow j) = 0$.

2.2 Selection of hydrophobic core

E.coli malate dehydrogenase (eMDH) consists of 312 amino acids with nine helices and eleven strands, which form into a homo-dimer. Its structure has

Figure 2: eMDH structure. Residues composing the core we focused in this study are denoted by thick line. This drawing was produced by MOLSCRIPT[20].

been determined by X-ray at 1.9Å with an R-value 0.195 [18]. Using a computer graphics system [19], we searched hydrophobic core of the enzyme by manually clipping the molecular and found four by monomer. We selected one out of the four (shown in Figure2) as the target in this study because it corresponds to the core of *Thermus flavus* malate dehydrogenase, a site experimentally shown to be one of determinants of the high thermal stability (unpublished data). The core consists of 10 hydrophobic residues at 158, 159, 169, 171, 186, 189, 192, 194, 199 and 202.

2.3 Mutagenesis

We conducted mutagenesis with a PCR method [21]. The nucleotide sequences of mutation points were confirmed by the M13 dideoxy chain termination method [22,23] with a DSQ-1000 (Shimadzu Co., Kyoto) DNA sequencer. The mutated gene was expressed in *E.coli* 5131-5 [24] using pUC18 as the expression plasmid. The cells were disrupted by sonication and eMDH variants were purified by Blue Sepharose CL-6B affinity column chromatography.

2.4 Circular dichroism

All circular dichroism (CD) spectra were measured with JASCO J-720 spectropolarimeters between 200 and 250 nm using proteins of about 100 μg/ml in 20 mM phosphate buffer (pH 7.0) at 25°C.

The changes in the CD spectra at θ_{222} nm of the proteins were compared by normalizing each transition curve to the apparent fraction of the unfolded form. Thermodynamic parameters, ΔG^{H_2O} and m, were extracted from the denatured curves by assuming that the unfolding reaction follows a two-state model $D \rightleftharpoons 2U$ [25]. ΔG is the change in free energy between D and U at a given GuHCl concentration, ΔG^{H_2O} is the value in the absence of denaturant and m is a fitting parameter that reflects the cooperativity of the unfolding transition.

2.5 Assay for the enzyme activity

Steady-state kinetic analyses were carried out by measuring the decrease in absorbance at 340 nm in reaction mixtures which contains 4 nM of the purified enzymes and 167 μM of oxaloacetate in 33 mM potassium phosphate buffer (pH 7.0) at 30°C as described previously [26].

3 Results

3.1 Designed sequences and their stability

Our programs produced eMDH variants with new amino acid sequences giving various free energy. The highest score of the free energy (14.17 kcal/mol) was obtained in can35 which had 5 amino acid replacements of Val158Leu, Val169Leu, Val189Ala, Phe194Leu and Val199Ile. This and other seven variants with various level of free energy (stability) were chosen for site-directed mutagenesis as described above. The amino acid sequences and their energy components are shown in Table 1 and 2, respectively. It can be seen from these results that the side-chain conformations of the predicted variant structures is very similar to those of the crystal structure and exhibit more tight packing (Figure 3). Except two variants, V169L/V199I and P_{ref}, other variants appeared to be more stable than the wild type at 25°C. P_{ref} was designed as a reference where 4 valine residues were simultaneously replaced with larger residues, Leu or Ile. For P_{ref}, unfavorable van der Waals contacts should occur because there was no room to contain four larger residues in the backbone structure which was fixed to that of the wild type. As for V169L/V199I, since the sum of the solvation term and the entropic term was a little larger than

Table 1: Designed sequences for eMDH type core region

variant	158	159	169	171	186	189	192	194	199	202
wild	V	A	V	V	L	V	V	F	V	L
V199I	V	A	V	V	L	V	V	F	i	L
V169L/V199M	V	A	*l*	V	L	V	V	F	*m*	L
V169L/V199I	V	A	*l*	V	L	V	V	F	*i*	L
V169L	V	A	*l*	V	L	V	V	F	V	L
can25	*l*	A	*m*	V	L	*a*	V	*l*	*m*	L
can26	*l*	A	*l*	*l*	L	*a*	V	*l*	*i*	L
V199M	V	A	V	V	L	V	V	F	*m*	L
can35	*l*	A	*l*	V	L	*a*	V	*l*	*i*	L
P_{ref}	*l*	A	*l*	*l*	L	V	V	F	*i*	L

Amino acids shown with lowercase are different from wild one.

Table 2: Predicted changes in free energy

protein	$\Delta\Delta H$ (kcal/mol)	$-T\Delta\Delta S$	$\Delta\Delta g$	$\Delta\Delta G^{cal}$
V199I	+1.61	−0.68	−0.22	+0.71
V169L/V199M	+6.18	−2.07	−3.01	+1.10
V169L/V199I	+2.27	−1.36	−1.37	−0.46
V169L	+0.66	−0.68	+0.15	+0.13
can25	+7.70	−3.21	−2.65	+1.85
can26	+11.25	−2.47	+0.53	+9.31
V199M	+7.24	−1.39	−1.99	+3.86
can35	+15.67	−1.79	+0.29	+14.17
P_{ref}	−101.48	−2.72	+0.40	−103.80

T = 298 K. $\Delta\Delta$g values from Ooi *et al.* $\Delta\Delta G^{cal}$ denotes a calculated $\Delta\Delta G$.

the enthalpic term, the stability was predicted to decrease. In general, replacement of a hydrophobic amino acid with the larger one increases the stability. However, it was not always the case in this study. In the variant V169L/V199I, the introduction of Ile at position 199 protected the surrounding residues from exposure to the solvent and thus caused a decrease in $\Delta\Delta g$.

3.2 Native structure for variants

We constructed the mutated eMDH genes for the expression of the nine variants described above and six variants of nine designed proteins were obtained. The other three mutants with replacement of Phe by Leu at position 194 were not produced with unknown reason. We assume that Phe 194 is an important residue for eMDH for correct folding to a native form. CD spectra of variants proved that all the variants had the same structure as the wild type

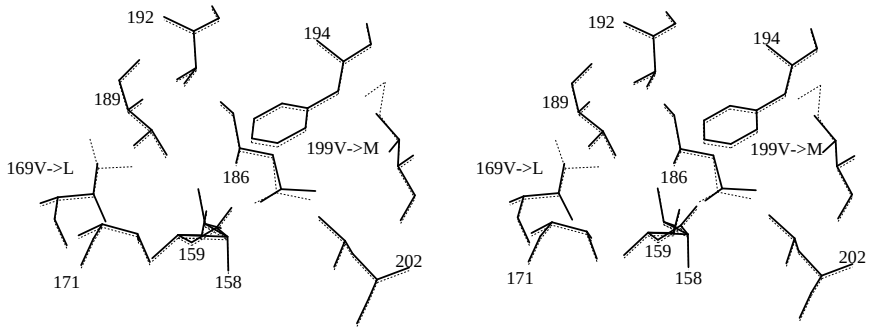

Figure 3: Stereo view of the core of V169L/V199M (dashed line) superimposed on that of the wild (solid line).

(data not shown). When the proteins were denatured with GuHCl and were refolded by dilution of the denaturant, all the proteins completely recovered their secondary structures and the enzyme activities (data not shown).

3.3 Stability

To quantitate the change in stability experimentally, CD spectra of each variant were measured as a function of GuHCl concentration at 25 °C, pH

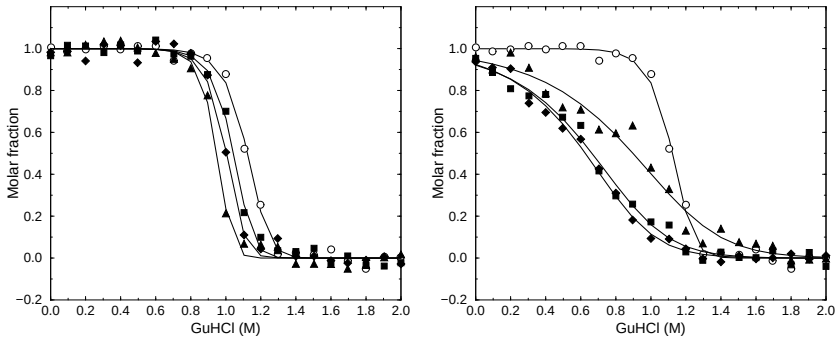

Figure 4: Molar fractions plotted against GuHCl concentration. Left figure shows wild type (circle), V169L (square), V169L/V199I (diamond) and V169L/V199I (triangle). Right figure shows wild type (circle), P_{ref} (square), V199M (diamond) and V199I (triangle).

Table 3: Thermodynamic parameters for eMDH and its variants.

protein	ΔG^{H_2O}	m^1	$\Delta\Delta G^{exp}$	$\Delta\Delta G^{cal}$	activity[2]
Wild	22.4 ± 1.7	13.2 ± 1.4	–	–	47.2
V169L	22.9 ± 2.7	14.7 ± 2.6	+0.5	+0.13	41.3
V199M	10.2 ± 0.5	4.2 ± 0.6	−12.2	+3.86	8.1
V199I	10.5 ± 0.5	3.3 ± 0.5	−11.9	+0.71	8.6
V169L/V199M	25.6 ± 3.2	19.0 ± 3.3	+3.2	+1.10	40.5
V169L/V199I	23.2 ± 2.7	15.6 ± 2.7	+0.8	−0.46	31.2
P_{ref}	10.1 ± 1.1	3.8± 0.9	−14.9	−103.8	9.8

[1] m : the slope of a denatured curve. $\Delta\Delta G^{cal}, \Delta\Delta G^{exp}$: calculated and experimental $\Delta\Delta G$ (kcal/mol), respectively. [2] unit = μmol/min

7.0 (Figure 4). The ΔG^{H_2O} and slope of denaturation curve were estimated by non-linear least-squares fitting. From the estimation, three variants in this study were found to be more stable than or equal to the wild type enzyme.

For V169L, V169L/V199M and V169L/V199I, which have similar denaturation cooperativity (m value) to the wild type, the calculated $\Delta\Delta G$ values between the wild type and variants were consistent with those determined by the experiments within experimental error (Table 3).

4 Discussion

Lee & Levitt predicted mutant stability based only on van der Waals energies of the folded state using a simulated annealing algorithm[8] without the energies of the unfolded state, hydration and entropic contributions. Lee applied a self-consistent ensemble optimization theory for prediction of mutant energetics[9], in which hydration and entropic contributions were also ignored. Although, their calculated energies by these methods appeared to correlate with the experimental thermostabilities of mutants of λ repressor, the energies were not quantitatively identical to those of experiment.

We previously examined with the reliability of our programs using a smaller protein, barnase, with a single domain (submitted) and the calculated $\Delta\Delta G^{H_2O}$ values of this protein and its mutants coincided with the experimental values (correlation coefficient 0.93). In this study, we applied our programs to measure a lager, multi-meric protein, eMDH, by consideration of both folded and unfolded state together with hydration, entropic contributions and van der Waals energy. The method enabled us to estimate the free energy change in between the native and unfolded state. Using this method, we designed a number of amino acid sequences with high stability. For the three proteins, V169L, V169L/V199M and V169L/V199I, which had similar denaturation cooperativ-

ity to that of the wild type, the calculated $\Delta\Delta G$ values between the wild type and variants coincided quantitatively with the values determined by the experiments within the experimental errors. In particular, the double mutant V169L/V199M was more stable than the wild type enzyme by 3.2 kcal/mol of $\Delta\Delta G$. This means that hydration and entropic contributions cannot be ignored in quantitative prediction because the stability changes are likely to be overestimated when only enthalpic energies were considered. In comparison with other methods mentioned above, our method can save computational cost because it can predict a set of sequence and their energies in a single run and does not have to sample many conformations for each set of sequences. Therefore our method may be a powerful tool to design new proteins with an increased stability.

However, the calculation was not applicable to estimate an exact $\Delta\Delta G$ value for the variants, V199I, V199M and P_{ref}, which show a different cooperativity pattern from that of the wild type. In consideration of the fact that our program may not be suitable to estimate $\Delta\Delta G$ values induced by replacement which could cause main chain movement, we may assume that the observed difference in $\Delta\Delta G$ values between the calculation and the experiment is partially attributed to the main chain movement in these variants. It is also likely that the decrease of specific activities in the variants may be caused by the slight distortion of the 3D structure. In addition, the specific activity showed that replacements of Val with Met or Ile at position 199 caused conformational change to some extent while the three variants, V169L, V169L/V199M and V169L/V199I, kept the structure similar to that of the wild type (Table 3). Their three dimensional structures must be determined to answer this question.

To be an alternative explanation, the failure in the prediction of $\Delta\Delta G$ for the three variants, V169L, V169L/V199M and V169L/V199I, with smaller $\Delta\Delta G$ than our calculation might be due to the fact that the denatured state of these variants was different from that of the wild type. Studies on staphylococal nuclease [27,28] indicated that the denatured state of variants with larger m value have a less structured denatured state with a greater exposure of nonpolar residues to solvent. This suggests an existence of multiple forms even in denatured state. To clarify these possibilities, determination of further characterization is required.

The loss of stability by more than 10 kcal/mol of $\Delta\Delta G$ for single variants, V199I and V199M, shows that the residue at 199 is a key residue for protein stability. However, there was no difference between each variant and the wild type in CD spectra at a range from 200 to 250 nm in the absence of the denaturant, suggesting that the global structure was not significantly distorted

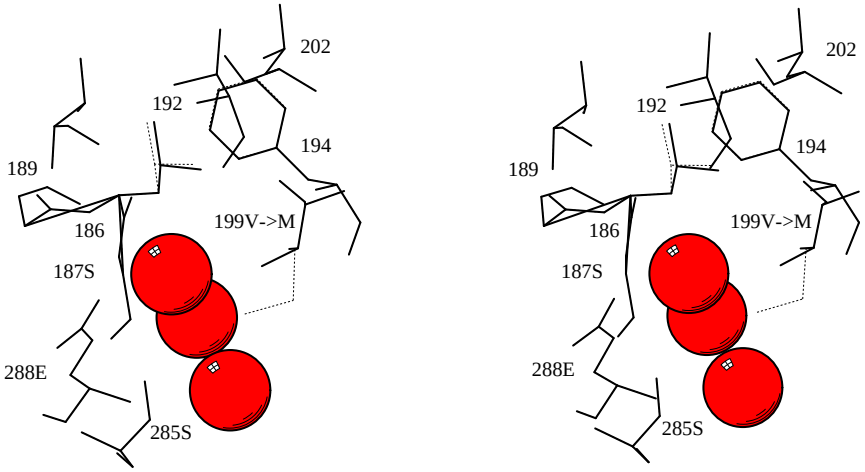

Figure 5: Stereo view of the cleft near the V199. In the crystal structure, three water molecules (sphere) in the cleft formed hydrogen bonds with S187, S285 and E288. Substitution of V199 with M (dashed line) can cause one of three waters push out.

in each variant. Although Val 199 locates at the N-terminus of the seventh helix and participates in the formation of one edge of the hydrophobic core, a cleft containing three water molecules is found in the vicinity of Val 199. We assume that one of the waters forming a hydrogen bond network with Ser 187, Ser 285 and Glu 288, would be pushed out by the substitution of Val 199 with Met or Ile and, hence, their stability of the variants were decreased (Figure 5). When additional substitution of Val 169 with Leu was introduced, the stability of the variants, V199I and V199M, were recovered. Since residue 169 located at a distance more than 15 Å from the residue 199 in the crystal structure, it is unlikely that these residues interact directly with each other. We suggest that some dynamic event may cause a repacking of the core.

Acknowledgments

We would like to thank Dr. T. Fujii for his assistant in measuring CD spectra. We also would like to thank M. Kinoshita, M. Kukimoto, N. Kobashi and N. Kudo for their help in engineering variants. Finally, we would like to thank Dr. Hao for carefully reading the manuscript.

References

1. W. F. DeGrado *et al*, *Science* **243**, 622-628 (1989).
2. C. Sander *et al*, *Proteins:Struct.Funct.Genet.* **12**, 105-110 (1992).
3. M. H. Hecht, *Proc.Natl.Acad.Sci.USA* **91**, 8729-8730 (1994).
4. A. Godzik, *Protein Engineering* **8**, 409-416 (1995).
5. D. Jones *et al*, *Proteins:Struct.Funct.Genet.* **24**, 502-513 (1996).
6. H. W. Hellinga and F. M. Richards, *Proc.Natl.Acad.Sci.USA* **91**, 5803-5807 (1994).
7. J. R. Desjarlais and T. M. Handel, *Protein Science* **4**, 2006-2018 (1995).
8. C. Lee and M Levitt,*Nature (London)* **352**, 448-451 (1991).
9. C. Lee, *J.Mol.Biol.* **236**, 918-939 (1994).
10. C. Wison *et al*,*J.Mol.Biol.* **220**, 495-506 (1991).
11. W. A. Lim and F. M. Richards, *Quarterly Reviews of Biophysics* **26**, 423-498 (1994).
12. H. Kono and J. Doi, *Proteins:Struct.Funct.Genet.* **19**, 244-255 (1994).
13. H. Kono and J. Doi, *J.Compt.Chem.* , in press (1996).
14. T. Ooi *et al*, *Proc.Natl.Acad.Sci.USA* **84**, 3086-3090 (1987).
15. G. Perrot*et al*, *J.Compt.Chem.* **13**, 1-11 (1992).
16. M. J. E. Sternberg and J. S. Chickos, *Protein Engineering* **7**, 149-155 (1994).
17. J. S. Weiner *et al*, *J.Am.Chem.Soc.* **106**, 765-784 (1984).
18. M. D. Hall *et al*, *J.Mol.Biol.* **226**, 867-882 (1992).
19. T. Ishii *et al*, *Transactions of Information Processing Society of Japan* **32**, 590-598 (1991).
20. P. Kraulis, *J.Appl.Cryst.* **24**, 946-950 (1991).
21. W. Ito *et al*, *Gene* **102**, 67-70 (1991).
22. F.S. Sanger *et al*, *Proc.Natl.Acad.Sci.USA* **74**, 5463-5467 (1977).
23. J. Messing and J. Vieira, *Gene* **19**, 269-276 (1982).
24. M. Nishiyama *et al*, *J.Biol.Chem.* **261**, 14178-14183 (1986).
25. M. S. Gittelman and C. R. Matthews, *Biochemistry* **29**, 7011-7020 (1990).
26. M. Nishiyama *et al*, *J.Biol.Chem.* **268**, 4656-4660 (1993).
27. D. Shortle and A. K. Meeker, *Proteins:Struct.Funct.Genet.* **1**, 81-89 (1986).
28. D. Shortle *et al*, *Biochemistry* **29**, 8033-8041 (1990).