Design and Synthesis of a Globin Fold†

Yasuhiro Isogai,*,‡ Motonori Ota,§ Tetsuro Fujisawa,|| Hiroyuki Izuno,⊥ Masahiro Mukai,‡ Hiro Nakamura,‡ Tetsutaro Iizuka,‡ and Ken Nishikawa§

The Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan, National Institute of Genetics, Yata, Mishima, Shizuoka 411-8540, Japan, The Institute of Physical and Chemical Research (RIKEN), RIKEN Harima Institute, Mikazuki-cho, Sayo, Hyogo 679-5143, Japan, and Department of Physics, Faculty of Science, Gakushuin University, Mejiro, Toshima-ku, Tokyo 170-0031, Japan

Received December 21, 1998; Revised Manuscript Received April 6, 1999

ABSTRACT: We propose a simple method to find an amino acid sequence that is foldable into a globular protein with a desired structure based on a knowledge-based 3D−1D compatibility function. An asymmetric α-helical single-domain structure of sperm whale myoglobin consisting of 153 amino acid residues was chosen for the design target. The optimal sequence to fit the main-chain framework has been searched by recursive generation of the protein 3D profile. The heme-binding site was designed by fixing His64 and His93 at the distal and proximal positions, respectively, and by penalizing residues that protrude into the space with a repulsive function. The apparent bumps among side chains in the computer model of the converged, self-consistent sequence were removed by replacing some of the bumping residues with smaller ones according to the final 3D profile. The finally obtained sequence shares 26% of sequence with the natural myoglobin. The designed globin-1 (DG1) with the artificial sequence was obtained by expression of the synthetic gene in *Escherichia coli*. Analyses using size-exclusion chromatography, circular dichroism spectroscopy, and solution X-ray scattering showed that DG1 folds into a monomeric, compact, highly helical, and globular form with an overall molecular shape similar to the target structure in an aqueous solution. Furthermore, it binds a single heme per protein molecule, which exhibited well-defined spectroscopic properties. The radius of gyration of DG1 was determined to be 20.6 Å, slightly larger than that of natural apoMb, and decreased to 19.5 Å upon heme binding based on X-ray scattering analysis. However, the heme-bound DG1 did not stably bind molecular oxygen as natural globins do, possibly due to high conformational diversity of side-chain structures observed in the NMR and denaturation experiments. These results give insight into the relationship between the sequence selection and the structural uniqueness of natural proteins to achieve biological functions.

De novo protein design is now thought to be an essential approach to elucidate the principles of protein architecture and has potential applications to yield novel molecules for medical and industrial aims such as drug discovery (for recent reviews, see refs *1*−*3*). The novel paradigm to design artificial sequences that fold into a desired three-dimensional (3D)[1] structure may originate from the conceptual proposal in the early 1980s, at nearly the same time that the term 'protein engineering' began to be used (*4*, *5*). Since then, many efforts have been made to determine the structural factors that govern the folding, stability, and functions of artificial and native proteins using both experimental and theoretical approaches. The first successful design of a well-defined global fold was achieved on a four-helix-bundle motif which is widely found in native protein structures (*6*−*8*). The methodology was based on manual model building using the simple binary patterns of polar and nonpolar residues along the sequences for the helices. The sequence selection for this fold has been made more sophisticated to increase the local structural specificity, and a native-like, highly ordered structure has been obtained (*9*−*11*). Besides symmetrical helical structures, successful, or at least partially successful, designs of β-sheets containing folds have been achieved by considering the geometry and statistical preferences of amino acids for the secondary structure (*12*−*14*). On the other hand, systematic and quantitative methods using computer algorithms have been developed to establish a general strategy for designing a desired, more challenging 3D structure. These design algorithms, which mainly originate from 3D structure prediction, have been tested and advanced in redesigning separate parts of native proteins

† This work was supported in part by the Biodesign Research Program and MR Science Program of RIKEN (to Y.I., H.N., and T.I.), by the Special Postdoctoral Researchers' Program of RIKEN (to M.M.), and by Grant-in-Aids for Scientific Research from the Ministry of Education, Science, Culture, and Sports of Japan (to Y.I. and to K.N.).

* To whom correspondence should be addressed. Fax: 81-48-462-4660. E-mail: yisogai@postman.riken.go.jp.

‡ The Institute of Physical and Chemical Research (RIKEN).

§ National Institute of Genetics.

|| RIKEN Harima Institute.

⊥ Gakushuin University.

[1] Abbreviations: CD, circular dichroism; C_m, midpoint denaturant concentration; 1D, one-dimensional; 3D, three-dimensional; DG, designed globin; Gd-HCl, guanidine hydrochloride; HPLC, high-performance liquid chromatography; K_d, dissociation constant; Mb, myoglobin; M_r, relative molecular mass; NMR, nuclear magnetic resonance; PAGE, polyacrylamide gel electrophoresis; p*I*, isoelectric point; R_g, radius of gyration; SCS, self-consistent sequence; SDS, sodium dodecyl sulfate; TFA, trifluoroacetic acid; UV, ultraviolet.

10.1021/bi983006y CCC: $18.00
Published on Web 05/15/1999

(*15*−*20*) and also in the entire sequence of a $\beta\beta\alpha$ protein motif, a small structural unit found as a part of the whole protein (*21*). These achievements open a possibility of de novo creation of artificial proteins with novel structures and functions.

The metal ions and other prosthetic groups have been introduced into the protein design for the purposes of probing/regulating the structure and dynamics (*22*−*27*) and of addition of oxidation−reduction functions to produce catalytically active proteins (*28*−*32*). Heme or iron protoporphyrin IX has been used mainly in the latter aspect and successfully introduced into designed four-helix bundles (*33*, *34*). The hemes are associated with the proteins by ligation of the two histidine residues and have shown visible absorption spectra typical of a six-coordinated state as seen in native *b*-type cytochromes. Such artificial heme proteins exhibit well-defined spectroscopic and electrochemical properties, and they provide simplified models of complicated native redox enzymes, which will enable specific functions of the cofactors in natural redox proteins to be understood.

Globins including myoglobin (Mb) and hemoglobin (Hb) are probably the most intensively studied proteins in all biochemical and biophysical aspects, including function, structure, stability, folding, molecular evolution, and the interrelationships between these (for reviews, see refs *35*−*37*). They bind hemes and are involved in storage and transport of molecular oxygen in a wide variety of organisms. As globins show remarkable sequence diversity, with less than 20% identity, between evolutionally distant pairs that have the common fold and function (*38*), only a small number of well-conserved residues in globins would be required to maintain the fold and its stability and also for binding and functionalizing the heme. Two of these residues, the so-called 'proximal' and 'distal' histidine residues (His93 and His64, respectively, in sperm whale Mb), are involved in coordination of the heme iron and in regulation of ligand-binding reactions, respectively. Mutagenesis experiments (e.g., refs *39*−*42*), as well as chemical modification (*43*) and module substitution (*44*), have been extensively performed to gain insights into the mechanisms of ligand recognition and heme autoxidation, the relationship between heme binding and protein stability, and the allosteric interactions in Hbs.

We have developed knowledge-based potentials from the protein structure database to evaluate the compatibility between protein tertiary structures and amino acid sequences (*45*). The compatibility score function is composed of the following four terms: side-chain packing, hydration, hydrogen bonding, and local conformation potentials, which are normalized referring to the random environmental state equivalent to the virtual-denatured state. This new normalization scheme, Minus Average Operation (MAO), leads to transformation of the 3D profile table, that is, a $20 \times N$ (sequence length) two-dimensional array representing the fitness of 20 amino acids to each structural site of a protein, into a score table of folding energy (ΔG) of every single-point mutant. The function has been successfully utilized for structural stability analysis of mutant proteins (*46*), the determination of structurally important sites in globins (*47*), and the search for sequences to fold into a target structure among the natural sequence/structure databases, i.e., the inverse-folding search (*48*), as well as the tertiary structure prediction (*49*).

In this paper, we propose a new algorithm with the compatibility function for finding the sequence that fits to a given protein backbone structure. The globin fold of sperm whale Mb was chosen for the target structure to test the validity of the algorithm for protein design. The reasons for choosing the globin fold were the following: (1) much information about the native globins and their mutants is available to compare with and evaluate the designed protein; (2) the globin fold is representative of single-domain globular proteins with independent structure and function; (3) the globins have an α-helical structure, which is considered to be more tractable than a β-sheet-containing structure at the first step of designing a sizable protein with asymmetric structure; (4) the globins bind heme, which can be used for spectroscopically probing the fold and microenvironments around the heme; (5) the globins have exquisite biological functions, which we set as the research goal of protein design. An artificial protein with the designed globin sequence was obtained by expression of the synthetic gene in *Escherichia coli*. The structural properties of the designed globin were extensively studied and compared to those of native Mbs. The relationships between the structural uniqueness and the distinctive roles of hydrophobic residues in protein folding are discussed with a view to designing functional proteins.

MATERIALS AND METHODS

Searching the Sequence. The 3D structure coordinates of sperm whale Mb, Protein Data Bank (PDB) code 1mbd (*50*), were used for the target backbone structure. An entire amino acid sequence to fold into the Mb structure was searched for by making the 3D profile recursively as shown in Figure 1. In the first step, the 3D profile of the native Mb was constructed using pseudo-energy functions according to the procedure of Ota and Nishikawa (*45*). In this profile (see Figure 2A), the best amino acid type to fit the structural environment at each position is shown in the left side column of the profile table. Thus, the output sequence constituting these highest-scoring residues is a candidate for the optimal sequence and is used as the input sequence for the next calculation step as shown in Figure 1. The new structural environments were formed by mounting this sequence on the 3D structure, and the second output sequence was obtained by the reconstruction of the new 3D profile (not shown). These operations were iterated in the same way until the sequences become identical. Note that in the present calculation of a profile table, the structural environment for a given residue implies not only purely structural features described with atomic coordinates but also sequence information of surrounding residues, because pairwise (two-body) potentials such as side chain−side chain interactions depend on the amino acid type of counterparts as well as on the relative disposition of two residues (this is, however, not the case if only single-body potentials are used). In short, our 3D profile implicitly contains sequence information besides the structural features of a protein. This is why an *initial* sequence was set at the beginning and then iterative calculations were carried out to obtain an *ideal* sequence that best fits a given structure. When some solutions oscillate between two sequences (input and output sequences in Figure 1), a few mutations were introduced into the relevant sites in order to escape from the local minima. The two His residues at positions 64 and 93, which play essential roles

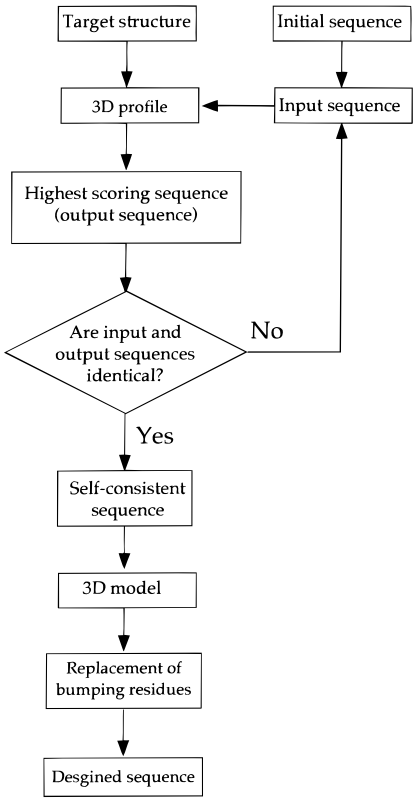

FIGURE 1: Flow diagram of the sequence design using the 3D profile. In the present study, the sperm whale Mb structure [PDB code: 1mbd (*50*)] and the native sequence were used for the target structure and for the initial sequence, respectively. His64 and His93 were fixed in the input sequences during the calculation. Residues protruding into the space for the heme-binding site were penalized with a repulsive function in the 3D profiles (see Materials and Methods).

in binding and functionalizing heme in the natural Mb, are fixed in the input sequences during the calculation. The space for the heme insertion was reserved by penalizing residues that protrude into the space with a repulsive function (*45*).

Molecular model building and molecular mechanics calculation were performed using Insight II (MSI) equipped with the optional modules Biopolymer and Discover on the basis of an extensible systematic force field with a minimization algorithm of the steepest descent system. Bumps among the residues were defined as van der Waals overlaps between the residue atoms of more than 10% of the sum of the van der Waals radii of the atoms.

Gene Construction, Cloning, and Expression. The designed globin-1 (DG1) gene was designed from the artificial amino acid sequence obtained here (see Results) using the *E. coli* optimal codons. Synthetic oligonucleotides for the gene fragments were phosphorylated on their 5′ termini, annealed, ligated, and cloned into a pRSET-C vector (Invitrogen). Clones were screened by restriction analysis, and the correct sequence was verified by DNA sequencing with a single primer extension/dye terminator method. The resultant full-length DG1 gene in the vector was expressed under the control of T7 promoter in *E. coli* strain BL21(DE3).

Purification. The harvested recombinant cells suspended in 10 mM Tris-HCl (pH 8.0) and 2 mM EDTA were treated with 0.1 mg/mL lysozyme, lysed by sonication, and centrifuged at 160000*g* for 1 h to separate the soluble fraction from insoluble materials. DG1 was found almost exclusively in the supernatant (see Results). This solution was loaded onto a Q Sepharose Fast Flow (Pharmacia Biotech) anion-exchange chromatography column equilibrated with a buffer containing 20 mM Tris-HCl (pH 8.0) and 50 mM NaCl. It was washed with the same buffer and was eluted with a buffer containing 20 mM Tris-HCl (pH 8.0) and 150 mM NaCl. The eluate was dialyzed against an aqueous solution containing 0.1% TFA and applied onto a C18 reversed-phase preparative HPLC column (Vydac 218TP1022). DG1 was purified to homogeneity with a gradient of 50−70% acetonitrile in the presence of 0.1% TFA. The DG1-containing fractions were collected, evaporated to remove acetonitrile, and dialyzed against a buffer solution used in the following experiments. The apparent molecular mass and the purity were examined by SDS−PAGE with 15% (w/v) polyacrylamide gel and C18 reversed-phase analytical HPLC with a Vydac 218TP54 column, respectively. The protein identity was verified by laser-desorption mass spectrometry (MALDI/TOFMS) and N-terminal amino acid sequencing. DG1 concentrations were determined spectrophotometrically using $\epsilon_{280} = 14.4\ \text{mM}^{-1}\ \text{cm}^{-1}$ based on $5.8\ \text{mM}^{-1}\ \text{cm}^{-1}$ for Trp and $1.4\ \text{mM}^{-1}\ \text{cm}^{-1}$ for Tyr at pH 8.

Horse metMb was purchased from Sigma and used to compare structural properties of DG1 with those of natural Mb. ApoMb was prepared from the metMb as follows. The metMb was solubilized in deionized water at several millimolar and dialyzed overnight against an aqueous solution containing 0.1% TFA. The resultant acid-denatured Mb was concentrated by centrifugation with Centriprep-10 (Amicon) and applied onto the C18 reversed-phase preparative HPLC column. ApoMb was purified with a gradient of 40−60% acetonitrile in the presence of 0.1% TFA to remove heme. The apoMb was collected, evaporated to remove acetonitrile, and dialyzed against a buffer solution used in the following experiments. Concentrations of metMb and apoMb were determined spectrophotometrically using $\epsilon_{409} = 157\ \text{mM}^{-1}\ \text{cm}^{-1}$ (*51*) and $\epsilon_{280} = 14.4\ \text{mM}^{-1}\ \text{cm}^{-1}$.

Introduction of Heme. Concentrated solutions of iron protoporphyrin IX (heme) were freshly prepared in the solvent dimethyl sulfoxide at several millimolar. Heme concentrations were determined spectrophotometrically using $\epsilon_{385} = 49\,000\ \text{M}^{-1}\ \text{cm}^{-1}$ in an aqueous solution at pH 8.0 with hemin concentrations between 1 and 5 μM. The heme was introduced into DG1 by titrating the protein in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl with the hemin solution in increments of 0.1−0.2 to a small excess amount. After each addition of heme, the mixed samples were incubated for at least 5 min at room temperature. The solvent and unbound hemin were removed by dilution with a buffer solution and by concentration with Centriprep-10. The heme-DG1 solution was centrifuged to remove insoluble aggregates, and the resulting supernatant was used in the following experiment. The dissociation constant (K_d) of DG1 with heme was determined as follows. Hemin solubilized in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl at several micromolar was titrated with concentrated protein solution (~1 mM) in the same buffer. After each addition of the protein solution, the mixture was incubated for 5 min at a room temperature, and the UV−visible absorption spectra were measured. The increase in the Soret absorbance peak was plotted against the protein concentration and analyzed using a single-site binding equation (*52*).

Size-Exclusion Chromatography. The chromatography was performed on a Hitachi L-6200 HPLC system with an L-4200 UV−visible detector using a Bio-Silect SEC 125-5 gel filtration column (Bio-Rad). It was operated with a mobile phase of 20 mM Tris-HCl (pH 8.0) and 200 mM NaCl at a flow rate of 1 mL/min. The column was standardized using the following commercially available globular proteins: γ-globulin (158 kDa), albumin (67.0 kDa), ovalbumin (43.0 kDa), carbonic anhydrase (29.0 kDa), triosephosphate isomerase (26.6 kDa), Mb (17.6 kDa), α-lactalbumin (14.4 kDa), ribonuclease (13.7 kDa), cytochrome *c* (12.4 kDa), and aprotinin (6.5 kDa).

Small-Angle X-ray Scattering (SAXS). The measurement of the SAXS pattern was done at RIKEN structural biology beamline I (BL45XU) of SPring-8 (*53*, *54*). The detailed description of SAXS optics will be given elsewhere (Fujisawa et al., manuscript in preparation). The half size of focus at BL45XU is 0.2 mm (vertical) × 0.4 mm (horizontal). The wavelength of the X-ray was 1.0 Å. The detector was an X-ray image intensifier with cooled CCD (XR−II+CCD) (*55*). For the SAXS experiments, samples were solubilized in 10 mM Tris-HCl (pH 8.0), 50 mM NaCl, and 0.1% octyl glucopyranoside at various concentrations. Mb did not show any scattering change by radiation damage until 10 s. The typical collection time was less than 2 s. Incident intensity was scaled by the current of an ionization chamber installed before the position of the sample. Prior to the measurements, all samples were centrifuged at 15 000 rpm to remove precipitates. Preliminary data processing was performed using a FORTRAN program, iisgnapr, which takes the circular average of sample and buffer images. Prior to the buffer subtraction, each image was subtracted by the dark current, and then scaled by the ion-chamber current. The R_g value was determined by the Guinier approximation: $I(S) = I(0) \exp(-4\pi^2 R_g^2 S^2/3)$, where S and $I(0)$ are the momentum transfer and intensity at the zero scattering angle, respectively (*56*). S is defined as $S = 2 \sin \theta/\lambda$, where 2θ and λ are the scattering angle and the X-ray wavelength, respectively. The S range used for R_g determination satisfied the condition $2\pi SR_g < 1.3$. The detector-to-sample length was determined by the meridian reflection of collagen (0.5 m).

Spectroscopy. UV−visible spectra were recorded with a Hitachi U-3000 spectrometer using quartz cuvettes of 1.0 cm in path length. Circular dichroism (CD) spectra were recorded at 20 °C with a JASCO J700 spectropolarimeter using rectangular quartz cuvettes of 0.2 cm in path length with protein concentrations between 2 and 10 μM under conditions in which the spectrum of DG1 was not affected by the protein concentrations. The resonance Raman spectra were measured with a single spectrophotometer (JASCO NR-1800) equipped with a cooled CCD device (Princetone Instruments). The excitation sources were Kr^+ laser (406.7 and 413.1 nm, Coherent). One-dimensional proton NMR measurements were performed on a Bruker ARX400 spectrometer.

Denaturation. Quantitative disruption of protein structures with the denaturant guanidine hydrochloride (Gd-HCl) was monitored by measuring the CD signal intensity at 222 nm or by measuring the Soret absorbance peaks of heme-bound proteins as a function of Gd-HCl concentration. Small volumes of a protein stock solution were added to 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl containing Gd-HCl to give a final protein concentration at 5 μM. The mixed solutions were incubated to reach equilibrium for at least 120 min, and the CD or absorption spectra were measured at 20 °C. The denaturation data were analyzed with a theoretical curve based on $\Delta G_{app} = -RT \ln K_{app} = \Delta G^\circ_{app} + m$[Gd-HCl] by assuming the two-state folding−unfolding transition with the equilibrium constant K_{app}. In this equation, ΔG_{app} and ΔG°_{app} are apparent free-energy changes from the folded state to the unfolded state in the presence and absence of denaturant, respectively, and m is the dependence of ΔG_{app} on the denaturant concentration, which measures the cooperativity of the two-state transition. We used 'apparent (app)' here because neither intermediate nor heme dissociated from unfolded protein in the equilibrium was considered for simplifying the calculation of thermodynamic parameters (see refs *57*, *58*).

(A) Original 3D profile of native myoglobin

N	I	H	S	R	O 3D profile table
1	V	3	e	16	PEKQTHSGDNRACYMVWIFL
2	L	8	e	2	PLIMCTFVYWSHGNAREKQD
3	S	1	e	2	DSTPNGEKHACQRYMLVWIF
4	E	1	a	2	PEDSKQRATGNHVCYLMWIF
5	G	1	a	11	EDSKNQPATRGMHCVYLWFI
6	E	4	a	1	EDWQARNTSKYHMGFLCPVI
7	W	3	a	4	EAQWMKDFLRTNHSVYICGP
8	Q	1	a	5	KEDAQRNTSGHMCPLYVWFI
9	L	6	a	4	WAMLFQCVIRYHTSNEDKGP
10	V	7	a	4	LWIVFMRYAQCTHEKNDSGP
11	L	4	a	6	EADQMLTRNKVIYCSGHWFP
12	H	6	a	10	EADQRKLNMHCTVIWYFSGP
13	V	9	a	3	ILVWMFAYCHTRQESDNGPK
14	W	7	a	1	WLFMYIHARQVCKTENDSGP
15	A	2	a	4	EKQADRNSTGMLHICYVFWP
•	•	•	•	•	•

(B) Final 3D profile of SCS1 on the Mb backbone structure

N	I	H	S	R	O 3D profile table
1	P	3	e	1	PEQKSGTHDNRAYCWVMFIL
2	P	8	e	1	PCVYLIFTHMSWNRGAEDQK
3	D	1	e	1	DSTNEPGKHAQCRYMLVWIF
4	P	1	a	1	PEDSKQATRGNHCVYLFMWI
5	E	1	a	1	EDKNQSRPTAGHMCYWILVF
6	R	4	a	1	RDEQNTHYVMWFKSCALPGI
7	K	3	a	1	KREQDMNWAYHLTFSVCIGP
8	K	1	a	1	KEQDRANSTGHMPLVCYWIF
9	R	6	a	1	RWFYMAHLNIQEVCKDSTGP
10	W	7	a	1	WFLMYHAQVICERTDNSKGP
11	E	4	a	1	EDQMANWKTRLYFICHVSGP
12	E	6	a	1	EMDLQARIWFNTYVCKHSGP
13	I	9	a	1	ILVWFMAYHCRTQESNDPGK
14	F	7	a	1	FWLMIYVACHRQETDNKSGP
15	K	2	a	1	KEQARDNSTGHMCLYWFVIP
•	•	•	•	•	•

FIGURE 2: 3D profiles made from the sperm whale Mb structure [PDB code: 1mbd (*50*)] with the native sequence (A) and with the converged self-consistent sequence SCS1 (B). The tables for the first 15 residue sites, which were excerpted from the complete tables consisting of the entire rows for the 153 residue sites of the Mb structure, are shown. The tables contain residue numbers (N), amino acids of the input and output sequences (I and O, respectively), hydration classes (H), secondary structures (S), and ranks of the input amino acids (R) besides the 3D-profile tables. In the profile tables, the 20 amino acids denoted by their one-letter code are arranged in order of the compatibility score from left to right at each site. The amino acids of the input sequences are highlighted in black. For a more detailed description of the 3D profile, see ref *45*.

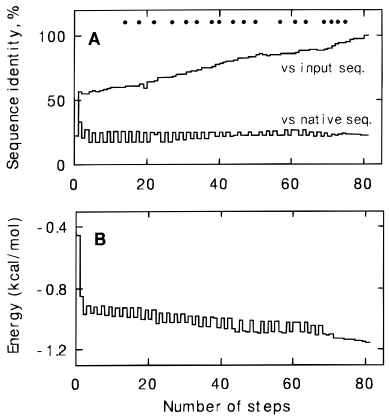

FIGURE 3: Convergence of sequences selected for the globin structure during iteration of the 3D profile calculation starting from the native sequence. Trajectories of sequence identities to the native sequence (lower trace) and to an input sequence (upper trace) and of the average residue score are shown in panels A and B, respectively. Dots in the upper part in panel A indicate the occurrence of sequence oscillation and the perturbation of sequences.

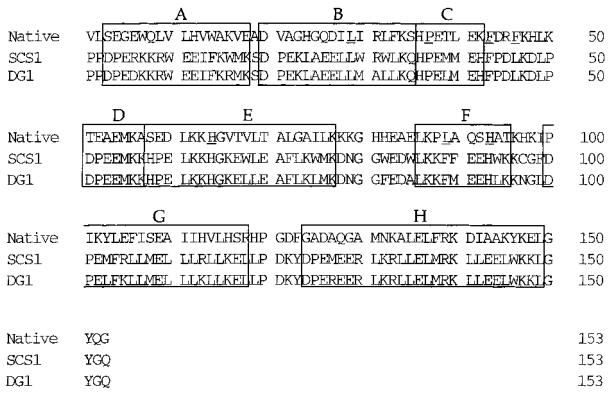

FIGURE 4: Designed globin sequences compared with the native sequence. The regions corresponding to helices A−H in the globin structure are shown in the boxes. The underlined sites on the native sequence indicate the highly conserved residues in natural globins (*38*). The shaded sites indicate the sites conserved between the native sequence and the artificial sequences.

RESULTS

Sequence Design. The self-consistent sequence-1 (SCS1) for the sperm whale Mb structure was computationally obtained by the recursive generation of an artificial sequence and its 3D profile for the structure as described under Materials and Methods. The initial and final 3D profiles with the native sequence and SCS1 as the input sequences are shown in Figure 2A,B, respectively. In the initial 3D profile (Figure 2A), the fitness rank position of the input sequence is about sixth on average, which is better than the random level (10.5th) although there still appears to be room for optimizing. Thus, the best scoring amino acids on the left side of the profile table were defined as the output sequence of this profile and were then used as the input sequence for the next step of the profile calculation as shown in Figure 1. After iterative calculation, the output sequences became identical with the total compatibility score minimized (see Figure 3), and the final 3D profile was obtained (Figure 2B). Note that all of the input amino acids are ranked as best in the 3D profile of SCS1, and the input and output sequences are the same except for the distal and proximal His positions at 64 and 93, which were fixed in all the input sequences and SCS1 to create the heme-binding site in the artificial globin. The space for the heme insertion was made by penalizing residues that protrude into the space in the iterative calculation with a repulsive function (*45*).

As shown in Figure 3A, the trajectory of identity of output and input sequences at each iteration step (Figure 3A) started from 23%, increased to 57% by the second step, and then continued to increase gradually, reaching 99% or SCS1 at the 80th step; whereas that of the output and the native sequences started from 23%, remained with slight fluctuations at around 25%, and finally reached the convergent point. On the other hand, the progress of optimizing was monitored by the trajectory of the compatibility scores shown in Figure 3B, which indicates oscillation but continuous decrease in the averaged residue energy. It decreased drastically in the first step from −0.45 to −0.85 kcal/mol of residue, gradually thereafter, and finally reached −1.16 kcal/mol of residue. This indicates that the trajectory dropped into the region surrounding the deep minima in the sequence space at the first step, and the 80 steps were needed to reach the convergent point.

Table 1: Amino Acid Compositions of Natural (Sperm Whale) and Designed Globins with Their Calculated Molecular Masses, Isoelectric Points, and Sequence Identities to the Native Sequence

amino acid	native	SCS1	DG1
Ala	17	2	4
Cys	0	1	0
Asp	7	10	11
Glu	14	25	25
Phe	6	7	6
Gly	11	6	6
His	12	5	5
Ile	9	1	1
Lys	19	24	26
Leu	18	27	34
Met	2	9	8
Asn	1	1	2
Pro	4	2	12
Gln	5	2	2
Arg	4	8	6
Ser	6	1	1
Thr	5	0	0
Val	8	0	0
Trp	2	10	2
Tyr	3	2	2
M_r, kDa	17.2	19.2	18.6
p*I*	9.5	6.0	5.8
identity to native, %	100	26.1	26.1

The designed sequence SCS1 was compared with the native sequence of sperm whale Mb and was found to share 26% of the sequence as shown in Figure 4 and Table 1. Many identical amino acids are found in the region from helices C to E, whereas helices A and G show rather low matches. The differences in amino acid composition (see Table 1) show significant increases in the contents of Glu (+11), Leu (+9), Pro (+8), Trp (+8), and Met (+7) and decreases in Ala (−15), Ile (−8), Val (−8), and His (−7). The preference for bulky amino acids in SCS1 is attributed to the score functions used here, which do not explicitly involve the side chain structures (*45*). The negative shift in the calculated value of p*I* from 9.5 to 6.0 is mainly due to the increase in the content of Glu, which has a high α-helical propensity.

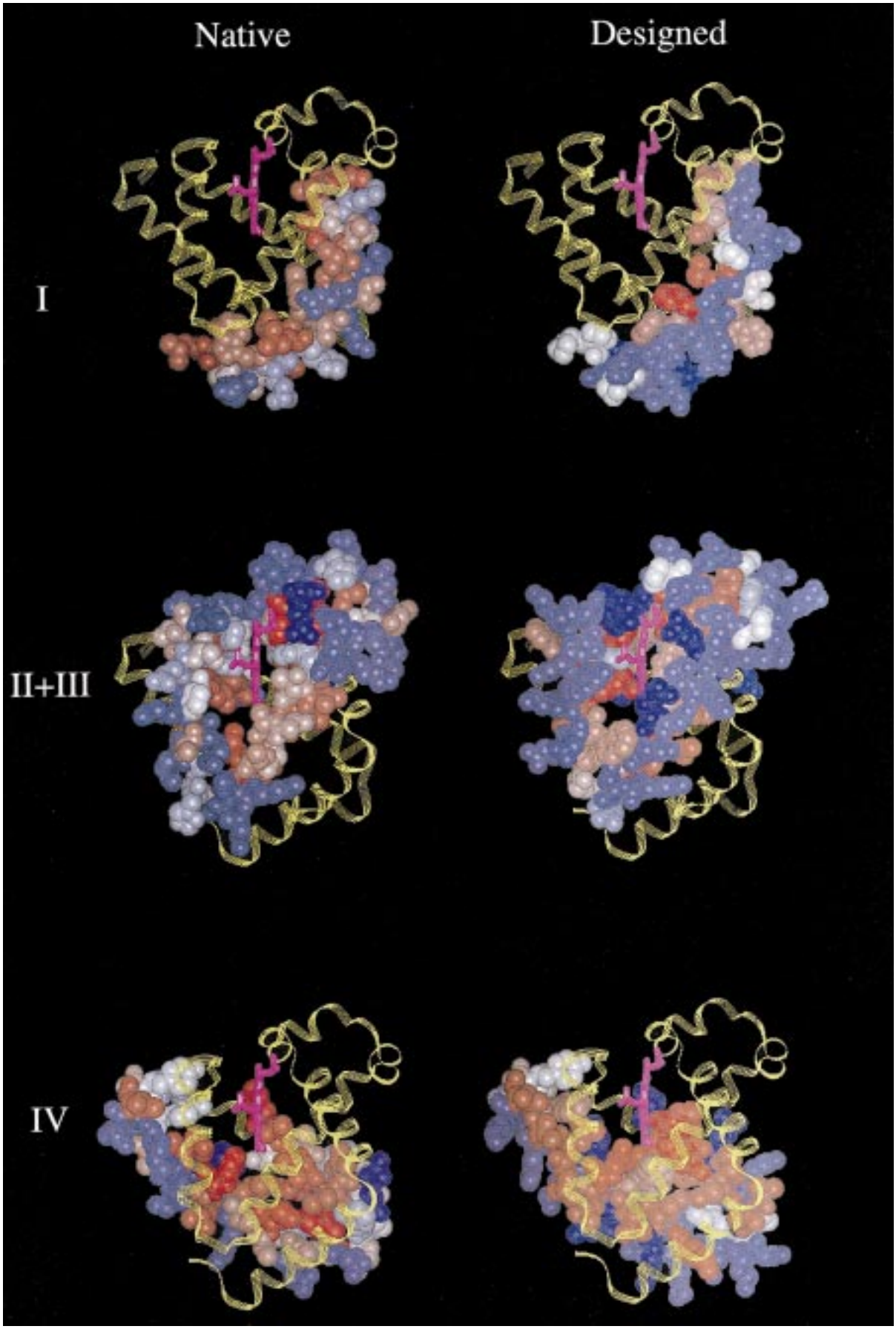

FIGURE 5: 3D model of DG1 compared with the crystal structure of sperm whale Mb (1mbd). The distributions of hydrophilic (blue), neutral (white), and hydrophobic (red) residues are shown in the CPK rendering. The sequences are divided into three regions: modules I, II+III, and IV (*59*), which are displayed separately on each model for clarity. The backbone structures are shown by yellow ribbons.

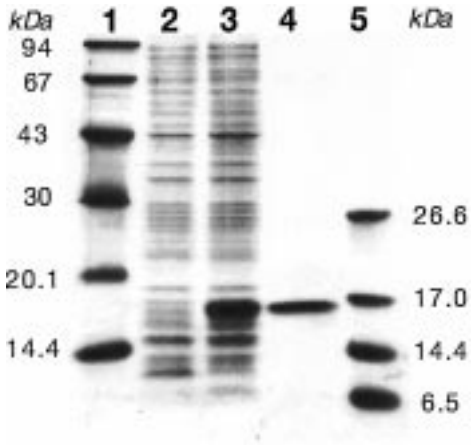

FIGURE 6: SDS−polyacrylamide gel electrophoresis showing the expression and purification of DG1. Lanes 2 and 3 show cell extracts of control and transformant *E. coli* cells, respectively, into which the vector pRSET-C with (lane 3) and without (lane 2) the DG1 gene was introduced. Lane 4 shows purified DG1 after preparative reversed-phase HPLC (see Materials and Methods). Lanes 1 and 5 are molecular size markers: phosphorylase *b* (94.0 kDa), albumin (67.0 kDa), ovalbumin (43.0 kDa), carbonic anhydrase (30.0 kDa), trypsin inhibitor (20.1 kDa), and α-lactalbumin (14.4 kDa) in lane 1; and triosephosphate isomerase (26.6 kDa), equine Mb (17.0 kDa), α-lactalbumin (14.4 kDa), and aprotinin (6.5 kDa) in lane 5.

An artificial-globin 3D model was constructed by mounting the designed sequence SCS1 onto the target backbone structure. The distribution of hydrophobic and hydrophilic residues in the model with SCS1 was consistent with the 3D structural topology, i.e., hydrophilic outside and hydrophobic inside. However, bumps among many residue pairs were found in the 3D model. This is easily expected from the calculated molecular mass of 19.2 kDa, which is larger than the natural Mb by 2.0 kDa or 12%. These bumps are removable within 0.12 Å of rms deviation of the backbone atoms from the target structure by replacement of several amino acids with smaller ones according to the 3D profile of SCS1 (Figure 2B). The finally obtained designed globin-1 (DG1) sequence has an average compatibility score of −1.03 kcal/mol of residue, slightly larger than −1.12 kcal/mol of residue of SCS1. These artificial sequences and their compositions are compared with those of sperm whale Mb in Figure 3 and Table 1. Bumps were removed by replacing bulky residues such as Trp with Leu or other smaller residues at hydrophobic sites; consequently, the molecular mass of DG1 decreased slightly to 18.6 kDa, and the Leu content increased to nearly 2 times that of natural Mb (Table 1). DG1 has 88% and 26% sequence identities with SCS1 and Mb, respectively. The 3D model of DG1 is displayed with the crystal structure of the natural Mb for comparison in Figure 5. In this model of DG1, there are no bumps between any nonbonded atoms as mentioned above, and intramolecular hydrogen bonds are detected more than those in the native Mb structure (270 vs 253).

Synthesis and Purification. A full-length DNA for DG1 was constructed from synthetic oligonucleotides, cloned, and expressed in *E. coli*. The recombinant cells efficiently synthesized the gene product, which is almost exclusively found in the soluble fraction, without formation of inclusion bodies (Figure 6). It has an apparent molecular mass of 17 kDa in SDS−PAGE analysis, and the smaller M_r value

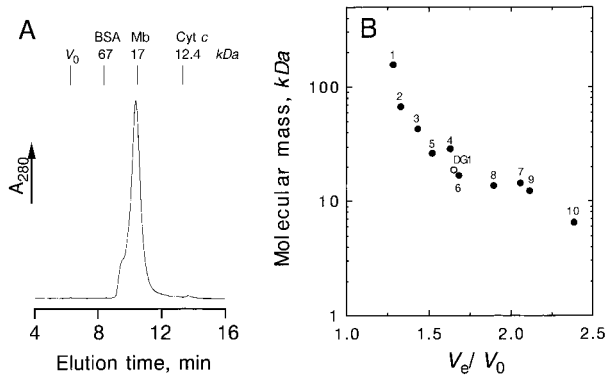

FIGURE 7: Size-exclusion chromatography of DG1. (A) Chromatogram monitored with absorbance at 280 nm. 100 μL of 5 μM protein in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl was loaded on the column. Elution times for the void volume (V_0), bovine serum albumin (BSA), horse Mb, and cytochrome *c* (Cyt c) are indicated at the top. (B) Correlation between molecular mass and elution volume (V_e/V_0) in the chromatography of DG1 (open circle) and natural globular proteins (closed circles) with various M_r and shapes. Natural globular proteins are γ-globulin (1), bovine serum albumin (2), ovalbumin (3), carbonic anhydrase (4), triosephosphate isomerase (5), horse Mb (6), α-lactalbumin (7), ribonuclease (8), cytochrome *c* (9), and aprotinin (10).

estimated by SDS−PAGE than the calculated value (18.6 kDa) is due to its negative net charges (theoretical pI = 5.8) under the experimental conditions. The product was purified to homogeneity as described under Materials and Methods with more than 95% purity judged by analytical reversed-phase HPLC, and the correct identity of the protein was confirmed using mass spectroscopy and also by chemical sequencing of several N-terminal amino acids. DG1 apparently binds heme in the cells as judged by the reduced CO minus oxidized difference absorption spectrum of the supernatant from cell lysis (data not shown). However, the heme was dissociated from the protein during purification with reversed-phase chromatography (see Materials and Methods). Purified DG1 or apoDG1 was used in the following experiments. DG1 was associated with heme in 1:1 stoichiometry by the addition of heme in vitro (see below).

Overall Molecular Shape. The elution volume of a globular protein in size-exclusion chromatography depends on both the size and molecular shape. In this chromatography, DG1 was eluted at almost the same volume as that of native Mb (Figure 7). The ratio of the elution volume (V_e) of DG1 to the void volume (V_o) of the column was plotted against the molecular mass, and the correlation between V_e/V_o and M_r was compared with those of natural globular proteins with a variety of sizes and shapes in Figure 7B, and was well consistent with that of Mb. This indicates that DG1 is monomeric and that the average overall shape of the DG1 monomer is similar to that of natural Mb under the experimental conditions and agrees with the results of solution X-ray scattering analysis mentioned below.

Solution X-ray scattering analysis provides the best quantitative measure of the overall structural dimensions of a globular protein (*56*). Native, molten globule, and unfolded conformations of natural Mbs have been extensively studied using small-angle X-ray scattering (SAXS) measurements (*60*−*62*). Analysis of DG1 showed that the aggregation effects are very small at concentrations less than 3 mg/mL,

Table 2: Experimental Structural Parameters of Natural and Designed Globins

	apo-form			holo-form		
	helical content[a] (%)	R_g[b] (Å)	$I(0)/C$[c]	helical content[a] (%)	R_g[b] (Å)	$I(0)/C$[c]
DG1	60.8	20.6 ± 0.6	331 ± 8	60	19.5 ± 0.5	332 ± 10
Mb						
native	61.6	19.7[d]	–	78.4	17.4 ± 0.7	282 ± 8
molten globule	45[d]	23.1[d]	–	–	–	–
unfolded	0	35.8[d]	–	–	–	–

[a] The values were determined by CD signal intensity at 222 nm. [b] Radius of gyration determined by small-angle X-ray scattering analysis. [c] Arbitrary unit. [d] Data were taken from ref *61*, in which the molten globule apoMb was formed and stabilized by trichloroacetate at pH 2, and the unfolded apoMb was formed by Gd-HCl.

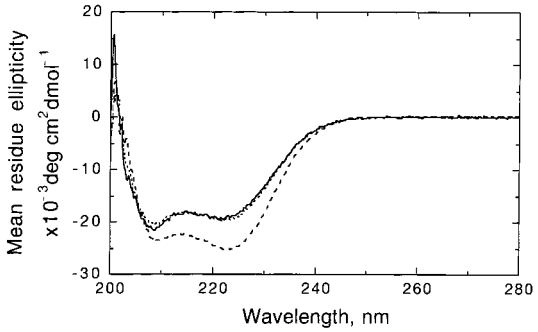

FIGURE 8: Circular dichroism spectra of DG1 (solid line), apoMb (dotted line), and holoMb (dashed line) recorded at 20 °C in 10 mM Tris-HCl (pH 8.0) and 200 mM NaCl. The spectrum of heme-DG1 (not shown) is almost the same as that of DG1. Protein concentrations were 5 μM as determined spectrophotometrically.

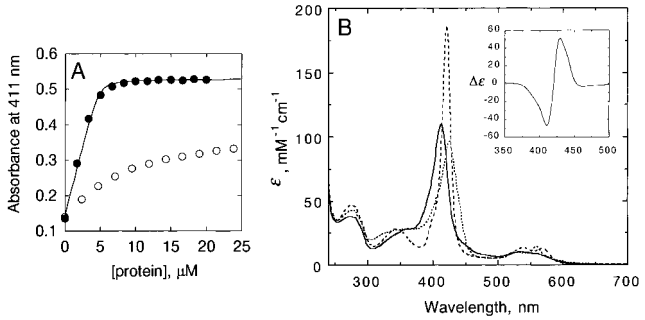

FIGURE 9: Association of heme with DG1. (A) Binding titrations of hemin with DG1 (closed circles) and a DG1(H93L) mutant (open circles). Hemin solubilized in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl at 4.8 μM was titrated with a concentrated protein solution in the same buffer. After each addition of the protein solution, the mixture was incubated for 5 min at room temperature, and the absorption spectrum was measured. The increase in the Soret absorbance peak was plotted against the DG1 concentration. The theoretical binding curve for DG1 was drawn based on a single-site binding equation with $K_d = 80$ nM. (B) Absorption spectra of the bound heme in ferric (solid line), ferrous (dotted line), and ferrous−CO (dashed line) forms recorded in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl at protein concentrations of 5 μM. Inset: The ferrous − ferric difference spectrum of heme-DG1.

where monomeric physical parameters were determined by extrapolating to infinite dilute. The forward-scattering intensity, $I(0)/C$, of DG1, where $I(0)$ and C are the intensity of zero scattering and the protein weight concentration, respectively, is comparable to that of metMb, which ensures the monomeric state of DG1 (Table 2). The R_g values of DG1 were determined to be 19.5 ± 0.5 and 20.6 ± 0.6 Å with and without heme, respectively, whereas the R_g value of metMb was determined as 17.4 ± 0.7 Å (Table 2). The estimated R_g values of DG1 are similar to the R_g value of apoMb (19.6 Å) and much smaller than that of a molten globule state (23.1 Å) and unfolded states (30−36 Å) previously reported by Kataoka et al. (*61*). These results show that DG1 folds into a monomeric globular protein with overall structural dimensions close to those of apoMb and that the dimensions decrease upon the binding of heme.

Secondary Structure. The secondary structure of a globular protein is readily probed by a far-UV circular dichroism (CD) measurement. The CD spectrum of DG1 shows its highly helical nature and is almost indistinguishable from that of apoMb (Figure 8). The mean residue ellipticities of DG1 and apoMb were determined to be −19 500 and −19 700 deg cm^2 dmol^{-1} at 222 nm, respectively. From these values, the helical contents of DG1 and apoMb under the experimental conditions are estimated at 60.8 and 61.6%, respectively, based on the ellipticity of −32 000 deg cm^2 dmol^{-1} for 100% helicity (*63*) (see also Table 2). This indicates that DG1 preserves the α-helical content to almost the same degree as that of native apoMb. In the present measurements, however, no change in the helical contents of DG1 was detected upon addition of heme, in contrast with the case of natural Mb in which the helical contents significantly increase upon addition of heme (Table 2).

Heme Binding. In the sequence design, His64 and His93 were fixed during the calculation, and DG1 has these residues which play essential roles in binding and functionalizing heme in natural globins. Then, we examined the binding of heme to DG1. Heme was successfully introduced into DG1 as shown in Figure 9. The titration data (Figure 9A) are well fitted by a theoretical curve using a single-site binding equation, indicating strictly stoichiometric binding of one heme per DG1 molecule. The dissociation constant was approximately estimated at 80 nM, which is comparable to $K_d = 5-200$ nM estimated for higher-affinity binding sites of designed four-helix bundle heme proteins (*33*, *34*) and to $K_d = 90$ nM for natural apocytochrome *c* with noncovalent Fe(II) heme (*64*). A DG1 mutant (H93L) in which His93 was replaced by Leu showed much lower affinity to heme and indeterminable binding stoichiometry as shown in Figure 9A. These results indicate that His93 coordinates the heme iron in DG1 as designed.

Heme Spectra. Heme-associated DG1 (heme-DG1) exhibited well-defined UV−visible absorption spectra as shown in Figure 9B. The ferric form showed visible absorption maxima at 413 and 535 nm, and the ferrous form at 425, 529, and 558 nm. The absorption peaks observed in the ferrous and ferric forms are characteristic of low-spin bis-His-coordinated heme iron, whereas the rather low absorption intensity of the Soret peak (425 nm) with the apparent

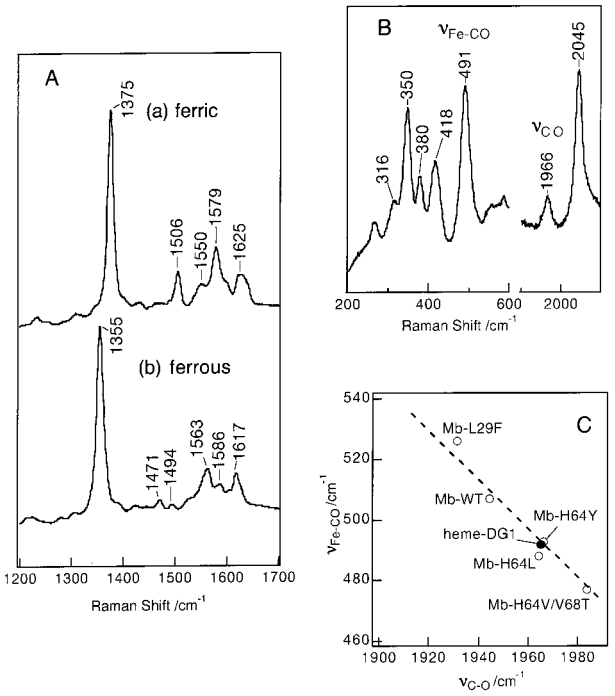

FIGURE 10: (A) Resonance Raman spectra of (a) ferric and (b) ferrous forms of heme-DG1 in the region 1200−1700 cm^{-1} by 406.7 nm excitation. The spectra were measured in 50 mM Tris-HCl (pH 8.0) and 200 mM NaCl at protein concentrations of 50 μM. The Raman cell was spun and kept below 10 °C by flushing with cold N_2 gas. (B) Resonance Raman spectra of the CO-bound form of heme-DG1 in the regions 200−600 and 1900−2100 cm^{-1}. Other conditions were the same as those in (A). (C) Correlation between the ν_{Fe-CO} and ν_{CO} frequencies of heme-DG1 (closed circle) and natural wild-type and mutant Mbs (open circles). The data of Mbs were taken from refs *68*−*70*.

shoulder at around 430 nm (see Figure 9B, inset) in the ferrous form is indicative in the presence of high-spin five-coordinated heme iron under the equilibrium between the five- and six-coordinated states. Reaction of ferrous heme-DG1 with molecular oxygen resulted in rapid oxidation of the heme iron, and the spectrum of a putative oxy form was not detected under the experimental conditions with a mixing time of several seconds. On the other hand, upon the addition of CO, ferrous heme-DG1 immediately formed a stable CO complex with absorption maxima at 345, 421, 539, and 568 nm, which are strikingly similar to those of the CO complex of natural globins (*51*) and are 3−4 nm larger than those of the CO-bound form of a designed four-helix bundle heme protein (*65*). Furthermore, the extinction coefficients of these peaks are similar to those of natural globins and are significantly larger than those of the designed four-helix bundle protein (*65*). These properties of heme-DG1 are analogous to those of the natural Mb H64V/V68H double mutant (*42*); the mutant has a low-spin six-coordinated heme iron with bis-His axial ligands in both the ferric and ferrous forms, and the ferrous Mb mutant is rapidly oxidized by O_2 and easily forms a stable complex with CO.

Resonance Raman spectroscopy has been used for determination of the redox state, spin state, and coordination structure of the heme (*66*, *67*). Figure 10A shows the resonance Raman spectra of the ferric (a) and ferrous (b) forms of heme-DG1 in the 1200−1700 cm^{-1} region. The ν_3 band, which is the marker band representing the coordination state and the size of heme, of the ferric form was observed

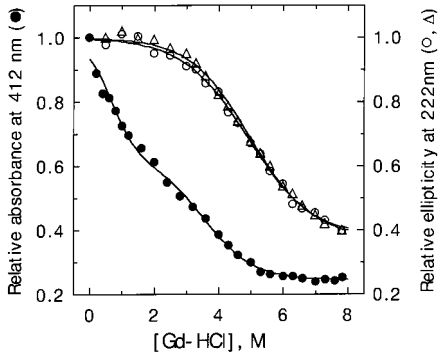

FIGURE 11: Denaturation curves of DG1. The proteins were quantitatively denatured by titration with Gd-HCl in 10 mM Tris-HCl (pH 8.0) and 200 mM NaCl at protein concentrations of 5 μM. Secondary structures of DG1 (open circles) and heme-DG1 (open triangles) were monitored with CD signal intensity at 222 nm. The association of heme in heme-DG1 (closed circles) was monitored with the Soret absorption peak at 412 nm.

at 1506 cm^{-1}. This frequency of the ν_3 band indicates that the ferric heme iron of heme-DG1 is in a six-coordinated low-spin state. On the other hand, in the case of the ferrous form, two ν_3 bands were observed, one at 1471 cm^{-1} and the other at 1494 cm^{-1}. The 1471-cm^{-1} band, which is the same frequency as the ν_3 band of deoxy-Mb, indicates that the heme iron is in a five-coordinated high-spin state, while the 1494-cm^{-1} band corresponds to a six-coordinated low-spin state. Thus, in the ferrous form of heme-DG1, there is coexistence of both five- and six-coordinated hemes, which is consistent with the inference from the UV−visible absorption spectra (Figure 9B).

Figure 10B shows the resonance Raman spectra of the CO-bound form of heme-DG1 in the regions of 200−600 and 1900−2100 cm^{-1}. An Fe−CO stretching (ν_{Fe-CO}) frequency band was observed at 491 cm^{-1}, which is lower than ν_{Fe-CO} = 507 cm^{-1} of the CO-bound Mb. On the other hand, a C−O stretching (ν_{CO}) frequency band was observed at 1966 cm^{-1}. It is known that the ν_{Fe-CO} and ν_{CO} frequencies of the CO adducts of heme proteins vary greatly according to the environments of heme pockets. Since these Raman bands of the CO-bound heme-DG1 are clearly single bands, it is considered that the heme pocket forms a single environment in the CO-bound form of heme-DG1. The relationship between the ν_{Fe-CO} and ν_{CO} frequencies of CO-bound heme proteins is called an inverse relationship which was first demonstrated by Yu et al. (*71*). The ν_{Fe-CO} value of heme-DG1 was plotted against the ν_{CO} in Figure 10C, where data from sperm whale Mb and its mutants were also plotted. The heme-DG1 data hold the same inverse relationship as was observed for the Mb mutants. The inverse relationship between ν_{Fe-CO} and ν_{CO} categorizes heme compounds into three groups by trans ligands: a weak trans ligand group, imidazole as the trans ligand (for example, mutant Mbs in Figure 10C), and a strong trans ligand group (for example, P-450_{cam}). Therefore, the trans ligand of CO-bound heme-DG1 can be attributed to imidazole of histidine, which agrees with the visible absorption data.

Stability. The quantitative disruption of the secondary structure of DG1 with the denaturant guanidine hydrochloride (Gd-HCl) was measured by monitoring the CD signal intensity at 222 nm to estimate the thermodynamic stability (Figure 11). The dependence of the mean residue ellipticity

Table 3: Thermodynamic Parameters for Gd-HCl Denaturation

	apo-form		holo-form	
	$\Delta G°_{app}$ (kcal mol^{-1})	m (kcal mol^{-1} M^{-1})	$\Delta G°_{app}$ (kcal mol^{-1})	m (kcal mol^{-1} M^{-1})
DG				
α-helix	2.8 ± 0.1	0.57 ± 0.01	3.1 ± 0.1	0.63 ± 0.02
bound heme	−	−	0.9 ± 0.1	1.4 ± 0.1
			2.7 ± 0.2	0.77 ± 0.07
Mb				
α-helix	3.0 ± 0.3	3.0 ± 0.3	7.1 ± 0.3	4.6 ± 0.2
bound heme	−	−	6.9 ± 0.5	4.5 ± 0.3

on the Gd-HCl concentration yielded a rather broad denaturation curve with a higher transition midpoint (C_m = 4.9 M) compared with that of native apoMb (C_m = 1.0 M). $\Delta G°_{app}$, the apparent free energy change from the folded state to the unfolded state in the absence of denaturant, was estimated as 2.8 kcal/mol with an m, the dependence of the free energy on the denaturant concentration, of 0.57 kcal/mol by assuming a simple two-state folding−unfolding transition. The association with heme slightly increased in resistance to the denaturation, and the heme-bound form showed a $\Delta G°_{app}$ of 3.1 kcal/mol with an m of 0.63 kcal/mol. These values are compared with those for horse Mb, which were determined under the same experimental conditions, in Table 3. The $\Delta G°_{app}$ values for DG1 are comparable with those for Mb whereas the m values for DG1 are much smaller than those for Mb, indicating that the unfolding−folding transition of DG1 is less cooperative. This agrees with NMR spectra of DG1 with and without heme (data not shown), which showed broader lines, poor chemical shift dispersion, and thus lower structural specificity than that of native Mbs (see Discussion).

The denaturation of heme-DG1 was also measured by monitoring the Soret absorption peak at 412 nm as a function of Gd-HCl concentration (Figure 11, Table 3). At least two phases were detected in the denaturation curve. The first denaturation phase ($\Delta G°_{app}$ = 0.95 kcal/mol and m = 1.4 kcal/mol) mainly occurred while the secondary structure was intact, and the second phase ($\Delta G°_{app}$ = 2.7 kcal/mol, m = 0.77 kcal/mol) approximately agrees with the denaturation curve of the secondary structure. In the spectral changes for the first phase, the absorbance intensity decreased over the whole wavelength range without clear isosbestic points (data not shown). This indicates that the heme is still associated with the protein and that the associated state was disturbed by the denaturant with the helices preserved. On the other hand, the spectral change in the second phase clearly shows the transition from bound heme into a free form. Such inconsistency between the disruption of the secondary structure and heme binding is in contrast with the good agreement observed in native Mbs (e.g., ref *58*; see also Discussion).

DISCUSSION

The main purpose of this study was to validate the 3D−1D compatibility function for de novo design of sizable globular proteins in the simple algorithm presented here. We computationally made a convergent artificial sequence starting from the native sequence of sperm whale Mb to find the best fit amino acids at each site for the backbone 3D structure (Figures 1−3). This operation was automatically performed without any bias such as composition constraint (*72*−*74*). The converged sequence thus obtained is self-consistent; i.e., the amino acids at every site are positioned first in its 3D profile (Figure 2B). The self-consistent sequence (SCS1) gives a molecular mass 12% larger than that of the target natural Mb, and therefore many atomic bumps among residues are observed in the 3D model. Since these bumps apparently impede the folding into the target structure, they were removed by replacing bumped bulky residues into smaller ones according to the 3D profile of SCS1. The 3D model with the "trimmed" DG1 sequence shows no atomic bumps within the molecule, a reasonable distribution of hydrophilic and hydrophobic residues, and more hydrogen bonds than those detected in the target Mb (Figure 5). The overpacking observed in the model of SCS1 is due to a preference of the compatibility functions for larger residues, which may be overcome by introduction of the volume constraint in sequence selection (*19*, *75*) and/or development of novel functions specified for the sequence design in the following design rounds.

SCS1 and the DG1 sequence preserve the amino acid residues that are highly conserved in the natural globin family at several sites (Figure 4), and both have 26% sequence identity with the native Mb (Table 1). Weak similarity of the artificial sequences was detected with many natural globins besides the sperm whale Mb, among which the squirrel monkey Mb has the highest sequence identity (29%) with DG1. The DG1 sequence also shows significant compatibility with many natural globin and phycocyanin (*76*) structures when it was threaded through these structures (results not shown). De novo self-consistent sequences have been obtained from the artificial initial sequences predetermined by the "first minimalist principles", i.e., helices comprised of binary patterns of Lys, Glu, and Leu along the sequence, and loops of all Gly, and also by a nonpairwise, one-body function consisting of local-conformation and hydration potentials using random sequences as the initial sequence. These self-consistent sequences (not shown) also preserve the residues that are well-conserved in natural globins and show more than 60% sequence identity to each other and to the DG1 sequence. Thus, the residues important for the target fold were not determined by the starting sequence but by the target structure. The experimental validation of new methodologies without using any natural sequences is the next essential subject to our research goal.

The synthetic gene encoding the DG1 sequence has been efficiently expressed in *Escherichia coli*, and the protein was generated in the soluble fraction of recombinant cells (Figure 6). Hydrodynamic analysis of purified DG1 with size-exclusion chromatography (Figure 7) and analytical ultracentrifugation (data not shown) showed that DG1 was solubilized in the monomeric form and that the overall molecular shape of DG1 resembles that of natural Mb. The helical contents of DG1 with and without heme were estimated at about 60%, which are almost the same as that of natural apoMb, based on CD spectroscopy. These results agree well with the quantitative estimation of the molecular shape of DG1 by SAXS analysis.

In the experiments of solution X-ray scattering, the aggregation state is crucial to the final interpretation of data and can be monitored with the forward scattering intensity,

$I(0)/C$, which is proportional to molecular mass. $I(0)/C = 330$ for DG1 was larger than $I(0)/C = 280$ for horse metMb by 18% (Table 2). Since the M_r value of apoDG1 (18.6 kDa) is larger than the value of metMb (17.6 kDa) by 5.7%, the net difference in $I(0)/C$ was 12%. The difference in surface judging from R_g, $(20.6/17.4)^2$, was approximately 40%, which could cause further hydration. Thus, the increase in $I(0)/C$ is due to the increase in hydration accompanying the increase in R_g. Therefore, the obtained data indicate that DG1 is in a monomeric state under the experimental conditions. The SAXS analysis of DG1 also showed that the R_g values are 19.5 and 20.6 Å with and without heme, respectively, which are close to the value of 19.7 Å for apoMb and larger than the value of 17.4 Å for holoMb (Table 2). ApoMb is known to preserve an almost intact holoMb backbone structure except for the disturbed region around the F helix based on NMR analyses (*77*), which should be responsible for the larger R_g value observed in SAXS analyses. The estimations of the R_g values of DG1 are roughly consistent with its helix content, 61%, which is almost the same as that of apoMb, according to the linear relationship between R_g and the helix contents observed in various conformational states of natural Mb (*61*). Interestingly, however, the heme binding of DG1 induced a reduction in the dimensions by 1 Å with the helical content almost unchanged. Thus, a rearrangement of the helix position rather than a change in secondary structure occurred with the heme binding in DG1. In conclusion, DG1 folds into a monomeric, Mb-like structure with a slightly expanded shape and has the ability to spatially rearrange helices to form a more compact structure upon the binding of heme.

Almost all natural globins preserve the two histidine residues, the so-called proximal (E7) and distal (F8) His in the E and F helices, respectively, which play essential roles in binding and functionalizing heme (e.g., ref *37*). Our previous analysis of natural globins using the 3D−1D compatibility method (*47*) showed that, in contrast with the other highly conserved residues, these two His residues are poorly suited to the sites. Thus, we concluded that they are not conserved by requirements for maintaining the globin folds but by those for the heme-related functions. In our sequence design, these two histidine residues were fixed during the calculation along with preservation of a space for heme insertion using a repulsive function (*45*) since the designed globin was intended to bind heme and hopefully to be functional. Synthesized DG1 successfully binds a single heme per protein molecule, and the heme-bound form shows well-defined spectroscopic features (Figures 9 and 10). The visible optical absorption and resonance Raman spectra of the ferric form are characteristic of low-spin six-coordinated heme iron with two His residues as the axial ligands, and the spectra of the ferrous form suggest a mixture of a high-spin five-coordinated state and a low-spin six-coordinated state. The mutagenesis experiment (Figure 9A) confirmed that His93 coordinates to the heme iron as one of the axial ligands, which is consistent with the conclusion that DG1 folds into globin-like tertiary structure. His64 is probably another axial ligand because this site is closest to, but sufficiently separated from, position 94 on the sequence for heme ligation (see ref *78*). The ferrous heme in DG1 forms a stable complex with CO, which gives optical absorption and resonance Raman spectra that are consistent with those of the heme−CO complex in natural Mbs (Figures 9B and 10B,C). On the other hand, the heme in DG1 is quickly oxidized by O_2, and the oxy form cannot be detected under the experimental conditions. This is in contrast with natural globins in which the heme iron forms a stable complex with O_2. The failure to realize this biologically essential function in DG1 is not surprising because the heme-binding pocket of DG1 was designed by just making a space for heme inside the protein using the repulsive function. Also, this failure may originate from the higher conformational diversity of side-chain structures compared with those of native globins (see below). It should be partly due to the same reasons that the dissociation constants for heme-DG1 (80 nM) and also for designed four-helix bundle heme proteins (*33*, *34*, *78*) are at least 10^5 larger than those of native Mbs (*58*).

In the denaturation experiments with Gd-HCl, the secondary structure of DG1 with and without heme was much more tolerant to the denaturant and exhibited a lower cooperativity in the folding−unfolding transition than that of natural Mbs (Figure 11, Table 3). The heme binding is less stable than the secondary structure in DG1, whereas they are quite coincident in natural Mbs. These results shed light on the general relationship between structures and functions of natural proteins; i.e., both the secondary and tertiary structures of natural proteins form just to maintain their functions during molecular evolution, whereas DG1 displays an excess stability of the structures. The low cooperativity in the denaturation transition of DG1 agrees with the ^{1}H NMR measurements; the NMR spectra (not shown) showed characteristic features of α-helical proteins; i.e., the range of the amide proton chemical shifts was between 6 and 9 ppm and the absence of Cα−H resonances at 5−6 ppm, with broader lines and poor chemical shift dispersion. Thus, DG1 adopts many different thermally accessible conformations that slowly interconvert on the proton chemical shift time scale, while it preserves the overall shape and the α-helical content that are similar to those of the target structure (Table 2). These properties of DG1 are homologous to those of artificial proteins previously designed (*14*, *26*, *34*, *79*), and they are characteristic of a "gemisch state", which is similar to, but distinguishable from, a molten globule state of natural proteins (*80*).

The lower structural specificity of DG1 at the high-resolution aspect is apparently caused by the significant preference of Leu in the sequence selection for hydrophobic sites in the present design. A total of 34 Leu residues were selected in the DG1 sequence, whereas only 18 Leu residues are in the sperm whale Mb. In the X-ray crystallographic structures of nine natural globins examined (1ash, 1eca, 1hlb, 1mbd, 1pbx, 2fal, 2gdm, 2hbg, 3sdh), Ile and Val residues (nine Ile and eight Val in 1mbd) are distributed dispersively in the hydrophobic region and constitute the core in combination with Leu and other hydrophobic residues. On the other hand, DG1 has only a single Ile and no Val (Table 1), and most of the hydrophobic core is composed of Leu in the 3D model. In a hydrophobic core that contains too many leucines, residues may not be distinguishable from each other and may not be able to find their own specific positions as one can imagine in the analogous situation of a jigsaw puzzle with many pieces of identical shape. Furthermore, recent analyses of natural and mutant globular proteins (*81*, *82*) have shown that in α-helices the β-branched side chains of Ile and Val contribute to the structural uniqueness by

decreasing residue conformational entropy cooperatively with neighboring bulky side chains of Trp, Phe, and Tyr. Thus, the redesign of DG1 will be based on the working hypothesis of the roles of the hydrophobic residues with different side chain shapes on protein folding; proper arrangements of these hydrophobic residues by considering the explicit side chain conformations are required to achieve the structural specificity and, ultimately, the functions with native-like qualities.

ACKNOWLEDGMENT

We thank Ms. Yasue Ichikawa (Biodesign DNA sequencing facility, RIKEN) for DNA sequencing, and Drs. Naoshi Dohmae, Masao Chijimatsu, and Koji Takio (Division of Biomolecular Characterization, RIKEN) for mass spectroscopic, N-terminal sequence, and ultracentrifugation analyses of proteins. Y.I. thanks Ms. Anna Ishii for preparing computer graphics.

REFERENCES

1. Desjarlais, J. R., and Handel, T. M. (1995) *Curr. Opin. Biotechnol. 6*, 460−466.
2. Bryson, J. W., Betz, S. F., Lu, H. S., Suich, D. J., Zhou, H. X., O'Neil, K. T., and DeGrado, W. F. (1995) *Science 270*, 935−941.
3. Cordes, M. H. J., Davidson, A. R., and Sauer, R. T. (1996) *Curr. Opin. Struct. Biol. 6*, 3−10.
4. Drexler, K. E. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 5275−5278.
5. Ulmer, K. M. (1983) *Science 219*, 666−671.
6. Ho, S. P., and DeGrado, W. F. (1987) *J. Am. Chem. Soc. 109*, 6751−6758.
7. Regan, L., and DeGrado, W. F. (1988) *Science 241*, 976−978.
8. Hecht, M. H., Richardson, J. S., Richardson, D. C., and Ogden, R. C. (1990) *Science 249*, 884−891.
9. Gibney, B. R., Rabanal, F., Skalicky, J. J., Wand, A. J., and Dutton, P. L. (1997) *J. Am. Chem. Soc. 119*, 2323−2324.
10. Betz, S. F., Liebman, P. A., and DeGrado, W. F. (1997) *Biochemistry 36*, 2450−2458.
11. Schafmeister, C. E., LaPorte, S. L., Miercke, L. J. W., and Stroud. R. M. (1997) *Nat. Struct. Biol. 4*, 1039−1046.
12. Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S., and Richardson, D. C. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 8747−8751.
13. Yan, Y., and Erickson, B. W. (1994) *Protein Sci. 3*, 1069−1073.
14. Tanaka, T., Kimura, H., Hayashi, M., Fujiyoshi, Y., Fukuhara, K., and Nakamura, H. (1994) *Protein Sci. 3*, 419−427.
15. Hellinga, H. W., and Richards, F. M. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 5803−5807.
16. Desjarlais, J. R., and Handel, T. M. (1995) *Protein Sci. 4*, 2006−2018.
17. Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. (1997) *J. Mol. Biol. 273*, 789−796.
18. Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. (1997) *Protein Sci. 6*, 1333−1337.
19. Dahiyat, B. I., and Mayo, S. L. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 10172−10177.
20. Kono, H., Nishiyama, M., Tanokura, M., and Doi, J. (1998) *Protein Eng. 11*, 47−52.
21. Dahiyat, B. I., and Mayo, S. L. (1997) *Science 278*, 82−87.
22. Handel, T., and DeGrado, W. F. (1990) *J. Am. Chem. Soc. 112*, 6710−6711.
23. Lieberman, M., and Sasaki, T. (1991) *J. Am. Chem. Soc. 113*, 1470−1471.
24. Ghadiri, M. R., Soares, C., and Choi, C. (1992) *J. Am. Chem. Soc. 114*, 825−831.
25. Mihara, H., Nishino, N., and Fujimoto, T. (1992) *Chem. Lett. 1992*, 1805−1808.
26. Gibney, B. R., Johansson, J. S., Rabanal, F., Skalicky, J. J., Wand, A. J., and Dutton, P. L. (1997) *Biochemistry 36*, 2798−2806.
27. Mihara, H., and Takahashi, Y. (1997) *Curr. Opin. Struct. Biol. 7*, 501−507.
28. Rabanal, F., Degrado, W. F., and Dutton, P. L. (1996) *J. Am. Chem. Soc. 118*, 473−474.
29. Gibney, B. R., Mulholland, S. E., Rabanal, F., and Dutton, P. L. (1997) *Proc. Natl. Acad. Sci. U.S.A. 93*, 15041−15046.
30. Scott, M. L., and Biggins, J. (1997) *Protein Sci. 6*, 340−346.
31. Pinto, A. L., Hellinga, H. W., and Caradonna, J. P. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 5562−5567.
32. Coldren, C. D., Hellinga, H. W., and Caradonna, J. P. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 6635−6640.
33. Robertson, D. E., Farid, R. S., Moser, C. C., Mulholland, S. E., Pidikit, R., Lear, J. D., Wand, A. J., DeGrado, W. F., and Dutton, P. L. (1994) *Nature 368*, 425−432.
34. Choma, C. T., Lear, J. D., Nelson, M. J., Dutton, P. L., Robertson, D. E., and DeGrado, W. F. (1994) *J. Am. Chem. Soc. 116*, 856−865.
35. Fermi, G., and Perutz, M. F. (1981) *Atlas of Molecular Structures in Biology, Vol. 2, Haemoglobin and Myoglobin*, Clarendon Press, Oxford.
36. Dickerson, R. E., and Geis, I. (1983) *Hemoglobin: Structure, Function, Evolution and Pathology*, Benjamin/Cummings, Menlo Park, CA.
37. Springer, B. A., Sligar, S. G., Olson, J. S., and Phillips, G. N., Jr. (1994) *Chem. Rev. 94*, 699−714.
38. Lesk, A. M., and Chothia, C. (1980) *J. Mol. Biol. 136*, 225−270.
39. Hughson, F. M., Barrick, D., and Baldwin, R. L. (1991) *Biochemistry 30*, 4113−4118.
40. Brantley, R. E., Smerdon, S. J., Wilkinson, A. J., Singleton, E. W., and Olson, J. S. (1993) *J. Biol. Chem. 268*, 6995−7010.
41. Hargrove, M. S., Krzywda, S., Wilkinson, A. J., Dou, Y., Ikeda-Saito, M., and Olson, J. S. (1994) *Biochemistry 33*, 11767−11775.
42. Dou, Y., Admiraal, S. J., Ikeda-Saito, M., Krzywda, S., Wilkinson, A. J., Li, T., Olson, J. S., Prince, R. C., Pickering, I. J., and George, G. N. (1995) *J. Biol. Chem. 270*, 15993−16001.
43. Shiro, Y., Iwata, T., Makino, R., Fujii, M., Isogai, Y., and Iizuka, T. (1993) *J. Biol. Chem. 268*, 19983−19990.
44. Wakasugi, K., Ishimori, K., Imai, K., Wada, Y., and Morishima, I. (1994) *J. Biol. Chem. 269*, 18750−18756.
45. Ota, M., and Nishikawa, K. (1997) *Protein Eng. 10*, 339−351.
46. Ota, M., Kanaya, S., and Nishikawa, K. (1995) *J. Mol. Biol. 248*, 733−738.
47. Ota, M., Isogai, Y., and Nishikawa, K. (1997) *FEBS Lett. 415*, 129−133.
48. Ota, M., and Nishikawa, K. (1999) *Protein Sci.* (in press).
49. Matsuo, Y., and Nishikawa, K., (1994) *Protein Sci. 3*, 2055−2063.
50. Phillips, S. E. (1980) *J. Mol. Biol. 142*, 531−554.
51. Antonini, E., and Brunori, M. (1971) *Hemoglobin and Myoglobin in Their Reactions with Ligands*, American Elsevier, New York.
52. Price, N. C., and Dwek, R. A. (1979) *Principles and Problems in Physical Chemistry for Biochemists*, Second ed., Oxford University Press, Oxford.
53. Yamamoto, M., Fujisawa, T., Nakasaoko, M., Tanaka, T., Uruga, T., Kimura, H., Yamaoka, H., Inoue, Y., Iwasaki, H., Ishikawa, T., Kitamura, H., and Ueki, T. (1995) *Rev. Sci. Instrum. 66*, 1833−1835.
54. Fujisawa, T., Yagi, Y., Inoue, Y., Oka, T., Iwamoto, H., and Ueki, T. (1997) *Annual Report of SPring-8 238*, JASRI, Japan.
55. Amemiya, Y., Ito, K., Yagi, Y., Asano, Y., Wakabayashi, K., Ueki, T., and Endo, T. (1995) *Rev. Sci. Instrum. 66*, 2290−2294.

56. Glatter, O., and Kratky, O. (1982) *Small-angle X-ray scattering*, Academic Press, New York.
57. Hughson, F. M., and Baldwin, R. L. (1989) *Biochemistry 28*, 4415−4422.
58. Hargrove, M. S., and Olson, J. S. (1996) *Biochemistry 35*, 11310−11318.
59. Go, M. (1981) *Nature 291*, 90−92.
60. Nishii, I., Kataoka, M., Tokunaga, F., and Goto, Y. (1994) *Biochemistry 33*, 4903−4909.
61. Kataoka, M., Nishii, I., Fujisawa, T., Ueki, T., Tokunaga, F., and Goto, Y. (1995) *J. Mol. Biol. 249*, 215−228.
62. Eliezer, D., Jennings, P. A., Wright, P. E., Doniach, S., Hodgson, K. O., and Tsuruta, H. (1995) *Science 270*, 487−489.
63. Pace, C. N., Shirley, B. A., and Thompson, J. A. (1989) in *Protein Structure: A Practical Approach* (Creighton, T. E., Ed.) pp 311−330, IRL, Oxford.
64. Dumont, M. E., Corin, A. F., and Campbell, G. A. (1994) *Biochemistry 33*, 7368−7378.
65. Gibney, B. R., Rabanal, F., Reddy, K. S., and Dutton, P. L. (1998) *Biochemistry 37*, 4635−4643.
66. Spiro, T. G., and Li, X.-Y. (1988) in *Biological Applications of Raman Spectroscopy* (Spiro, T. G., Ed.) Vol. 3, pp 1−37, Wiley, New York.
67. Kitagawa, T. (1988) in *Biological Applications of Raman Spectroscopy* (Spiro, T. G., Ed.) Vol. 3, pp 97−131, Wiley, New York.
68. Ray, G. B., Li, X.-Y., Ibers, J. A., Sessler, J. L., and Spiro, T. G. (1994) *J. Am. Chem. Soc. 116*, 162−176.
69. Li, T., Quillin, M. L., Phillips, G. N., Jr., and Olson, J. S. (1994) *Biochemistry 33*, 1433−1446.
70. Nakashima, S., Kitagawa, T., and Olson, J. S. (1998) *Chem. Phys. 228*, 323−336.
71. Yu, N.-T., Kerr, E. A., Ward, B., and Chang, C. K. (1983) *Biochemistry 22*, 4534−4540.
72. Shakhnovich, E. I., and Gutin, A. M. (1993) *Protein Eng. 6*, 793−800.
73. Jones, D. T. (1994) *Protein Sci. 3*, 567−574.
74. Godzik, A. (1995) *Protein Eng. 8*, 409−416.
75. Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1997) *Protein Sci. 6*, 1167−1178.
76. Holm, L., and Sander, C. (1993) *FEBS Lett. 315*, 301−306.
77. Eliezer, D., and Wright, P. E. (1996) *J. Mol. Biol. 263*, 531−538.
78. Rojas, N. R. L., Kamtekar, S., Simons, C. T., Mclean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S., and Hecht, M. H. (1997) *Protein Sci. 6*, 2512−2524.
79. Handel, T. M., Williams, S. A., and Degrado, W. F. (1993) *Science 261*, 879−885.
80. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995) *Protein Sci. 4*, 561−602.
81. Nakamura, H., Tanimura, R., and Kidera, A. (1996) *Proc. Jpn. Acad., Ser. B 72B*, 149−152.
82. Furukawa, K., Oda, M., and Nakamura, H. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 13583−13588.

BI983006Y