Structural requirement of highly-conserved residues in globins

Motonori Ota[a,*], Yasuhiro Isogai[b], Ken Nishikawa[a]

[a] National Institute of Genetics, Yata, Mishima, Shizuoka 411, Japan
[b] The Institute of Physical and Chemical Research (RIKEN), Hirosawa, Wako, Saitama 351-01, Japan

Received 8 July 1997; revised version received 8 August 1997

Abstract **Globins have remarkable sequence diversity, and yet maintain a common fold. In spite of the diversity, there are highly-conserved residues at several sites. The conserved residues were examined in terms of the structural stability, by employing the pseudo-energy functions of the structure/sequence compatibility method. The fitness of each residue type to the structural environment was evaluated at seven highly-conserved sites: the Leu (at the B10 site), Phe (CD1), and Leu (F4) residues were found to fit their respective sites due to hydrophobic interactions; Pro (C2) stabilizes the N-terminal edge of an α-helical structure; and Phe (CD4) is stabilized by backbone hydrogen-bonding to Phe (CD1). On the other hand, the other two residues, His (E7) and His (F8), are poorly suited to the sites from a structural viewpoint, suggesting that their conservation clearly results from a heme-related functional requirement. The invariant Phe residue (CD1) has been suggested to be important for supporting the heme. The present analysis revealed that this residue is also well suited to the site in terms of energy.**

Key words: Structure/sequence compatibility evaluation; Pseudo-energy potential; Sequence conservation; Protein structure; Protein function

1. Introduction

The design of artificial protein sequences that fold into a desired structure is the main goal of protein engineering [1]. A first step in this direction is to elucidate the reasons why specific amino acids are adopted at certain sites of some folds, as an inductive way to understand the principles of protein architecture and sequence determination. There are a number of possible reasons for the sequence of a native protein to be adopted by a particular fold. Among them, functional and structural requirements are two major determinants. The functional requirement is rather easy to detect, as mutations introduced into such sites would cause a decrease or loss of activity. On the other hand, the structural requirement is much more difficult to analyze, because protein structures can often accommodate various amino acid substitutions [1,2]. It is widely known that both requirements are often, but not always, incompatible, i.e. a mutation introduced into the active site decreases the activity, but increases the stability, and vice versa [3]. In the present study, we addressed whether the conservation of amino acids observed within a protein family can be explained in terms of the structural stability.

The globin family was chosen as the subject, since several X-ray structures as well as a large amount of sequence data are available. The characteristic features of globin sequences have been extensively studied by Lesk and Chothia [4], as well as by Bashford et al. [5]. In the latter analysis, 226 globin sequences were aligned against site numbers uniquely indexed to the globin family, and the number of appearances of each amino acid type at each site were tabulated. According to Table 3 of [5], there are seven sites at which a specific amino acid appears more than 200 times (greater than 90%): Leu (B10), Pro (C2), Phe (CD1), Phe (CD4), His (E7), Leu (F4), and His (F8). This conservation may be partly attributable to a role in heme-binding [4]. Recently, Hargrove et al. [6,7] reported that substitutions of Phe (CD1) caused a decrease not only in heme-binding affinity but also in apomyoglobin stability. The paper by Hargrove and Olson [8] reported that the substitution of Thr for Val at E11 (not a query site in this paper) increased the heme-binding affinity, but decreased the apomyoglobin stability. A qualitative correlation between heme-binding affinity and holomyoglobin stability was observed, and thus they concluded that the stability of the holomyoglobin resulted from the heme interaction, rather than the apomyoglobin stability [8], indicating that the heme contributes to both the structural stability and the functional center of the globins.

Recently, we developed a computational method [9] to analyze protein structural stability, using the three-dimensional (3D) profile [10,11]. The compatibility of 20 amino acids to a given site of a structure is evaluated with pseudo-energy potentials [11], and the computation shows which type of amino acid is best fitted to a given site. Similar methods, based on structure/sequence compatibility [12], for estimating the structural stability have been proposed by other research groups [13–15]. Our method was applied to a number of mutants of ribonuclease HI [9]: the energy difference between a mutant and the wild-type protein (ΔΔG) was estimated and compared with the experimentally determined difference in the melting temperature, ΔTm. A significant correlation was observed between the stabilities determined by the computations and the experiments. The computations also revealed the stabilizing mechanism of several mutants.

The same method can be applied to the analysis of conserved amino acid residues of a protein as well, instead of mutated residues. We focused on the above mentioned seven conserved residue sites of the globins, and examined whether the amino acid conserved at a given site is optimal in terms of the structural stability. Concerning the dual roles of heme mentioned above, our analyses were performed on the globin part of the holo-form (excluding the interaction with heme), to distinguish the direct contribution of the conservative residues to the stability from the indirect contribution through the heme-binding. If an amino acid residue is stabilizing in this situation, then we could say that the residue fulfils the structural requirement in a strict sense.

*Corresponding author. Fax: +81 (559) 81-6889.
E-mail: mota@genes.nig.ac.jp

PII S0014-5793(97)01110-1

Fig. 1. a–g: Transfer free energy from the random environmental state to the folded state at the site environment for each residue type (ΔG). The ΔG was individually estimated for nine globin structures with dissimilar sequences, and was averaged over them at each query site. Residue types with lower ΔG values are advantageous to the site. Black and shaded bars represent the conserved residue at the site and the other residues, respectively. The order of the residue types along the horizontal axis is according to the hydrophobicity measured by the hydration function employed here [11,21]. h–m: Contribution of each energy term decomposed from the total ΔG. Each figure, representing the energy of the conserved residues, corresponds to its adjacent one on the left side (a–g). SCP, Hyd, Hbd, and Loc indicate the side-chain packing, hydration, hydrogen-bonding, and local conformational functions, respectively. The main contributor is indicated by the black bar.

→

2. Materials and methods

Globin structures with mutually dissimilar sequences (less than 25% sequence identity) were taken from the PDB_SELECT database released in November, 1996 [16]. The nine globin structures shown in Table 1 were selected. The structural alignment was made in accordance with the FSSP database [17]. The residue sites of a query structure, along with the native amino acid type at each site, are shown in Table 1.

The 3D profile [10] expresses the fitness of the 20 amino acids to each site of a protein structure. By introducing a slight modification into the definition, the profile can indicate the transfer energy (ΔG) from the denatured state to the folded state for an amino acid at a site [9]. This modified type of 3D profile is calculated from the X-ray coordinates of a protein using empirically derived pseudo-energy functions. Four energy functions are considered: side-chain packing, hydration, backbone hydrogen-bonding, and local conformation [11]. The side-chain packing and the backbone hydrogen-bonding functions are the pairwise terms, while the others are determined by the match between a single amino acid and the structural environment. The heme moiety was ignored in the calculation, except that the repulsive energy [11] was considered between the amino acid side-chains and the heme.

After the 3D profiles were obtained for each of the nine globin structures, an 'average 3D profile' was synthesized by taking the average over the nine profiles at those residue sites judged as completely available from their structural alignment.

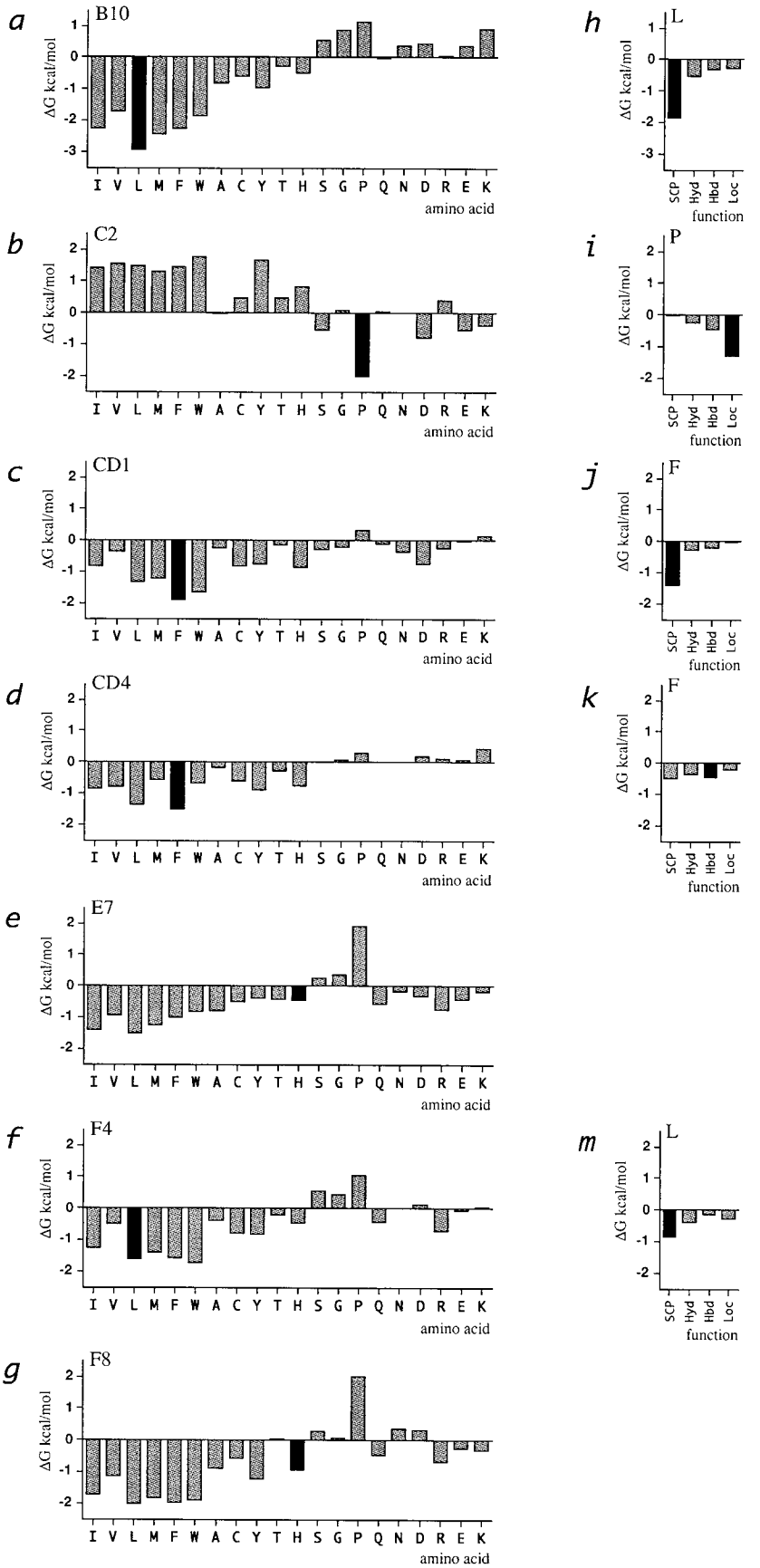
3. Results and discussion

The ΔG value tabulated in the average 3D profile is indicative of the free energy difference between the folded and unfolded states [9]. A large, negative value of ΔG is advantageous to the site environment in terms of the structural stability. The energy values of the 20 amino acids at the conservative residue sites are shown in Fig. 1a–g. The scores of the highly-conserved residues are represented by black bars, and the rest by shaded bars. At the B10, C2, CD1 and CD4 sites, the highly-conserved residues (i.e. Leu, Pro, Phe, and Phe, respectively) are the most advantageous among the 20 amino acids. These calculations were carried out by neglecting the heme moiety (see Section 2), and therefore they meet the requirement of the structural stability for the globin fold itself. In other words, the conservation of these residues is supported without the interaction with the heme group. The importance of Phe (CD1) for stability in the apomyoglobin refolding was suggested by Hargrove et al. [6]. Our calculation agrees with their experiments. At the F4 site, the highly-conserved residue is Leu, which occupies the second position of the fitness ranking next to Trp, and the energy difference between them is small. We could also consider Leu (F4) as satisfying the requirement of the structural stability. Thus, all of the highly-conserved sites of the globins, except for two (His at sites E7 and F8), fulfil the structural requirement. On the other hand, Fig. 1e and g clearly show that the two His residues are conserved only for the functional (heme-binding) reason, at the expense of the structural stability, in the apo-form [4]. His (F8) is directly coordinated to the heme iron atom and therefore is invariant throughout the known globin sequences. His (E7), located on the other side of the heme from His (F8), is involved in controlling the ligand-binding and the stabilization of the O_2-heme complex. This residue is conservative, but is not invariant (Table 1).

The results in Fig. 1 seem to clearly discriminate the conserved residues into either structural or functional types. In order to see whether our method works well in more general cases, the method was applied to the active site residues of typical enzymes. The results for six enzymes are shown in Table 2. The top example shows that one of the active site residues, His-12, of bovine ribonuclease A was ranked 6th among the 20 amino acids by the 3D profile calculation at this site, with an energy value of −2.2, while Phe had the lowest energy, −3.0 and Pro had the highest, +2.7. The rankings of the active site residues in Table 2 indicate a similar tendency toward those of the two His residues in the globins, rather than those of the other five residues. It is particularly interesting to see that each enzyme contains at least one active site residue that is ranked relatively poorly, at lower than 10th place. These rankings clearly suggest that the residues are indispensable for the enzymatic activity, but are not fitted to the sites in terms of structural stability. On the other hand, some of the residues (e.g. His-119 of ribonuclease A, Asp-70 of ribonuclease HI, and Asp-20 of T4 phage lysozyme) are more or less fitted to the structural environment. These data

Table 1
List of nine globins of known structure, and their well-conserved sites used in the analysis

PDB		Residue site (amino acid type)						
Code	Protein	B10 (L)	C2 (P)	CD1 (F)	CD4 (F)	E7 (H)	F4 (L)	F8 (H)
1ash	hemoglobin (E. coli)	31 (Y)	39 (P)	45 (F)	48 (R)	65 (Q)	93 (L)	97 (H)
1eca	erythrocruorin	24 (L)	32 (P)	38 (F)	41 (F)	58 (H)	83 (F)	87 (H)
1hlb	hemoglobin (sea cucumber)	39 (F)	47 (P)	53 (F)	56 (M)	73 (H)	100 (L)	104 (H)
1mls	myoglobin (sperm whale)	30 (L)	38 (P)	44 (F)	47 (F)	65 (H)	90 (L)	94 (H)
1pbxA	hemoglobin (Antarctic fish)	29 (L)	37 (P)	43 (F)	46 (W)	59 (H)	84 (L)	88 (H)
2fal	myoglobin (sea hare)	29 (L)	37 (P)	43 (F)	46 (F)	63 (V)	91 (F)	95 (H)
2gdm	leghemoglobin	30 (F)	38 (P)	44 (F)	47 (L)	63 (H)	93 (L)	97 (H)
2hbg	hemoglobin (marine bloodworm	31 (L)	39 (P)	45 (F)	47 (F)	58 (L)	86 (V)	90 (H)
3sdhA	hemoglobin (ark clam)	36 (M)	44 (Q)	50 (F)	53 (L)	68 (H)	96 (F)	100 (H)


a
B10
b
C2
c
CD1
d
CD4
e
E7
f
F4
g
F8
h
L
i
P
j
F
k
F
m
L
ΔG kcal/mol
I V L M F W A C Y T H S G P Q N D R E K
amino acid
SCP Hyd Hbd Loc
function

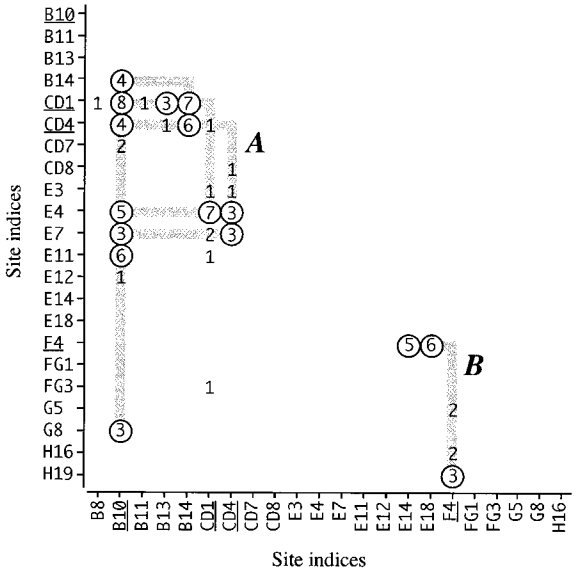


Fig. 2. Counterparts of the side-chain packing interaction of B10, CD1, CD4, and F4 (underlined). The interactions evaluated below −0.3 kcal/mol are counted, and the number of observations among the nine globins is added up at each site pair. Interactions observed three times or more are circled and connected. Two individual networks (A and B) are shown.

led us to classify the conserved residues into three categories: conserved for functional but not structural reasons, conserved for structural but not functional reasons, and conserved for both functional and structural reasons. Of these, the first two types may be more intriguing.

The 3D profile calculation for the globins is more reliable than that for the enzymes in Table 2, because the average profile was used for the former. Returning to the globins, we analyzed the contributions to the structural stability more closely. The stabilization energy, ΔG, was decomposed into four types of energy components. The results are shown in Fig. 1h–m. At the B10, CD1, and F4 sites, the structural stability is mainly derived from the side-chain packing interaction, whereas at the C2 and CD4 sites, the local conformational term and the hydrogen-bonding term are the main contributors to the total ΔG, respectively. The conservation of Pro at the C2 site is easy to understand, since the local conformational function is determined by a single residue at the site. In the function, we consider the fitness of a single residue to a local conformation over five successive sites around C2. The pentapeptide conformation around C2 is expressed as *ae***g**gg, with the C2 site at the center with the conformation *g* (bold), and where *a* is α-helical, and *e* and *g* have the dihedral angles of α- and β-structures, but are assigned as 'coil' by the DSSP program [18]. We considered the pentapeptide divided into three triplets, *ae***g**, *e***g**g, and **g**gg. The potential function was derived for each triplet and was then combined. In each triplet, Pro is the most plausible residue at the C2 site, with scores of −1.35, −1.1, and −0.6 for *ae***g**, *e***g**g, and **g**gg, respectively. As the first two contribute more than the last, Pro seems to be suited to the N-terminus of the C-helix. Lesk and Chothia [4] reported that the conservation of Pro at this site cannot be explained solely by the dihedral angles of the backbone. The present analysis indicates that the empirical potential function is able to explain the conservation of Pro at the C2 site, by taking the local conformations of the five residues into account. Hemoglobin variants on this site were known to increase oxygen affinity. Among them, Hb Linköping (Pro changed to Thr) destabilizes mildly in the propanol test [19,20], while Hb Chicago (Pro to Ser) does not show significant instability [21,22]. These results are consistent with Fig. 1b.

At the CD4 site, the backbone hydrogen-bonding (HB) energy is the dominant stabilization factor over the other energy terms (Fig. 1k). Indeed, the rankings of Phe among the 20 amino acids, evaluated by the individual functions, are 3, 4, 1, and 3 for side-chain packing, hydration, hydrogen-bonding and local conformational functions, respectively. Except for one structure, 2hbg, the backbone oxygen of CD4 binds the backbone nitrogen of CD1. In our parameter set, a backbone HB bond between two Phe residues, separated by three residues along a chain, is one of the most stable HB bonds among all pairs of X-Phe or Phe-X. Along with the fact that Phe (CD1) is invariant in all globin sequences, whereas Phe (CD4) is not (Table 2), the remarkable HB stabilization of CD4 is explainable because for the CD4 site, the counterpart of the HB bond is always Phe, while the same is not true for the CD1 site.

As pairwise interactions are the main contributors to the structural stability, except for Pro (C2), the counterpart of a residue pair would be important. At the B10, CD1, CD4, and F4 sites, the counterparts of the side-chain packing interaction are shown in Fig. 2. The sites with contributions to the side-chain packing term below −0.3 kcal/mol were counted in each

Table 2
Stability analysis of active site residues of six enzymes

Code	Protein	Active site	Amino acids in the order of fitness	Ranking	ΔG (lowest, highest)
3rn3	ribonuclease A	12 (H)	FMWYLHIVCRQAKENTGSDP	6	−2.2 (−3.0, 2.7)
3rn3	ribonuclease A	41 (K)	PVCYDIGFHKMSRNLATEWQ	10	−0.2 (−1.2, 0.7)
3rn3	ribonuclease A	119 (H)	VYWHRIKFCTQMLGSEANPD	4	−0.4 (−0.8, 1.1)
2rn2	ribonuclease HI	70 (D)	YWNQDPHSETFGAVCLIKRM	5	−0.5 (−1.4, 1.2)
2rn2	ribonuclease HI	124 (H)	PDSENGKTAQRHCVYMLFIW	12	0.5 (−1.0, 2.4)
1mbA	barnase	72 (E)	WYFRVMIHCLKNQEASTPDG	14	0.0 (−3.2, 0.8)
1mbA	barnase	101 (H)	GDKNSQETHAMRCPLYWVIF	9	0.7 (−1.6, 3.4)
1lyd	T4 lysozyme	11 (E)	YWLFCHAIMVNERTGSQKDP	12	−0.2 (−1.8, 3.0)
1lyd	T4 lysozyme	20 (D)	GCDSHNTVAKYQRIELWMFP	3	−1.0 (−1.1, 5.4)
1lz1	human lysozyme	35 (E)	LMIWFYVAHRQCKTNGESDP	17	0.7 (−2.0, 3.6)
1lz1	human lysozyme	53 (D)	FYVIHWTCKNQMRLPEGDSA	18	0.5 (−0.5, 0.6)
4ptp	β trypsin	40 (H)	PTSCNHADEGVKFQYLRIMW	6	−0.1 (−0.7, 0.9)
4ptp	β trypsin	84 (D)	LVIFYCTAMQRPKHWSGDEN	18	0.6 (−1.0, 0.8)
4ptp	β trypsin	177 (S)	IPFCVLYDQSNHATREKGWM	10	0.0 (−0.7, 0.9)

sample, and were added up for each interaction pair of the nine globins. If these interactions occur in at least three samples out of nine samples, then they are circled in Fig. 2. All of the circled sites constitute two individual networks: One consists of B10, CD1, CD4, B14, E4, E7, E11, and G8 (network A in Fig. 2), and the other consists of E14, E18, F4, and H19 (network B). Interestingly, the sites within the networks tend to have conserved amino acids. According to Table 3 of Bashford et al. [5], a single amino acid type is observed in more than 50% of 226 samples at these sites: Leu (at site B14, 125 proteins), Val (E4, 131), Val (E11, 198), Ala (E14, 138), Ala (E18,113), and Leu (H19, 167). Since almost all of these residues are hydrophobic, and all of the sites were assigned buried, according to Bashford et al. [5], they would be very important in building the hydrophobic core of the globin structure. It should be noted that His (F8) is not involved in these networks. Another functionally-conserved residue His (E7) is a part of network A, but almost all of the pairwise interactions for the E7 site, counted in Fig. 2, arose from non-conserved samples (2fal and 2hbg, see Table 1). For structural stability, hydrophobic residues would be preferred at this site (Fig. 1e). These interaction networks confirm that the structurally-conserved residues participate in the formation of the structural core in the globin portion. On the other hand, the two His residues (E7, F8) are functionally essential, and contribute to the stability only through the heme [8]. The same analyses were carried out for the apomyoglobin structure (1bvd), and we obtained results similar to these shown in Figs. 1 and 2. This confirms our conclusions about the roles of the two His, as well as the others.

Acknowledgements: We are very grateful to Dr. Hideki Morimoto for the comments about abnormal hemoglobins.

References

[1] Cordes, M.H.J., Davidson, A.R. and Sauer, R.T. (1996) Curr. Opin. Struct. Biol. 6, 3–10.
[2] Chothia, C. and Gerstein, M. (1997) Nature 385, 579–581.
[3] Meiering, E.M., Serrano, L. and Fersht, A.R. (1992) J. Mol. Biol. 225, 585–589.
[4] Lesk, A.M. and Chothia, C. (1980) J. Mol. Biol. 136, 225–270.
[5] Bashford, D., Chothia, C. and Lesk, A.M. (1987) J. Mol. Biol. 196, 199–216.
[6] Hargrove, M.S., Krzywda, S., Wilkinson, A.J., Dou, Y., Ikeda-Saito, M. and Olson, J.S. (1994) Biochemistry 33, 11767–11775.
[7] Hargrove, M.S., Wilkinson, A.J. and Olson, J.S. (1996) Biochemistry 35, 11300–11309.
[8] Hargrove, M.S. and Olson, J.S. (1996) Biochemistry 35, 11310–11318.
[9] Ota, M., Kanaya, S. and Nishikawa, K. (1995) J. Mol. Biol. 248, 733–738.
[10] Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) Science 253, 164–170.
[11] Ota, M. and Nishikawa, K. (1997) Prot. Eng. 10, 339–351.
[12] Wodak, S.J. and Rooman, M.J. (1993) Curr. Opin. Struct. Biol. 3, 247–259.
[13] Miyazawa, S. and Jernigan, R.L. (1994) Prot. Eng. 7, 1209–1220.
[14] Gilis, D. and Rooman, M. (1996) J. Mol. Biol. 1996, 1112–1126.
[15] Topham, C.M., Srinivasan, N. and Blundell, T.L. (1997) Prot. Eng. 10, 7–21.
[16] Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) Protein Sci. 1, 409–417.
[17] Holm, L. and Sander, C. (1994) Nucleic Acids Res. 22, 3600–3609.
[18] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[19] Jeppsson, J., Källman, I., Lindgren, G. and Fägerstam, L.G. (1984) J. Chromatogr. 297, 31–36.
[20] Ward, C.M., Fay, K.C., Brennan, J., Lowrey, I. and Blacklock, H.A. (1992) Aust. N. Z. J. Med. 22, 390–391.
[21] Rahbar, S., Louis, J., Lee, T. and Asmerom, Y. (1985) Hemoglobin 9, 559–576.
[22] Ota, M. (1996) Ph.D. Dissertation, Waseda University.