Protein Engineering vol.14 no.8 pp.557–564, 2001

Knowledge-based potential defined for a rotamer library to design protein sequences

Motonori Ota[1,2], Yasuhiro Isogai[3] and Ken Nishikawa[1]

[1]National Institute of Genetics, Mishima, Shizuoka 411-8540 and
[3]The Institute of Physical and Chemical Research (RIKEN), Wako, Saitama 351-0198, Japan

[2]To whom correspondence should be addressed.
E-mail: mota@genes.nig.ac.jp

A knowledge-based potential for a rotamer library was developed to design protein sequences. Protein side-chain conformations are represented by 56 templates. Each of their fitness to a given structural site-environment is evaluated by a combined function of the three knowledge-based terms, i.e. two-body side-chain packing, one-body hydration and local conformation. The number of matches between the native sequence and the structural site-environment in the database and that of the virtually settled mismatches, counted in advance, were transformed into the energy scores. In the best-14 test (assessment for the reproduction ability of the native rotamer on its structural site within a quarter of 56 fitness rank positions), the structural stability analysis on mutants of human and T4 lysozymes and the inverse-folding search by a structure profile against the sequence database, this function performs better than the function deduced with the conventional normalization and our previously developed function. Targeting various structural motifs, *de novo* sequence design was conducted with the function. The sequences thus obtained exhibit reasonable molecular masses and hydrophobic/hydrophilic patterns similar to the native sequences of the target and act as if they were the homologs to the target proteins in BLASTP search. This significant improvement is discussed in terms of the reference state for normalization and the crucial role of short-range repulsion to prohibit residue bumps.

Keywords: 3D profile/*de novo* design/inverse folding/stability/threading

Introduction

The search for a protocol to design a foldable sequence into a desired structure is one of the major issues in structural biology. In the early stages of this endeavor, a simple fold of the four-helix bundle was frequently employed as a target motif (Regan and DeGrado, 1988; Hecht *et al.*, 1990). For this structure, a manual sequence design succeeded well, based on the amino acid propensities for the secondary structure formation and the binary pattern of polar and non-polar residues (Betz *et al.*, 1997). Designing sequences for structures containing a β-sheet was more difficult (Quinn *et al.*, 1994), yet some progress was reported recently (de Alba *et al.*, 1999). To design more complicated folds, we cannot rely on the manual building method and thus we should adopt a computational approach. Stringent work along this path was carried out by Dahiyat and Mayo (Dahiyat and Mayo, 1997). They used a computational algorithm to design a sequence of about 30 amino acids in length targeting the ββα zinc finger motif and confirmed its validity by an NMR structure analysis. They employed a rotamer library, a set of frequently observed side-chain templates, to represent the protein side-chain conformations (Ponder and Richards, 1987). After fixing the backbone coordinates of the target structure, the optimal sequence was sought by a sophisticated algorithm, the dead-end elimination (Desmet *et al.*, 1992). The target function for the optimization consists of a physico-chemical Lennard–Jones potential for each atom pair and an empirical hydration function determined by a QSAR-like procedure (Dahiyat and Mayo, 1996).

We sought to develop a method for designing a sizable protein with a length of more than 100 amino acids. Since there is an enormously large number of different sequence candidates (20^{100}), rigorous physico-chemical approaches could not be employed for the calculation to be completed within a realistic computational time. Thus, a more practical way was attempted (Shakhnovich and Gutin, 1993; Jones, 1994; Godzik, 1995) to utilize a knowledge-based potential of amino acids developed for protein fold recognition by the threading technique (Ota and Nishikawa, 1997). Progress in this area was compiled in the Proceedings of CASP3, Critical Assessment of Techniques for Protein Structure Prediction (Murzin, 1999). In the structure prediction, called the forward-folding protocol, a query sequence is searched for its compatible fold in the structural library, whereas in the reverse case, for designing the sequence (inverse-folding protocol), compatible sequences for a target fold are sought in the sequence space (Kocher *et al.*, 1994). Although the two protocols look symmetrical at a glance, the inverse-folding protocol revealed that more attention should be paid to the reference state problem, i.e. how to define the zero level of the function (Rooman and Wodak, 1995; Miyazawa and Jernigan, 1999; Ota and Nishikawa, 1999).

In our previous work (Ota and Nishikawa, 1997), the energy of the reference state was adjusted by utilizing the three-dimensional (3D) profile, in which the fitness of each amino acid type to every structural site is shown as a two-dimensional array of 20 (number of amino acid types)×N (length of protein) (Bowie *et al.*, 1991). Employing the 3D profile of ribonuclease HI, the difference of the folding energy (ΔG) of a mutant relative to that of the wild-type ($\Delta\Delta G$) was calculated and the computation was correlated with the experimental value (ΔT_m) (Ota *et al.*, 1995). The design of human lysozyme mutants more stable than the native was further carried out and nine mutants were selected which were predicted to be strongly stabilized. However, significant stabilization was observed in only one of the mutant proteins (Takano *et al.*, 1999a). This failure was partly attributed to the poor estimation of the amino acid side-chain volume, because the large-sized amino acids, e.g. Trp, Phe, Met and Leu, tend to augment the stability in the calculation [see Table I of Takano *et al.* (1999a)], even if the reference state was adjusted to exclude atomic overlaps among large side chains (Ota and Nishikawa, 1997).

Table I. Proteins used for the best14 test

PDB	Length	Type[a]	Fold[a]
135l	129	α+β	Lysozyme-like
1a27	285	α/β	NAD(P)-binding Rossmann-fold domains
1a6m	151	α	Globin-like
1apyA	161	α+β	N-Terminal nucleophile aminohydrolases (Ntn hydrolases)
1bbpA	173	β	Lipocalins
1btl	263	α+β	β-Lactamase/D-Ala carboxypeptidase
1cdkB	343	α+β	Protein kinases (PK), catalytic core
1ede	310	α/β	α/β-Hydrolases
1hfs	160	α+β	Zincin-like
1hvc	203	β	Acid proteases
1ilr1	144	β	β-Trefoil
1mai	119	β	PH domain-like
1nar	289	α/β	β/α (TIM)-barrel
1osa	148	α	EF hand-like
1pmy	123	β	Cupredoxins
1sbp	308	α/β	Periplasmic binding protein-like II
1vid	213	α/β	*S*-Adenosyl-L-methionine-dependent methyltransferases
1ycc	108	α	Cytochrome *c*
256bA	106	α	Four-helical up-and-down bundle
2ayh	214	β	ConA-like lectins/glucanases
2dri	271	α/β	Periplasmic binding protein-like I
2rn2	155	α/β	Ribonuclease H-like motif
2sfa	191	β	Trypsin-like serine proteases
4ccx	291	α	Heme-dependent peroxidases
5p21	166	α/β	P-loop containing nucleotide triphosphate hydrolases

[a]According to the SCOP database (Murzin *et al.*, 1995).

The issue of side-chain volume also arose in the *de novo* design of protein sequences. The optimal sequence was sought to fit the globin fold by generating the 3D profile recursively (Isogai *et al.*, 1999). The computationally selected sequence, SCS1, contained many Trp, Met and Leu, but few Val, Ile and Ala, and its molecular mass was larger by more than 10% as compared with the native globin [see Table 1 of Isogai *et al.* (1999)]. Bumping residues were trimmed manually using 3D modeling and many large-sized residues were substituted with smaller ones, based on the 3D profile. The refined sequence (DG1) was synthesized and its structural features were evaluated extensively. It was concluded that the secondary structure and the global shape of DG1 resemble the targeted globin fold; however, it lacks the structural uniqueness at the side-chain level: its NMR spectrum is broad and it exhibits poor folding cooperativity (Isogai *et al.*, 1999). The structural uniqueness was improved by manual replacements of Leu and Met in DG1 with β-branched amino acids, Ile and Val (Isogai *et al.*, 2000). Following these results, we were urged to develop a theoretical method to improve the side-chain packing, because part of our failure was attributed to the poor estimation of the steric interactions between side-chains.

In this study, a knowledge-based potential for a rotamer library, based on the structures in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977), was developed to improve the side-chain packing estimation for designing a protein sequence. Thus far, potential functions for threading have usually been defined for each residue or residue pair, but detailed side-chain conformations were ignored (Miyazawa and Jernigan, 1985; Sippl, 1990). We introduced 56 rotamers in total (Ponder and Richards, 1987), comprising three rotamers depending on the χ_1 angle for each amino acid type, except for Ala and Gly. Each rotamer is represented by a virtual atom situated at the centroid of the rotamer conformation. The side-chain packing function was defined to depend on the distance between the rotamer centroids. It also depends on the types of interacting amino acid residues, but does not depend on the rotamer types. For instance, the three Leu rotamers (Leu1, Leu2 and Leu3) share the same potential function; the only difference characterizing these three rotamers is the position of the side-chain centroid relative to the backbone coordinates. Clearly, our knowledge-based function differs from the functions derived for each atom pair (Melo and Feytmans, 1997; Samudrala and Moult, 1998).

Materials and methods

Structural library

We selected 683 structures deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). These structures, determined within 2.5 Å resolution with no omitted side chains, have sequence lengths of more than 100 residues and are mutually dissimilar, sharing less than 30% of identical sequences.

Rotamer library

A rotamer library consisting of 56 templates was employed (Ponder and Richards, 1987). The side-chain conformations of amino acids other than Ala, Gly and Pro are divided into three types, referring to their χ_1 angle (type 1, 0 to 120°; type 2, 120 to –120°; type 3, –120 to 0°). Pro was divided into three types, referring to its χ_1 and χ_2 angles (type 1, $0 < \chi_1 < 60°$ and $-60 < \chi_2 < 0°$; type 2, $-60 < \chi_1 < 0°$ and $0 < \chi_2 < 60°$; type 3, $\chi_1 \times \chi_2 > 0$). Apparently, Ala and Gly have a single template. Although the same library proposed by Dunbrack and Karplus (Dunbrack and Karplus, 1992), consisting of 112 templates, was employed at the beginning of this study, a library consisting of many rotamers was unsuitable for the knowledge-based approach, mainly for statistical reasons, and therefore this minimal set was used.

Compatibility function

Three knowledge-based terms, the side-chain packing, hydration and local conformational functions, constitute the

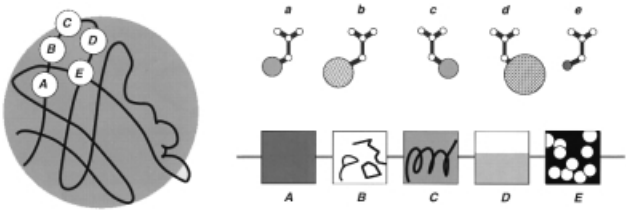

Fig. 1. Schematic diagram to illustrate the native match/mismatch (NMM) model. On the left-hand side, there is a globular protein with structural environments *A*, *B*, *C*, *D* and *E* at different residue sites. On the right-hand side, *A*, *B*, *C*, *D* and *E* are shown as boxes with different patterns, indicating their different structural features. Above them, native rotamers corresponding to *A*, *B*, *C*, *D* and *E* are shown as *a*, *b*, *c*, *d* and *e*, respectively. If rotamer *a* is put in its native environment *A*, then we call it the native match. If rotamer *a* is put in any non-native environment (other than *A*), then we call it the mismatch.

compatibility function. The hydration function was the same as used in the previous work (Ota and Nishikawa, 1997). The local conformational function also followed the previous definition, except that it was defined for each rotamer instead of each amino acid. The side-chain packing function is explained below.

To compile the side-chain packing function, we employed the native match/mismatch model (NMM model), which is illustrated in Figure 1. In the left-hand view, the environment at each residue site (*A*, *B*, *C*, *D* and *E*) is extracted from the PDB entry and is shown on the bottom of the right-hand view and they are aligned against the native rotamers at the corresponding sites, *a*, *b*, *c*, *d* and *e* (top of the right-hand view). After mounting a native rotamer on its own environment (*a* on *A*, *b* on *B*), the distance between the rotamers (centroid distance between the native rotamer and a rotamer situated near the site) was measured. Also, we virtually settled the mismatch, corresponding to misalignments, as *a* on *B*, *a* on *C*, etc. and measured the distance between the rotamers. Each rotamer centroid is situated at the expected center of the gravity of side-chain atoms except Cβ. For Ala and Gly, Cβ and Cα atoms are regarded as the rotamer centroids, respectively. The number of observations in which rotamers of residue pair *xy* is located at a distance *r* under the native-match and mismatch events is counted and normalized by each of the total number in the contact region, to give the potential function:

$$\Delta E^{xy}_{\mathrm{NMM}}(r) = -\left[1 - p^{xy}_{nc}(r)\right] RT \ln \frac{f^{xy}_{N}(r)}{f^{xy}_{M}(r)} \qquad (1)$$

where $f^{xy}_{N}(r)$ and $f^{xy}_{M}(\mathrm{r})$ are the frequencies of occurrence for rotamers, each of which belongs to residue *x* or *y*, located at distance *r* in the native match and mismatch, respectively. $p^{xy}_{nc}(r)$ is part of the damping factor: the non-contact probability of rotamers of residue pair *xy* at *r*. If some atoms in each of side chains are within 5 Å in the native structure, then they are defined to be in contact. *R* is the gas constant and $T = 300$ K. The side-chain packing function is defined when the sequence separation is more than four residues and sparse data correction as proposed by Sippl was applied (Sippl, 1990). The discretization of the spatial distance *r* is carried out using the bins of 1 Å width and linear smoothing: if *r* is 4.8 Å, 0.7 and 0.3 portions are counted for bins of 4–5 and 5–6 Å, respectively. Equation 1 indicates that the mismatch plays the part of the reference state. The same formulation is also introduced to the interactions between a main-chain atom (N, Cα, C, O or Cβ) and the rotamer centroid as a supplement of the side-chain packing. A combined function composed of this function, the hydration function and local conformational function is abbreviated as 'rotamer'.

Another side-chain packing function originally introduced by Sippl (Sippl, 1990), was also compiled (the acronym of the combined function is 'Sippl_r'). This formulation considers all rotamer pairs of any amino acid type at a distance in the native structures as the reference state for the normalization:

$$\Delta E^{xy}_{\mathrm{Sp}}(r) = -\left[1 - p^{xy}_{nc}(r)\right] RT \ln \frac{f^{xy}_{N}(r)}{f(r)} \qquad (2)$$

where $f(r)$ is the probability of any rotamer pair having a centroid separation of *r* (maximum *r* is 10 Å). This is a distinct point compared with the NMM model, in which not only native but also non-native (mismatch) rotamers are taken into account.

Results and discussion

Best-14 test

The 3D profile was constructed for the 25 proteins listed in Table I using the compatibility function of 'rotamer' derived from the remaining 658 structures in the structural library. A part of the 3D profile of human lysozyme is shown in Figure 2. The 56 rotamers are arranged in order of the fitness score from left to right. The native rotamer is highlighted and generally appears on the left-hand side (preferable to its own site). We examined how many native rotamers were positioned within the best-14 ranks amongst the 56 rotamers and estimated the success rate (best-14 score) for 5024 sites of the 25 proteins. Fourteen is a quarter of 56 and therefore this test, called the best-14 test, is comparable to the previously conducted best-five test for the 20 amino acid types (Ota and Nishikawa, 1997). It was assumed that the best-14 score should be higher as the function improves, given that almost all native rotamers (residues) fit well to the native site.

The results of the best-14 test, conducted for rotamer and Sippl_r functions, are shown in column 2 in Table II. The best-14 score of the rotamer function is 75.7%. This score is three times higher than the random level (25%) and better than the score of the Sippl_r function (64.2%). Each function is also asked whether the native rotamer is correctly detected as having the lowest energy among the three alternatives of the same amino acid residue (except Ala and Gly). For instance, if the native rotamer is Leu1 and it has the lowest score among Leu1, Leu2 and Leu3, then the assignment is correct. This type of test was applied only to residues at buried sites, where the hydration class defined by the number of surrounding heavy atoms is ⩾7 (Ota and Nishikawa, 1997). The prediction accuracy was 78.6% by the rotamer function (33% by the random assignment), as shown in column 3 of Table II. Apparently, it performs better than the Sippl_r function (70.1%). Gilis and Rooman (1997) reported that the relative weights for the local and non-local terms depend on the hydration class, so that the statistics of the rotamer positions were suspected to depend on the hydration class (buried, partially buried in protein or exposed to water). To check this point, counting and compilation were carried out individually on the hydration class on the side-chain packing function or the local conformational function. However, the dependence on the hydration class has little influence on the results (data not shown). Obviously, the introduction of a rotamer succeeded

PDB	N	Rt	L	H	En	Rk	Eb	profile table	Ew
1lz1	1	K2	e	3	-0.62	4	-0.87	R1K1R2K2S3T2Q1G1T3R3S2Q2H1N2S1T1K3D2N1H2Y2E2Q3A1C2M1V1W2F2E3C1V3N3C3I2V2M2E1M3W3L2W1H3I3D3I1P3P1D1L3Y3P5L1F1Y1F3	1.22
1lz1	2	V2	e	1	0.80	42	-1.05	P1P3P5K1Q1T2S3K3N2S2E1E2G1K2Q2R1Q3T3T1E3H1S1D2N1N3D3R2R3H2A1M1H3V3D1V1M2Y1L1F1Y2W2V2C1M3W1C2I1C3L3F2I2I3W3L2F3Y3	1.36
1lz1	3	F3	e	8	-1.43	1	-1.43	F3Y3L3P3W3C2C3H3M3I3V3A1P5P1V2T2S2T3C1S3V1Q3G1K3T1R3N2N3S1I1I2E3D3D2Q2M1M2H1N1L1Q1E2E1D1W1K1R1L2F1Y1H2R2K2W2F2Y2	7.11
1lz1	4	E3	e	1	0.23	20	-1.21	S1D2N1T1P3P5N2P1S2T2D1K1K3H2G1E2T3Q3R3E3E1R1H3D3S3M1N3K2A1Q1C1C2V3Q2Y3W3L1R2V1L3M3W2F3I1C3M2I2V2L2H1F2Y1Y2W1I3F1	2.60
1lz1	5	R2	a	5	0.10	27	-0.74	P3L2Q2E3D3I2P5W3E2M2K2A1E1V1C3P1T3S3K1V2H2H1Q3Y1Y3R1R2M3F3L3S1N3I1S2D1I3W1T2G1H3N2T1K3M1R3V3Q1F1C1D2C2W2L1N1F2Y2	1.59
1lz1	6	C3	a	1	0.30	13	-0.35	E1D1S3S1P3Q1A1N3D3G1S2K1C3P5T3E3N1N2R1H1P1T1Q3T2C2H3W1E2M1V1L1Y1K3R2V3V2D2Y3I2I1M3Q2F1M2I3K2F3C1W2R3W3L3Y2F2H2L2	2.84
1lz1	7	E2	a	3	-0.71	1	-0.71	E2K3R2K2Q2H2R3N3D3Q3Y2S3T3E3N2A1L2H3F2W2M3S2M2S1D2I3L3V2Y3T1G1C3W3C2P3F3T2I2V1V3N1M1P5E1Q1I1C1K1R1L1D1P1W1H1Y1F1	6.05
1lz1	8	L2	a	9	-1.84	1	-1.84	L2F2M2I3I2V2Y2W2A1H2C2V1M3T3C3V3Q2I1S3S2L3M1G1K2E2S1C1R2T1L1Q3N2T2E3N3D3D2R3Q1W1E1N1R1K3K1W3P3D1H1H3P5Y1P1F1F3Y3	5.64
1lz1	9	A1	a	5	-0.83	1	-0.83	A1G1I3E3M3Q3V2S3N3V3H3T3C3L3S1V1E2T1S2M2D3F3C1Q2C2W3K3I1M1I2T2W2Y3H2N1K2N2Q1R3P3D1D2E1R2L2L1P5F2Y2H1W1P1K1R1Y1F1	4.82
1lz1	10	R3	a	2	-0.39	5	-0.61	E2E3K3Q2R3N3Q3R2K2S3D3T3A1D2Y2K1H2S2R1M3W2M2N2S1H3F2L2G1L3V2E1Q1T1I3C3W3N1T2M1C2I2V1V3L1F3Y3P3D1I1C1W1H1P5Y1P1F1	2.40
1lz1	11	T3	a	6	-0.41	12	-1.09	I3L3M2M3W3L2W2V2Y2F2Q2T3E2I2C2V1F3R3H2Q3A1C3R2Y3H3D3V3N3S3K2K3E3I1G1C1T1S2D2N2S1Q1M1T2E1L1K1N1R1D1P3W1H1P5P1Y1F1	4.40
1lz1	12	L3	a	9	-1.65	1	-1.65	L3M3I3F3V2C3M2A1W3I2Y3V3T3C2L2V1S3H3Q2Q3I1G1N3M1S2T1S1D3E3L1C1E2K3N2T2R2K2R3H2Q1D2W2E1F2K1N1P3P5D1W1R1H1P1Y2Y1F1	4.44
1lz1	13	K2	a	4	-0.71	2	-0.80	R2K2Q2L2H2W2E2Y2A1F2V2M2T3N2S3C2S2D2C3S1Q1I2I3M1V1G1T1V3N3Q3E1E3M3T2L1I1H3R1K1C1W3N1D3L3K3W1D1Y3R3F3P3P5Y1H1P1F1	3.48
1lz1	14	R1	a	3	0.14	21	-0.66	K2E3R2D3E2N3Q2K3Q3S1E1K1R3S3A1Q1H2H3D2T1R1M3T3Y2Y3N2L3F3M1V1T2W2G1L2V3F2I1M2C2I3S2C3V2I2N1W3L1C1D1H1W1P5Y1P3F1P1	3.54
1lz1	15	L3	g	6	-1.31	1	-1.31	L3H3F3Y3M3Q3C3N3A1C1K3W3R3T1V3E3S1S3I1D3V1M1Q1N1G1E1I3M2E2D2N2C2I2D1V2F2H2K1S2Q2H1L2Y2W1T2T3W2L1P5K2R1P3R2P1Y1F1	4.09
1lz1	16	G1	l	4	-1.21	1	-1.21	G1N2N3K3D2R3H3Q3E3S3H1D3E1H2C3F3S2R2C2P5N1Q2A1L1K2R1M1Q1W1Y3S1T3W3K1L3T2C1M3V3L2F1V1D1Y2W2Y1E2T1I2M2F2V2P3I3I1P1	2.27
1lz1	17	M3	g	8	-0.14	8	-0.77	V3V1T1C3C1S1L3M3I2S3T2A1H3V2C2I3Q3Y3E3I1F3T3G1N3N1W3D3K3M2D1F2L2W2E2Y2R3N2P3S2D2Q2P1K2R2E1H2P5M1Q1H1L1K1W1R1F1Y1	6.30
1lz1	18	D2	e	4	-0.25	7	-0.65	E2N2Q2W1K2Q1D2S2S3T2Y2R2W2H2A1D1S1C1T3G1F2T1E1I2V1Y1V3E3F1M2N1P5I1V2C2L2N3P1D3P3H1C3Q3L1R3I3M3M1K3H3K1R1F3Y3L3W3	1.85
1lz1	19	G1	l	1	-1.42	1	-1.42	G1N2N3D3S3K3D2Q3D1E3S2K2P5Q1T2P3T3R3H3R2E2H2E1R1P1A1N1K1W3M1S1Q2H1M3Y3T1L1C3F3L2C2I1L3M2V1W1F2C1I2Y2W2Y1F1V2I3V3	2.71
1lz1	20	Y2	e	9	-1.48	1	-1.48	Y2M1C1F2W2T3T1V1M2L1C3V2H2A1C2S3N2H1Q2G1L2S1S2D2Q1Q3E2N1I2E3I3D1V3T2Y1K1R2M3N3I1P1R3L3F1E1W1K2R1P3P5W3D3K3H3F3Y3	6.96
1lz1	21	R3	l	5	0.64	22	-1.13	N3D3G1Y3C3W3S3F3H3N2M3Q3E3Q2L2L3S2F2A1K3D2R3M1P5V3N1T3I2T2C1C2V2W2Y2V1Q1H2M2S1H1P3R2K2W1L1I1T1I3F1P1E1D1E2Y1K1R1	4.13
1lz1	22	G1	l	3	-1.19	1	-1.19	G1N3N2S3D3K2Q3K3H3R3Y3Q2E3F3E2A1P5Q1N1D1M3W3T1C3K1H1D2R1M1V3E1T2T3F2L1L3R2S2H2C1W1Y1C2M2V2Y2W2V1P3F1I2S1P1I1I3L2	2.67
1lz1	23	I3	e	7	-0.56	5	-1.05	F3Y3I2L3I3V1I1V2M3M2C1H3C2T2W3M1V3C3Q3S3K1T1A1L1R1S2N2N3G1H1S1D2Q2L2E3H2K3T3Q1D3E2Y1F2K2R3W1P3R2E1W2N1P5F1P1D1Y2	1.95
1lz1	24	S1	e	4	-1.08	1	-1.08	S1D2T1N2N1P3H2D1S2P5P1T2K3C2R3C1E1G1T3E2Q3K1E3H3R1M1Y3D3K2Q1N3C3A1S3L3W3V3W2R2F3L1V1Q2Y2F2M3L2V2I2I1M2H1I3W1Y1F1	2.25
1lz1	25	L2	a	7	-2.17	1	-2.17	L2I2M2V1V2H1C2T3Q2F2S3H2A1I3E2Y1C3I1S1W2W1N2T1K2P5Y2E1M3M1C1S2T2V3P3P1D3G1E3F1R2R1K1L1Q3N3D2L3R3Q1N1D1W3K3H3F3Y3	4.26
1lz1	26	A1	a	4	-0.47	2	-0.65	E1A1Q1S1G1R1S2T3D1H1P3K1N1Q3S3P5M1Y1V2E3W1L1D3V1C1T1P1N3C3V3I1F1T2I3R3M3H3C2L3K3Q2W3I2N2E2Y3F3M2D2R2L2H2K2W2F2Y2	7.31
1lz1	27	N3	a	7	-0.22	11	-1.06	L3M3I3V2A1W3C3Q3F3S3N3T3D3H3I2E3Y3M2V1C2R3K3G1V3L2S2C1S1Q2T1R2K2E2H2W2T2F2I1N2N1E1Y2Q1P3D2M1D1P5K1L1H1P1R1W1F1Y1	6.35
1lz1	28	W3	a	9	-0.56	5	-1.34	L3M3I3V2W3F3C3A1V3L2M2F2V1C2I2T3W2Y3S3I1H3C1Q2N3Q3S1S2H2M1D3Y2T1G1E3T2E2Q1N2P3K3E1L1K2R3D2K1R2P5N1W1R1D1H1P1Y1F1	4.52
1lz1	29	M3	a	8	-1.02	4	-1.60	I3L3L2M3V2C3M2W2H2T3Y2I2S3V3C2Q2N3F2A1S2Q3D2H3F3S1E2C1D3V1T1E3R2I1N2G1K2T2W3Y3N1E1K3R3Q1D1M1P3L1P5H1R1K1W1P1Y1F1	6.03
1lz1	30	C2	a	7	-1.30	1	-1.30	C2A1G1C1Q2H2M2M3E2S1S2C3S3M1T3N2V2I3Q1H3L2K2V3E1Q3N1D2T1V1H1E3I2K1T2N3D1R2W3W2D3R1I1P3F2R3K3W1P5L1L3Y2P1F3Y1F1Y3	3.85
1lz1	31	L2	a	7	-0.98	1	-0.98	L2M2I3V2F2R2I2Q2A1K2T3M3V1S2C2S3W2H2C3V3E2Y2Q3M1I1L3N2T2C1Q1N3S1T1D3G1K1H3D2K3E3R1L1R3E1F3W3N1W1Y3D1P3H1P5Y1F1P1	3.88
1lz1	32	A1	a	9	-1.05	1	-1.05	A1V2V3C3I3V1M3C2I1S3T3S2L3G1M1I2M2C1L1T1S1N3Q3W3H3F3Q2E3Y3H2D3L2T2N2Q1W2E1E2K2D2K3W1K1P3H1R3R2N1P5R1F2D1Y2P1Y1F1	3.62
1lz1	33	K2	a	5	-0.28	11	-0.88	F2Y2H2W2R2E2L2Q2A1M2K2V2I3M3C2Q3R3W3T3G1E3I2K3D2N2S2H3V1C3L3S3V3S1D3N3T1Y3T2I1Q1F3M1C1K1N1E1R1L1P3D1H1P5W1P1Y1F1	4.69
1lz1	34	W3	a	5	-0.33	3	-0.45	F3D3W3A1H3I3W2S3F2Y3Y2T3N3V2M3M2C2I1T1H2V3S2E3Q3I2S1G1L1L3V1C3H1K2M1P3D2Q1N2Q2T2L2N1K3D1E1C1P5E2R3F1W1K1R2P1Y1R1	1.97
1lz1	35	E3	a	8	0.44	16	-0.48	L3A1I3C3T3M3S3V2Q3N3V3H3S1G1T1E3C2C1R3K3D3S2V1W3F3I1T2Q1N1Y3Q2E1M1E2L1I2K1D1R1M2D2N2P3R2H1P5K2H2L2P1W1W2F1Y1F2Y2	8.19
1lz1	36	S1	a	7	-0.29	2	-0.38	T1S1A1C1S3C3G1P5V2V3T3C2N3V1N1T2M3Q2Q3H3S2N2P1P3I3M2D3E3E2I2I1Q1D1H1Y1W1D2L3E1K3R2W3M1K2F1H2R3L2Y3L1K1W2R1F3F2Y2	8.88
1lz1	37	G1	l	2	-0.97	2	-1.06	N2G1N3D2Q3R3H3E3T1A1P5K1K3T2M1S3Q1Q2N1R1D3H1S2L1E1W1Y1F1R2D1C2E2C1H2S1C3W3Y3K2M3Y2V1W2M2I2P3V3P1L2I1T3F2F3L3V2I3	3.36
1lz1	38	Y3	l	4	-0.10	4	-1.30	G1N3F3Y3D3W3H3S3C3R3N1A1N2Q3P5M3C1L3E3Q1V3T3R1C2L1V1K3H1M1T2D1Q2E1V2I1I2T1S2W1K1I3S1P3Y1E2P1F1M2D2K2R2H2L2W2F2Y2	6.44
1lz1	39	N2	e	2	-1.13	1	-1.13	N2S2D2T1T2S1G1A1K2K1Q3R3S3H2E2Q1K3H3Q2C1R2N1E3C2R1N3E1F2P1V3M2Y2V1M1D1I3W3V2D3M3I2P3T3H1I1W2P5L3C3Y3F3L2L1Y1W1F1	2.83
1lz1	40	T1	g	5	-0.41	4	-0.46	P3S1N1T1Q3C1K3V3N3V1S3C3A1D1L3E3D3R3M3H3T2T3G1V2I1P1W3I3P5S2Q1E1Q2F3N2I2Y3H1C2M2W1M1K1D2L1E2L2R1Y1F1K2H2R2W2F2Y2	5.37

Fig. 2. Part of the 3D profile of human lysozyme (PDB code 1lz1) constructed with the rotamer function. Rotamers are ordered according to fitness from left to right in the profile table. Native rotamers are highlighted. N, the site number; Rt, the native rotamer; L, the local structure; H, the hydration class; En, score of the native rotamer; Rk, ranking position of the native rotamer amongst 56 alternatives; Eb, score of the fittest rotamer; Ew, score of the worst rotamer [see also Ota and Nishikawa (1997) for details].

Table II. Various assessments for the potential functions

Function	Best-14 score (%) 25 proteind	Rotamer detection (%) 25 proteins	Correlation human lysozome	Correlation T4 lysozyme
Rotamer	75.7	78.6	0.58	0.69
ON97	58.9[a]	–	0.48	0.56
Sippl_r	64.2	70.1	0.61	0.67

[a]Best-5 score.

well in the best-14 test (75.7%) owing to the improvement in the packing, as compared with the previous function, ON97, that was defined for the amino acid pairs, yielded 58.9% in the best-5 test for the same 25 samples (Ota and Nishikawa, 1997).

Stability analysis

To estimate the relative stability of a mutant protein to the wild-type ($\Delta\Delta G$), we have to know the energy level of the denatured state. Regarding the ensemble of non-native (mismatch) occurrences specified by the residue-pair (xy) and their distance (r) as the denatured state, the NMM model directly provides the relative side-chain packing energy (Equation 1) between the native and denatured states. This approximation is allowable if we virtually define the denatured state of given x, y and r, maintaining the structurally compact state but lacking the structure (r)–sequence (xy) relationships observed in the native state (Rooman and Wodak, 1995). The stability of the mutant proteins relative to the wild-type ($\Delta\Delta G_c$) was calculated and compared with the experimentally determined energy $\Delta\Delta G_e$. Since the present approach with side-chain rotamers requires the 3D structures of the mutant proteins, mutants of human lysozyme and T4 lysozyme were employed, for which substantial stability and structural data are available. Seventy-seven single mutants of human lysozyme, analyzed mainly by Yutani and co-workers (Herning *et al.*,

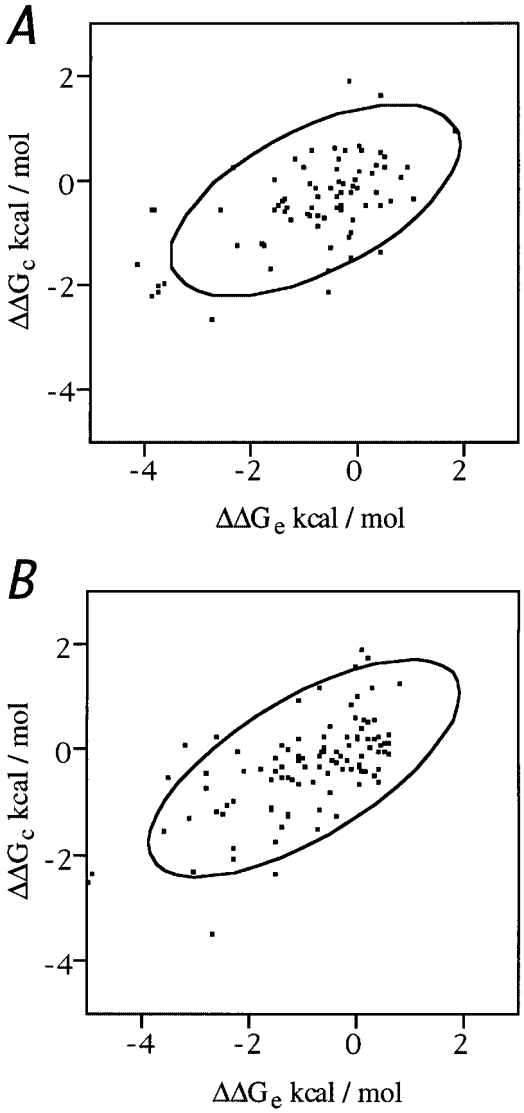

Fig. 3. Scatter plots of the experimental stability ($\Delta\Delta G_e$) of a mutant protein and its calculated stability ($\Delta\Delta G_c$) by the rotamer function, for human lysozyme (**A**) and T4 lysozyme (**B**). Positive values indicate stabilization in comparison with the wild-type; 90% density ellipses are shown.

1993; Takano *et al.*, 1995, 1997a,b, 1999a,b; Yamagata *et al.*, 1998; Funahashi *et al.*, 1999), and 103 point mutations of T4 lysozyme, analyzed by Matthews (Matthews, 1995), were employed. The T4 lysozyme data are a mixture of point mutants of the wild-type and point mutants of lysozyme lacking two cysteines (Matthews, 1995). We adopted both the real native protein (PDB code 3lzm) and pseudo-native, the double mutants of C54T and C97A (PDB code 1l63), as parent structures and mutants generated from either of the parent structures are called 'mutants' and were treated in the same way. Each data set is a mixture of various mutants whose stability changes were attributed to several reasons, e.g. hydrophobicity, secondary structure propensity or hydrogen bonding (Matthews, 1995; Funahashi *et al.*, 2001). We examined the total ability of functions for the stability analysis regardless of the reason for the stabilization of each mutant.

Figure 3A and B are the scatter plots of the theoretical values ($\Delta\Delta G_c$), estimated with the rotamer function, against the experimental values ($\Delta\Delta G_e$) for the human and T4 lysozymes, respectively. The two values correlate: the correlation coefficients for the human and T4 lysozymes are 0.58 and 0.69, as shown in Table II. The results indicate a notable improvement as compared with the previous study, where the correlation coefficients for the human lysozyme data and T4 lysozyme data were 0.48 (Takano *et al.*, 1999a) and 0.56, respectively (Table II). Since this calculation contains no adjustable parameters, the correlation coefficients are not comparable directly with other studies in which many factors were determined by regression (Gromiha *et al.*, 1999; Funahashi *et al.*, 2001). In this test, the Sippl_r function performs as well as the rotamer function (Table II). We classified single-residue substitutions into four groups: between hydrophobic (C, F, I, L, M, V, W, Y)–hydrophobic residues (denoted HH), hydrophilic (D, E, H, K, N, Q, R, S, T)–hydrophilic residues (PP), hydrophobic–hydrophilic residues (HP) and those involving Ala, Gly or Pro (AGP). Substitutions between the residues of the same character, e.g. HH or PP, do not yield a good correlation between the computation and the experiment, whereas substitutions between different types, HP and AGP, significantly correlate: the correlation coefficients are 0.65 (human lysozyme) and 0.76 (T4 lysozyme). Also, correlation at the buried sites is remarkable: 0.64 (human lysozyme) and 0.73 (T4 lysozyme). It is noteworthy that the energy scale is almost identical for the ordinate and the abscissa in Figure 3 (ratio 0.6). In other words, the theoretical and experimental values are similar on the absolute scale (Mirny and Shakhnovich, 1996), although the knowledge-based function has frequently been criticized regarding this point (Thomas and Dill, 1996; Ben-Naim, 1997).

Inverse-folding search

The structure–recognize–sequence protocol, named the inverse-folding search, was conducted with the rotamer and Sippl_r functions. A benchmark set composed of 382 structures used in the previous work was also employed (Ota and Nishikawa, 1999); 177 of them have at least one related protein belonging to the same superfamily of the SCOP database (Murzin *et al.*, 1995) and the structures of 165 queries (<400 residues long) sought their homologous sequences (including the native one) in the sequence library. Here 657 true pairs and 62 373 false pairs constitute all of the 3D–1D matches.

Employing so-called the jack-knife procedure, the 3D profiles of query structures were accomplished by each of the functions derived from this 382 data set. Assigning the native type of the rotamer to each sequence, e.g. V3L3S1E2-G1E3W3Q2L3V2 . . . for the sperm whale myoglobin (1mbd) sequence, 3D–1D alignment using the 'rotamer'-based 3D profile was carried out with the optimized gap penalties. Since the conservation of the rotamer type (χ_1 angle) is not obvious in the remote homologous proteins, e.g. only 41% of rotamer types are identical between 1mbd and 1ash (ascaris hemoglobin domain I) according to FSSP alignment (Holm and Sander, 1996), the assignment of the native rotamer type into the sequence does not always guarantee an improvement of the performance. The alignment score was just regarded as the compatibility score (Ota and Nishikawa, 1999) and transformed into the SD (standard deviation unit) score.

Figure 4 is a sensitivity (number of the true positives/number of the trues)–specificity (number of the true positives/number of the positives) plot, which shows how many of the possible trues are detected at a given confidence level. The curve drawn toward the upper-right corner is excellent. A clear improvement is seen for the rotamer function (thick line) from

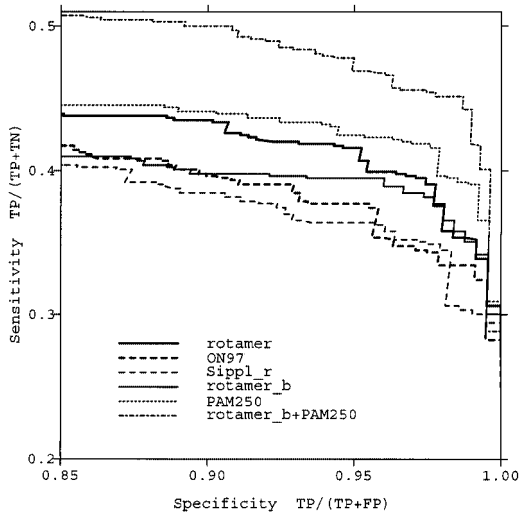

Fig. 4. Sensitivity (number of true positives/number of true positives and true negatives)–specificity (number of true positives/number of true positives and false positives) plot of the inverse-folding search. 3D profiles made by several functions, rotamer (thick line), ON97 (thick broken line) and Sippl_r (thin-broken line), were employed. Rotamer_b (thin line) represents a search results using a residue-based 3D profile made by the rotamer function, in which among the scores of the rotamers belonging to a given residue, the best of them was chosen. PAM250 (thin dotted line) shows the results of the sequence similarity search using Dayhoff PAM250 substitution matrix (Dayhoff *et al.*, 1978). Rotamer_b + PAM250 (thin one-point broken line) is the search results obtained by the combined profile consisting of rotamer_b (weight 0.7) and PAM250 matrix (–0.15). In all the searches a simple dynamic programming algorithm was employed using gap opening and extension penalties and global–local alignment scheme, i.e. gaps on the sequence termini are not penalized (Needleman and Wunsch, 1970). In the 3D profile searches gap opening penalties were altered according to the hydration class (Ota and Nishikawa, 1997, 1999). The optimized gap penalties for each search are as follows: gap opening penalty for the most exposed site (hydration class 1: go1), gap opening penalty for the most buried site (hydration class 9: go9) and gap extension penalty (ge) = 2.2, 4.4 and 0.28 kcal/mol, respectively, for rotamer, Sippl_r, rotamer_b and rotamer_b + PAM250 profile searches; and 2.4 (go1), 4.8 (go9) and 0.3 (ge) kcal/mol for ON97 profile search (Ota and Nishikawa, 1999). For the partially buried sites the gap opening penalties are determined by linear interpolation of go1 and go9. In the sequence similarity search, gap-opening and extension penalties are –11 and –2, respectively.

the previous function, ON97 (thick broken line) (Ota and Nishikawa, 1997, 1999). In this analysis, an ~3% difference in the sensitivity is statistically significant (Ota and Nishikawa, 1999). The Sippl_r function (thin broken line) performs as well as or slightly worse than the ON97 function. To estimate the ability for the actual sequence database search with the rotamer function, where the rotamer type of the sequence is unknown, we transformed the 'rotamer'-based 3D profile to a 'residue'- based 3D profile. Among the scores of the rotamers belonging to a given residue, the best of them was chosen. The results of this search using these shrunk 3D profiles are shown with the thin line (rotamer_b). The sensitivity–specificity curve of rotamer_b overlaps with the curve of the rotamer above 0.95 specificity and appears to perform better than the previous results (ON97) above 0.90 specificity. Even the rotamer profile search (rotamer) is still slightly inferior to the sequence similarity search using the Dayhoff PAM250 matrix (PAM250) (Dayhoff *et al.*, 1978); however, if it is properly combined with the sequence information (Domingues *et al.*, 2000; Panchenko *et al.*, 2000), it outperforms the sequence similarity search (rotamer_b + PAM250).

De novo design

A *de novo* sequence design of eight proteins chosen from Table I was conducted by rotamer, ON97 and Sippl_r functions. Using only the one-body function (hydration and local conformation), an optimal sequence was determined in advance. Mounting this sequence on the protein backbone structure, a 3D profile was produced by adding the two-body function. The next, tentative optimal sequence was derived from this 3D profile and used for the next calculation (as the sequence mounting on the structure). This cycle was iterated recursively until the self-consistent sequence (SCS) was given (Isogai *et al.*, 1999). When oscillation occurred between two fixed sequences, we mixed them randomly and continued the calculation. Fifty SCS sequences were generated for each of target proteins and the one with the best compatibility score was selected.

The quality of a designed protein is not easy to assess, other than by experiments and synthesis of the designed protein. However, the molecular mass and the match of the hydrophobic (H)/hydrophilic (P) amino acids with the native sequence are useful criteria to check the rationality of the designed sequences prior to experiments, as shown by Isogai *et al.* (Isogai *et al.*, 1999). In Table III, the match of the hydrophobicity in the designed sequences and their molecular masses are shown. HPid is the identical ratio between the native and designed sequences, each of which is represented by H/P symbols, and HPr is the correlation coefficient of both H/P patterns. Here we regard A, C, F, I, L, M, V, W and Y as hydrophobic.

The average molecular mass of the designed sequences by the rotamer function is 17.0 kDa and is slightly higher than the native sequence (16.5 kDa). The sequences designed by the other two functions were significantly heavy, >19 kDa. Especially the molecular mass of the sequence designed by targeting the 1mbd template by the rotamer function, 18.2 kDa, is comparable to DG1 (18.6 kDa), which was made by manual modification of the automatically designed sequence, SCS1 (19.2 kDa in Table III), and was experimentally shown to fold into a globin-like structure (Isogai *et al.*, 1999). It is clear that the introduction of the side-chain rotamers resulted in elimination of the atomic collisions and maintaining an appropriate molecular mass. The average HPid (75%) and HPr (0.48) of the rotamer function are also the highest among the results, indicating that an H/P pattern similar to the native sequence was generated by the rotamer function. For instance, sperm whale myoglobin (1mbd) and ascaris hemoglobin domain I (1ash) share 74% H/P pattern and their correlation is 0.47. Therefore, 75% HPid and 0.48 HPr obtained by the rotamer function are promising, implying that the function can generate a sequence similar to the homologs for the target protein. To clarify this point further, a BLASTP search (Altschul *et al.*, 1997) was performed with each of the sequences designed by the rotamer function against the SwissProt database release 39 (Bairoch and Apweiler, 2000). The first hit of this search is shown in Table IV. All of them are homologs of the target protein, aligned with the query (designed) sequences over half of the total length (see alignment length in Table IV), and almost all of them have significant E values. In addition, the phylogenetic tree among the designed globins, true globins and phycocyanins (remote homologs of globin) was constructed by CLUSTALW (Thompson *et al.*, 1994) and is shown in Fig. 5. The designed globins made a cluster and appear to relate with the globin family as well as phycocyanins. Hence we successfully designed a fake of the homologs by the rotamer function.

Table III. Molecular weight, HP pattern identity (HPid) and correlation coefficient (HPr) of the designed proteins

Target	Native MW (kDa)	Rotamer			ON97			Sippl_r		
		MW (kDa)	HPid (%)	HPr	MW (kDa)	HPid (%)	HPr	MW (kDa)	HPid (%)	HPr
1mbd	17.2	18.2	75	0.48	19.2[a]	78[a]	0.56[a]	21.6	71	0.39
135l	14.2	13.9	78	0.56	17.4	74	0.50	17.0	71	0.41
1bbpA	19.6	20.4	75	0.47	23.4	68	0.41	23.5	64	0.23
1ilr1	16.2	16.0	71	0.40	19.1	66	0.37	19.0	68	0.31
1osa	16.6	18.0	80	0.58	18.6	72	0.40	19.5	76	0.48
256bA	11.8	13.2	75	0.46	14.1	72	0.39	14.7	70	0.35
2rn2	17.6	17.2	72	0.45	20.9	65	0.35	21.6	74	0.44
5p21	18.8	19.3	73	0.45	22.2	68	0.41	22.0	69	0.36
Average	16.5	17.0	75	0.48	19.4	70	0.42	19.9	70	0.37

[a]Data for the SCS1 sequence in Isogai *et al.* (Isogai *et al.*, 1999).

Table IV. BLASTP search from the designed sequence against SwissProt

Query[a]	First hit				
	Code	Protein	Alignment length	% ID[b]	*E* value
d_1mbd	MYG_LAGLA	Myoglobin	90	28	7×10^{-5}
d_135l	LYC3_ANAPL	Lysozyme C-3	120	22	0.007
d_1bbpA	VEGP_PIG	Lipocalin-1	97	29	0.12
d_1ilr1	IL1X_BOVIN	Interleukin-1 receptor antagonist	97	26	0.055
d_1osa	CATR_SCHDU	Caltractin	139	28	7×10^{-15}
d_256bA	C562_ECOLI	Cytochrome B562 precurser	91	24	7×10^{-4}
d_2rn2	RNH_SALTY	Ribonuclease H	138	25	0.046
d_5p21	RASN_XENLA	P21/N-RAS	87	25	2×10^{-4}

[a]d_PDB code: designed sequence using a template (PDB code) by rotamer function.
[b]Sequence identity defined in the aligned region.

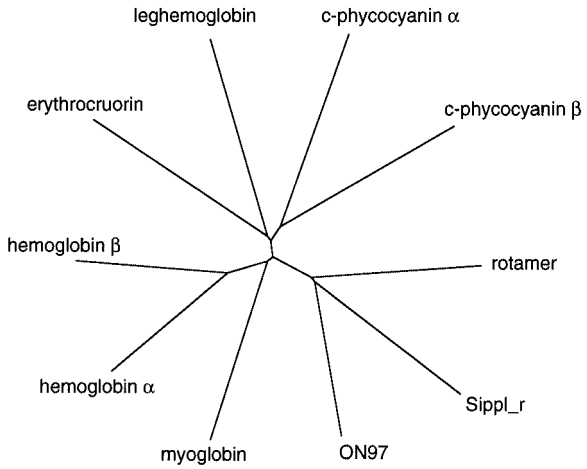

Fig. 5. The phylogenetic tree among the designed globins (by rotamer, ON97 and Sippl_r functions), true globins and phycocyanins constructed by CLUSTALW (Thompson *et al.*, 1994). Except for two hemoglobins, natural sequences are distantly related with having <25% sequence identity. SwissProt codes of the sequences are MYG_PHYCA (sperm whale myoglobin), HBA_HUMAN (human hemoglobin α), HBB_HUMAN (human hemoglobin β), GLB3_CHITH (erythrocruorin), GLB1_GLYDI (leghemoglobin), PHA1_FREDI (c-phycocyanin α) and PHB1_FREDI (c-phycocyanin β).

Reference state problem

The only difference between the side-chain packing functions of rotamer and Sippl_r is the denominator in the argument of the logarithmic function (see Equations 1 and 2), i.e. the reference state for the normalization (Rooman and Wodak, 1995; Miyazawa and Jernigan, 1999; Ota and Nishikawa, 1999). The former employs the distribution of the rotamer distance under virtually settled mismatch according to the NMM model [$f_M^{xy}(r)$ in Equation 1], whereas the latter employs the distribution of any rotamer pairs observed in the native proteins [$f(r)$ in Equation 2]. To investigate this difference further, we plotted scores of each function [$\Delta E_{\mathrm{NMM}}^{xy}(r)$ and $\Delta E_{\mathrm{Sp}}^{xy}(r)$] in Fig. 6. Surprisingly, corresponding score values over a 3 Å distance are very similar and clustered around 0 kcal/mol (blue dots). The other values, scores within 3 Å (red dots), are significantly affected by the choice of the reference state. Almost all red dots are on or under the diagonal, indicating that the Sippl_r function yields better scores within a 3 Å rotamer distance. In other words, the repulsive part of Sippl_r is very weak. This is the essential difference between the rotamer and Sippl_r functions. Rotamer pairs within 3 Å distance are rare in the real protein structures, whereas this is not the case in mismatch, resulting in a large $f_M^{xy}(r)$. Referring to this virtual situation, rotamer bumps are properly inhibited with the rotamer function. To confirm the effect of this repulsive term, we made a modified Sippl_r function in which only the scores within a 3 Å distance are exchanged by that of the rotamer function. The best-14 score of this function is 71% and it predicts rotamer type with 76% accuracy. This superiority over the results of the Sippl_r function (Table II) indicates the crucial role of steric hindrance in the short-range interaction. Interestingly, most of the scores that are frequently used in structural stability analysis are the scores of over 3 Å rotamer distance (data not shown). This is the reason why the rotamer and Sippl_r functions behave equivalently in the stability analysis (Table II). Each of the mutant proteins examined here folds into its unique structure

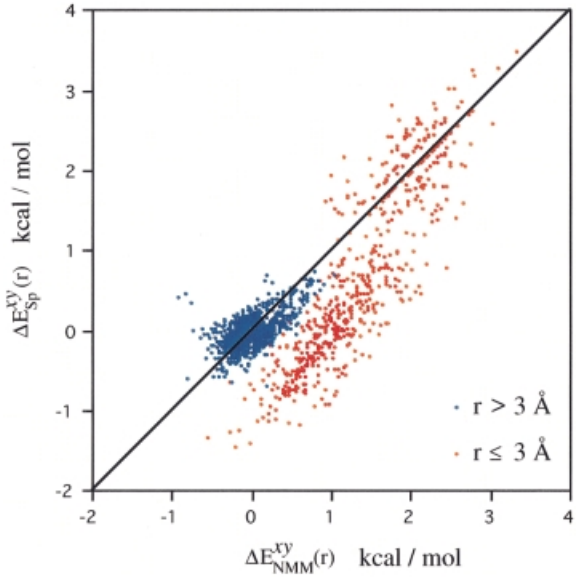

Fig. 6. Scatter plots of the interaction parameters of Sippl_r side-chain packing function $[\Delta E^{xy}_{Sp}(r)]$ against rotamer side-chain packing function $[\Delta E^{xy}_{NMM}(r)]$. Red and blue dots are the scores for rotamer–rotamer pair distance separation less than or over 3 Å, respectively.

with least deviation from its native structure, and therefore it is unnecessary to evaluate the unrealistic collisions within a 3 Å rotamer distance.

In conclusion, the knowledge-based side-chain packing function is improved by the introduction of rotamers if it is normalized by the NMM model. This new function performs better than either the previous function (ON97) or the function normalized by the ordinal formula (Sippl_r) in the best-14 test, the stability analysis of mutant proteins and the inverse folding search. This improvement was achieved by the proper estimation of packing, especially the repulsive part of short-range interactions. The function developed here might be suitable to evaluate designed sequences on a target fold and to investigate the importance of the packing in protein folding or on structural uniqueness. We are validating it also by experiments.

Acknowledgements

We are grateful to Hiroyuki Izuno for his help in obtaining data in the first stage of this work and Noriko Ito for her assistance. We also thank Katsuhide Yutani, Kazufumi Takano and Jun Funahashi for helpful discussions and their kindness in offering their latest data. This work was supported by a grant-in-aid to M.O. from the Ministry of Education, Science, Sports and Culture, Japan.

References

Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48.
Ben-Naim,A. (1997) *J. Chem. Phys.*, **107**, 3698–3706.
Bernstein,F.C., Koetzle,T.F., Williams,G.T.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Betz,S., Liebman,P. and DeGrado,W. (1997) *Biochemistry*, **36**, 2450–2458.
Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Dahiyat,B. and Mayo,S. (1996) *Protein Sci.*, **5**, 895–903.
Dahiyat,B. and Mayo,S. (1997) *Science*, **278**, 82–87.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *A Model of Evolutionary Change in Proteins*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345–352.
de Alba,E., Santoro,J., Rico,M. and Jimenez,M. (1999) *Protein Sci.*, **8**, 854–865.
Desmet,J., De Maeyer,M., Hazes,B. and Lasters,I. (1992) *Nature*, **356**, 539–542.
Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) *J. Mol. Biol.*, **297**, 1003–1013.
Dunbrack,R.L. and Karplus,M. (1992) *J. Mol. Biol.*, **230**, 543–574.
Funahashi,J., Takano,K., Yamagata,Y. and Yutani,K. (1999) *Protein Eng.*, **12**, 841–850.
Funahashi,J., Takano,K. and Yutani,K. (2001) *Protein Eng.*, **14**, 127–134.
Gilis,D. and Rooman,M. (1997) *J. Mol. Biol.*, **272**, 276–290.
Godzik,A. (1995) *Protein Eng.*, **8**, 409–416.
Gromiha,M.M., Oobatake,M., Kono,H., Uedaira,H. and Sarai,A. (1999) *Protein Eng.*, **12**, 549–555.
Hecht,M., Richardson,J., Richardson,D. and Ogden,R. (1990) *Science*, **249**, 884–891.
Herning,T., Yutani,K., Inaka,K., Kuroki,R., Matsushima,M. and Kikuchi,M. (1993) *Biochemistry*, **32**, 7077–7085.
Holm,L. and Sander,C. (1996) *Science*, **273**, 595–602.
Isogai,Y., Ota,M., Fujisawa,T., Izuno,H., Mukai,M., Nakamura,H., Iizuka,T. and Nishikawa,K. (1999) *Biochemistry*, **38**, 7431–7443.
Isogai,Y., Ishii,A., Fujisawa,T., Ota,M. and Nishikawa,K. (2000) *Biochemistry*, **39**, 5683–5690.
Jones,D.T. (1994) *Protein Sci.*, **3**, 567–574.
Kocher,J.A., Rooman,M.J. and Wodak,S.J. (1994) *J. Mol. Biol.*, **235**, 1598–1613.
Matthews,B. (1995) *Adv. Protein Chem.*, **46**, 249–278.
Melo,F. and Feytmans,E. (1997) *J. Mol. Biol.*, **267**, 207–222.
Mirny,L. and Shakhnovich,E. (1996) *J. Mol. Biol.*, **264**, 1164–1179.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Miyazawa,S. and Jernigan,R. (1999) *Proteins*, **36**, 357–369.
Murzin,A. (1999) *Proteins*, **37**(S3), 88–103.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
Ota,M. and Nishikawa,K. (1997) *Protein Eng.*, **10**, 339–351.
Ota,M. and Nishikawa,K. (1999) *Protein Sci.*, **8**, 1001–1009.
Ota,M., Kanaya,S. and Nishikawa,K. (1995) *J. Mol. Biol.*, **248**, 733–738.
Panchenko,A.R., Marchler-Bauer,A. and Bryant,S.H. (2000) *J. Mol. Biol.*, **296**, 1319–1331.
Ponder,J.W. and Richards,M. (1987) *J. Mol. Biol.*, **193**, 775–791.
Quinn,T., Tweedy,N., Williams,R., Richardson,J. and Richardson,D. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 8747–8751.
Regan,L. and DeGrado,W.F. (1988) *Science*, **241**, 976–978.
Rooman,M.J. and Wodak,S.J. (1995) *Protein Eng.*, **8**, 849–858.
Samudrala,R. and Moult,J. (1998) *J. Mol. Biol.*, **275**, 895–916.
Shakhnovich,E. and Gutin,A. (1993) *Protein Eng.*, **6**, 793–800.
Sippl,M.J. (1990) *J. Mol. Biol.*, **213**, 859–883.
Takano,K., Ogasawara,K., Kaneda,H., Yamagata,Y., Fujii,S., Kanaya,E., Kikuchi,M., Oobatake,M. and Yutani,K. (1995) *J. Mol. Biol.*, **254**, 62–76.
Takano,K., Funahashi,J., Yamagata,Y., Fujii,S. and Yutani,K. (1997a) *J. Mol. Biol.*, **274**, 132–142.
Takano,K., Yamagata,Y., Fujii,S. and Yutani,K. (1997b) *Biochemistry*, **36**, 688–698.
Takano,K., Ota,M., Ogasahara,K., Yamagata,Y., Nishikawa,K. and Yutani,K. (1999a) *Protein Eng.*, **12**, 663–672.
Takano,K., Yamagata,Y., Kubota,M., Funahashi,J., Fujii,S. and Yutani,K. (1999b) *Biochemistry*, **38**, 6623–6629.
Thomas,P. and Dill,K. (1996) *J. Mol. Biol.*, **257**, 457–469.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
Yamagata,Y., Kubota,M., Sumikawa,Y., Funahashi,J., Takano,K., Fujii,S. and Yutani,K. (1998) *Biochemistry*, **37**, 9355–9362.

Received January 12, 2001; revised March 28, 2001; accepted April 22, 2001