PROTEINS: Structure, Function, and Bioinformatics 54:271–281 (2004)

Effective Scoring Function for Protein Sequence Design

Shide Liang[2] and Nick V. Grishin[1,2]*
[1]*Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas*
[2]*Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas*

ABSTRACT **We have developed an effective scoring function for protein design. The atomic solvation parameters, together with the weights of energy terms, were optimized so that residues corresponding to the native sequence were predicted with low energy in the training set of 28 protein structures. The solvation energy of non-hydrogen-bonded hydrophilic atoms was considered separately and expressed in a nonlinear way. As a result, our scoring function predicted native residues as the most favorable in 59% of the total positions in 28 proteins. We then tested the scoring function by comparing the predicted stability changes for 103 T4 lysozyme mutants with the experimental values. The correlation coefficients were 0.77 for surface mutations and 0.71 for all mutations. Finally, the scoring function combined with Monte Carlo simulation was used to predict favorable sequences on a fixed backbone. The designed sequences were similar to the natural sequences of the family to which the template structure belonged. The profile of the designed sequences was helpful for identification of remote homologues of the native sequence. Proteins 2004;54:271–281.** © 2003 Wiley-Liss, Inc.

Key words: protein design; Monte Carlo simulation; atomic solvation parameters; profile; homology detection

INTRODUCTION

De novo protein design involves the construction of a sequence not directly related to that of any natural protein and intended to fold into a precisely defined three-dimensional (3D) structure.[1] Recently, protein design has emerged as a powerful method for understanding the underlying principles that dictate protein folding.[2] Most design studies are aimed at the generation of novel hydrophobic cores for proteins.[3–7] In such cases, using hydrophobic residues and considering only packing specificity are sufficient to design well-folded proteins.[8] The first fully automated design and experimental validation of a novel sequence for an entire protein was described by Dahiyat and Mayo.[9] However, the algorithm restricted core positions to hydrophobic residues and surface positions to hydrophilic residues, whereas both hydrophobic and hydrophilic residues were considered at boundary positions. More recent computer programs tend to use no restrictions.[10,11] Solvation energy and amino acid correction baseline factors were used in these programs to avoid considering the protein's unfolded state.

Because solvent affects protein structure, calculating solvation energy is important for protein design. Explicit modeling of protein-solvent interaction is impossible because of prohibitively intensive computation. Two types of approximate methods are frequently used: the atomic solvation parameter (ASP) models[12,13] and structure-based solvation parameters.[14,15] In ASP models, the solvation energy is estimated as the product of the accessibility of the atom and its atomic solvation parameter, which is derived by using the octanol-water or gas-water transfer free energy for each amino acid. The structure-based potential derived from an ensemble of experimentally determined protein structures consists of computing frequencies of structural features and converting these frequencies into free energy.[16] The structure-based potential or the solvation energy calculated by ASPs derived from octanol-water transfer free energy overlap with the terms of common force fields and are, therefore, difficult to use in molecular modeling. Nevertheless, some workers added the solvation energy calculated by the ASPs derived from gas-water transfer free energy directly to the force field of molecular mechanics.[17,18] Others tried to find an appropriate weight between the calculated solvation energy and the force-field terms.[10,19] However, the ASPs may not be optimized when combined with force-field terms because molecular mechanics is a linear combination of simple empirical terms and does not always capture physical reality. For example, when a low electrostatic dielectric constant is chosen, a larger penalty should be used against buried hydrophilic surfaces. Dahiyat and Mayo[5] obtained the optimal ratios for the force-field energy, buried polar, and nonpolar surfaces by fitting them to the experimentally determined stability of designed peptides. Raha et al.[20] determined the strength of ASPs relative to force-field terms from a coarse grid search over combinations of the parameters. The values that gave the best overall results in terms of the resemblance between the designed sequences and natural members of the protein family were accepted. More recently, Das and Meirovitch[21] optimized

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

*Correspondence to: Nick V. Grishin, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9050. E-mail: grishin@chop.swmed.edu

Received 23 January 2003; Accepted 1 June 2003

the ASPs in such a way that the modeled loop structure with the global minimum resembled the X-ray structure.

We previously developed a side-chain modeling program by optimizing the weights of the energy terms.[22] In the course of optimization, for every residue, its side-chain was replaced by varying rotamers, the representative conformations of the amino acid, whereas conformations for all other residues were kept as they appeared in the crystal structure. The weights were optimized to achieve the minimal average root-mean-square deviation (RMSD) between the lowest energy rotamer and the real side-chain conformation on a training set of high-resolution protein structures. Kuhlman and Baker[10] developed a scoring function for protein design by using a similar procedure. In their study, the solvation energy calculated by using the Lazaridis–Karplus solvation model and other energy terms were balanced by a conjugate-gradient-based optimization method.[23] Here, we derived the ASPs together with the weights of empirical energy terms. As a result, the predicted unfolding $\Delta\Delta G$ of T4 lysozyme mutants correlated with their experimental values.

THEORY AND METHODS

The Rotamer Intrinsic Energy and Rotamer Library

The modified backbone-dependent rotamer library of Dunbrack is used in this study.[22,24] Polar hydrogen atoms are added. χ_2 of Ser, Thr, and χ_3 of Tyr are assigned values of $-60°$, 60°, and 180°. Three protonation states of His with the same expected frequencies are considered: $N_{\delta 1}$ protonated, $N_{\epsilon 2}$ protonated, and both. χ_2 of Asn, His, and χ_3 of Gln are flipped 180° to make new rotamers to correct for the lack of defined rotameric states in the Dunbrack library. As a result, the total number of rotamers increases to 412 from 320 in the original library. Given backbone conformation, the intrinsic energies of rotamers are represented by $-\ln(f_1 \times f_2)$. f_1 is the expected rotamer frequency of a particular amino acid.[24] f_2 is the frequency of the amino acid given backbone ϕ,ψ angles, which is derived by statistical analysis of 1344 peptide chains with $<20\%$ sequence identity and > 2.2 Å resolution. The number of each amino acid is counted in 10° by 10° ϕ,ψ bins centered ($-180°$, $-170°$, …, 0°, 160°, 170°). The frequency is calculated as the number of a particular amino acid divided by the number of total amino acids falling in the bin. f_2 represents not only the tendency of an amino acid to adopt the given backbone conformation but also the abundance of the amino acid in native proteins. In regions where amino acids are rarely distributed, the boundaries of the bin are extended by 10° in both directions until the number of total amino acids is >100.

The Scoring Function

The scoring function is a linear combination of the following terms: 1) the contact surface and overlapped volume between the rotamer and surrounding protein atoms[22]; 2) hydrogen bond energy; 3) electrostatic interactions using a distance-dependent dielectric constant; 4) desolvation energy; 5) the rotamer intrinsic energy; 6) disulfide bond energy; and 7) reference value for each amino acid. The parameters of CHARMM polar hydrogen model are used in the energy calculation unless specifically indicated.[25] The definition of hydrogen bond is more stringent than in our previous work[22] to avoid unfavorable hydrogen bonding geometry in the designed strucure:

$$\begin{aligned} &2.3\ \text{Å} < R < 3.5\ \text{Å} \\ &\theta > 100° \\ &|\phi - 109.5°| < 71.5°,\ |\psi - 109.5°| < 71.5° \quad \text{for } sp^3 \text{ acceptor} \\ &\phi > 80°,\ \psi > 80° \quad \text{for } sp^2 \text{ acceptor} \end{aligned} \tag{1}$$

where R is the distance between donor and acceptor of a hydrogen bond, θ is the donor-hydrogen-acceptor angle, ϕ is the hydrogen-acceptor-base angle (the base is the atom attached to the acceptor), and ψ is the donor-acceptor-base angle. The hydrogen bond energy is calculated as: $\cos\theta - \cos 100°$.

Because the bond length is very sensitive to the discrete errors of rotamer analogy, we neglect it in the hydrogen-bond potential. Similarly, the bonding geometry is not considered in the disulfide bridge potential; we only consider it if the modeled cysteine rotamer forms a disulfide bridge with another cysteine residue.[22] The electrostatic interactions between the modeled rotamer and the protein environment are calculated as follows:

$$\begin{aligned} &\Sigma\Sigma(q_i \times q_j)/r^2 \\ &r = R_{ij} \quad \text{if } 0.8 \times (r_i + r_j) \le R_{ij} \le 12 \\ &r = 0.8 \times (r_i + r_j) \quad \text{if } R_{ij} < 0.8 \times (r_i + r_j) \end{aligned} \tag{2}$$

where indices i and j refer to the atoms of the rotamer and the environment, respectively, q_i and q_j are partial charges and r_i and r_j are atom radii from CHARMM. R_{ij} is the distance between the two atoms. The summation is over all atoms i and j for which $R_{ij} \le 12$. Four terms are used for solvation energy: buried hydrophobic surface, buried hydrophilic surface, fraction of buried surface of non-hydrogen-bonded hydrophilic atoms, and solvent-exclusion volume of charged atoms. Solvent-accessible surface area is calculated as described by Zou et al.[26] The probe radius is set to 1.2 Å. The radii of polar hydrogen atoms are set to 1.0 Å. The radii of other atoms are taken from CHARMM and scaled by 0.8. The solvation energy of non-hydrogen-bonded hydrophilic atoms are expressed in a nonlinear way:

$$(S_{buried}/S_{total})^{30} \tag{3}$$

S_{total} is calculated as $4\pi(r + 1.2)^2$. Here, r is the atom radius. S_{buried} is calculated as $S_{\text{total}} - S_{\text{accessible}}$. $S_{\text{accessible}}$ is the solvent-accessible surface. Especially when a buried non-hydrogen-bonded surrounding hydrophilic atom forms a hydrogen bond with the modeled rotamer, the desolvation energy is calculated as $-(S_{\text{buried}}/S_{\text{total}})^{30}$, where S_{buried} is the buried surface calculated absent of the modeled rotamer. The solvent-exclusion volume around charged atoms (H/O atoms

of Asp, Glu, Lys, Arg, and charged His) is calculated similar to the Lazaridis–Karplus model[23]:

$$\sum_i \sum_j \frac{2C_iV_j \exp(-(r_{ij}/\lambda)^2)}{4\pi\sqrt{\pi}\lambda r_{ij}^2} \quad (4)$$

Here, C_i is charge index of H/O atoms of the charged residues. The charge is equally distributed (e.g., the $C_{i=1,2,3}$ of lysine H atoms is 1/3, whereas $C_{i=1,2}$ of O atoms of aspartic acid is 1/2. V_j is the volume of a surrounding non-hydrogen atom). r_{ij} is the distance between the surrounding atom and the charged atom. The correlation length λ is set to 5. The additional solvent-exclusion volume for the surrounding charged residues due to the modeled rotamer is also calculated. Multiple desolvation energy terms may be summed up for a hydrophilic atom. For example, the desolvation energy of buried hydrophilic surface, solvent-exclusion volume of charged atom, and fraction of buried surface of non-hydrogen-bonded hydrophilic atom can be considered for the O atom of aspartic acid side-chain simultaneously.

The training proteins are chosen according to the following criteria. Sequence identity cutoff is set to 20%, the resolution cutoff is set to1.8 Å, and the R-factor cutoff is set to 0.2. A total of 641 chains that met the criteria were downloaded from ftp://fccc.edu/dunbrack/pub/culledpdb on December 26, 2001. Only single-chain proteins with 100–500 residues and containing no incomplete side-chains or ligands were kept. Twenty-eight proteins meeting all the requirements were selected: 153l, 1a12, 1agj, 1ako, 1amm, 1arb, 1bd8, 2sga, 1cem, 1cex, 1chd, 1dhn, 1edg, 1ifc, 1iib, 1koe, 1kpt, 1mla, 1mml, 1npk, 1thv, 1whi, 2baa, 2cpl, 2end, 2pth, 2rn2, and 4eug. The program REDUCE[27] was used to add hydrogen atoms to all proteins. Nonpolar hydrogen atoms were deleted.

The weights of the different energy terms are optimized by 25 cycles of Monte Carlo annealing simulation. The initial temperature is set to 10 and is scaled by 0.8 after each cycle. At maximum, 30,000 substitutions are tried at each cycle. The cycle is terminated sooner, and the simulation goes to the next cycle if 3000 substitutions are accepted.

Evaluation Methods

The free energy of a particular amino acid on the modeled position is calculated as:

$$-\ln(\Sigma \exp(-E(r_i))) \quad (5)$$

where r_i is a rotamer of the amino acid and i is the index of the rotamers. The conformation of the rotamer with the lowest energy is compared to the crystal structure of the native residue if they belong to the same residue type. C_β is included in RMSD calculation and hydrogen atoms are excluded. If the χ_1 angle of a predicted residue is within 40° of the experimental value, the residue is considered correctly predicted until χ_1. χ_{1+2} is considered correctly predicted when both χ_1 and χ_2 are within 40° of their experimental values. Residues with <20% solvent accessibility are considered as core residues.

Predicting Sequences on the Fixed Backbone

All 20 amino acids are considered at each sequence position. Again, we use Monte Carlo annealing simulation to search sequence space. The residues and their conformations are initialized randomly. Then, a residue substitution is made at a selected position. The energies of the old and new residue are calculated by Eq. 5. The new residue is accepted with the probability $\exp[-(E_{new} - E_{old})/T]$. The initial temperature T is set to 10 and is scaled by 0.8 after each cycle. A total of 18 cycles are repeated. For each cycle, 50N substitutions or 5N successful substitutions are carried out. Here, N is the residue number of the modeled protein. The side-chain conformation of the accepted new residue or the old residue is then determined by a random procedure. The probability to accept a rotamer is $\exp[-(E(r_{random}) - E(r_{low}))/T]$, where r_{low} is the rotamer with the lowest energy and r_{random} is the rotamer selected at random. The procedure continues until one rotamer is chosen.

RESULTS AND DISCUSSION

The Derived Scoring Function

Starting from random numbers, the weights of energy terms and reference values of each amino acid were determined by minimizing the sum of the following formula over 5792 positions of the training set of 28 proteins by Metropolis Monte Carlo simulation[28]:

$$-\ln \frac{\sum_i \exp(-E(r_i))}{\sum_n \exp(-E(r_n))} \quad (6)$$

where i was the index of rotamers of the native residue type at a position and $E(r_i)$ was energy of the rotamer r_i; the partition function in the denominator was over all rotamers of 20 amino acids. The formula was similar to that used by Kuhlman and Baker.[10] In the optimization procedure, only one residue is changed at a time, whereas conformations for all other residues were kept in their native conformation. We repeated the optimization procedure four times, and the values of the objective function to be minimized fell in a narrow range (7560.3–7562.2). For the four independent calculations, the variances of weights for several energy terms were very small [e.g., the weight of backbone dependency (lnf) fell in from −0.915 to −0.92]. The weights of several other energy terms were correlated. For example, when the weight of electrostatic interaction increased, the weight of hydrogen bond decreased, which had small overall effect on the objective function. The weights of these correlated energy terms could vary significantly. Similar variation was observed for contact surface, buried hydrophobic solvent-accessible surface, and buried hydrophilic solvent accessible surface. We accepted the parameters when the objective function value was the lowest and the derived scoring function was found to be

$$E = -0.143 \times S_{contact} + 0.724 \times V_{overlap} + 1.72 \times E_{hbond} + 28.6 \times E_{elec} - 0.0467 \times \Delta S_{pho} + 0.0042 \times \Delta S_{phi}$$

$$+ 1.14 \times \Delta(F_{\text{phi}})^{30} + 7.95 \times V_{\text{exclusion}} - 0.919 \times \ln(f_1 \times f_2) - 4.3 \times N_{\text{ssbond}} - \Delta G_{\text{ref}} \quad (7)$$

where S_{contact}, V_{overlap}, E_{hbond}, E_{elec}, ΔS_{pho}, and ΔS_{phi} were contact surface, overlapped volume, hydrogen-bonding energy, electrostatic interaction energy, buried hydrophobic solvent-accessible surface, and buried hydrophilic solvent-accessible surface between the rotamer and other parts of the protein, respectively; F_{phi} was the fraction of buried surface of non-hydrogen-bonded hydrophilic atoms; $\Delta(F_{\text{phi}})^{30}$ meant the difference between the rotamer positioned in the protein environment and the isolated form. $V_{\text{exclusion}}$ was the normalized solvent-exclusion volume around charged atoms; f_1 was the observed frequency of the rotamer, and f_2 was the observed frequency of the amino acid given a backbone conformation; N_{ssbond} was the flag of disulfide bridge(1 or 0); ΔG_{ref} was assumed as the difference between the free energy of the rotamer in solvent and denatured protein (Table I).

It is of interest that F_{phi} played a much more important role than S_{phi}. We thought the buried surfaces of hydrogen-bonded hydrophilic atoms or partially buried hydrophilic atoms had little adverse effect. But the totally buried

TABLE I. Reference Values of the 20 Amino Acids

Residue	ΔG_{ref}	Residue	ΔG_{ref}
Ala	−1.35	Lys	−3.06
Arg	−3.37	Met	−3.88
Asn	−1.73	Phe	−4.78
Asp	−1.32	Pro	−3.01
Cys	−1.78	Trp	−5.12
Gln	−2.12	Val	−3.52
Glu	−2.03	Ser	−0.72
Gly	0	Thr	−1.59
Ile	−4.48	Tyr	−3.93
Leu	−4.27	His	−3.00

TABLE II. Prediction Results for the 28 Training Proteins

% Residue Correct[a]		% χ_1 Correct[b]		% χ_{1+2} Correct[b]	
All	Core	All	Core	All	Core
58.9	78.9	95.3	98.3	88.3	93.2

[a]The percentage of positions in which the observed residues were predicted as the most favorable.
[b]Side-chain conformations were evaluated only when the residues were correctly predicted.

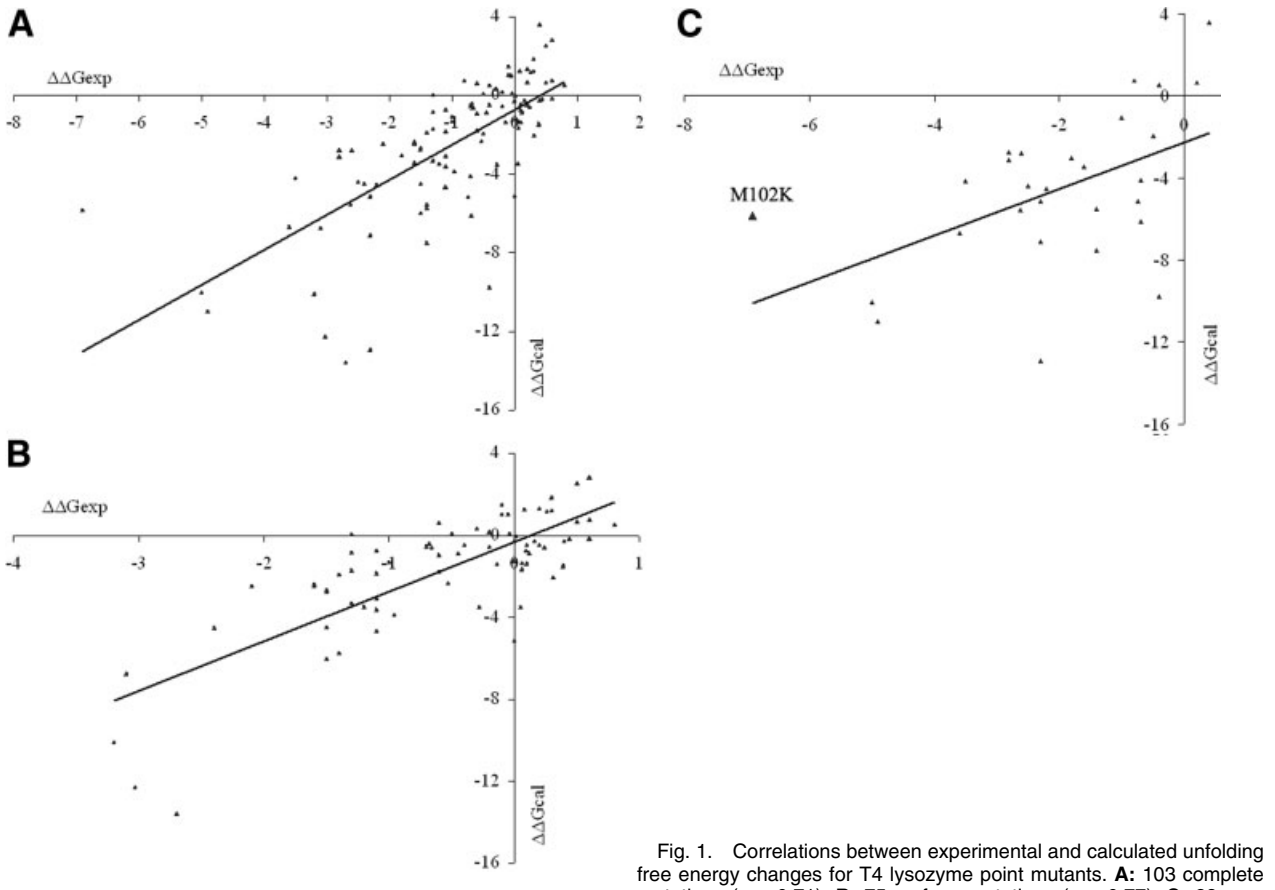

Fig. 1. Correlations between experimental and calculated unfolding free energy changes for T4 lysozyme point mutants. **A:** 103 complete mutations (r = 0.71). **B:** 75 surface mutations (r = 0.77). **C:** 28 core mutations (r = 0.46).

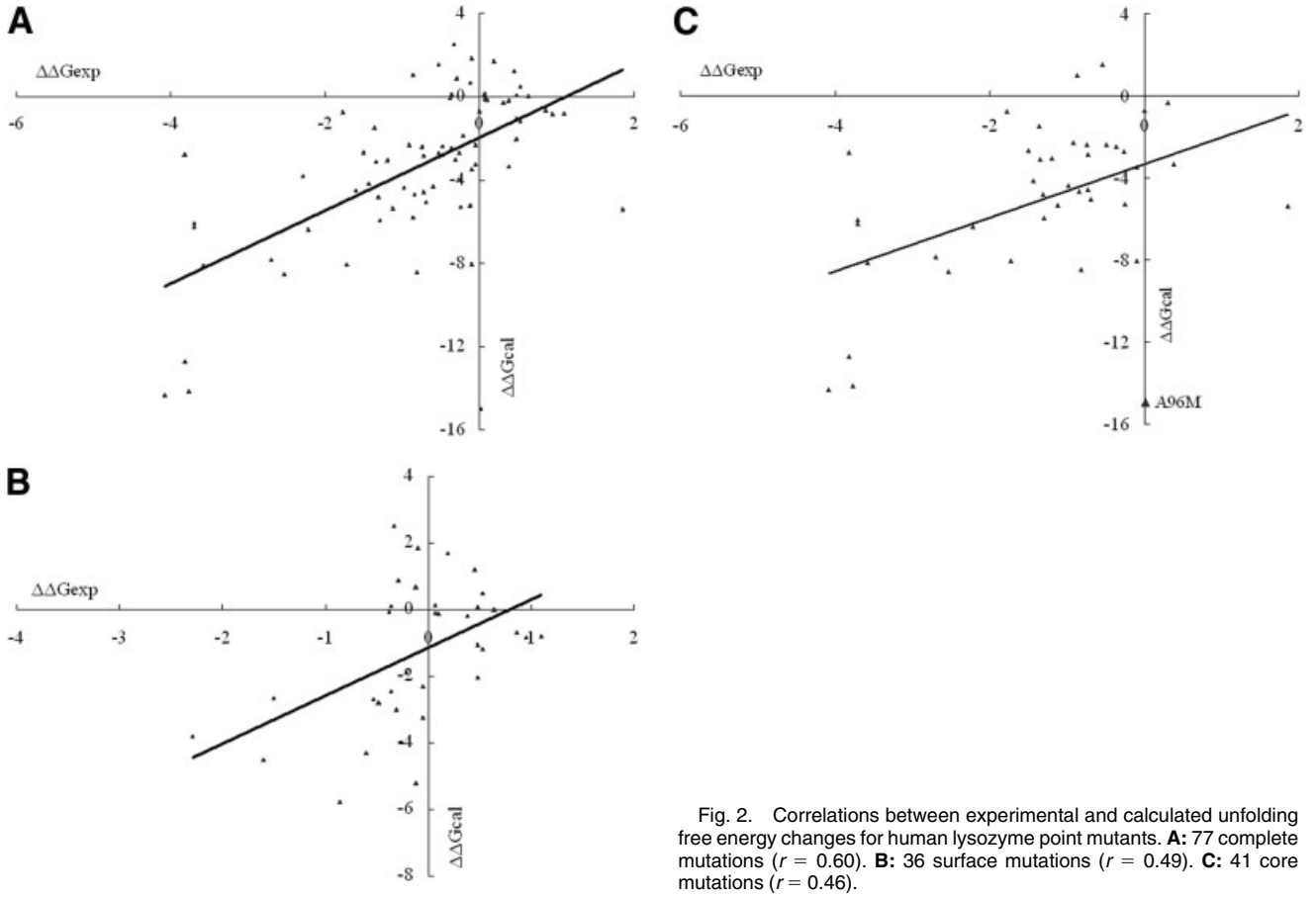

Fig. 2. Correlations between experimental and calculated unfolding free energy changes for human lysozyme point mutants. **A:** 77 complete mutations (r = 0.60). **B:** 36 surface mutations (r = 0.49). **C:** 41 core mutations (r = 0.46).

TABLE III. Effect of Different Solvation Models

Evaluation methods	$(F_{phi})^{30}$	$(F_{phi})^{1}$	Wesson–Eisenberg	Lazaridis–Karplus	No solvation terms
Object function value	7560.3	7739.5	8036.4	8293.6	8530.4
% Residue correct	58.9	58.1	55.9	55.2	54.0
Correlation coefficient (r)	0.71	0.69	0.63	0.62	0.60

F_{phi}, the fraction of buried surface of non-hydrogen-bonded hydrophilic atoms. The expression $(F_{phi})^{30}$ was used for the standard algorithm, whereas linear expression was used in comparison. For both cases, atomic solvation parameters were derived together with the weights of other energy terms. In contrast, the solvation energy calculated by the Wesson–Eisenberg or Lazaridis–Karplus models[18,23] was balanced with other energies as one term. Three criteria were used to evaluate the scoring function: 1) the minimized object function value as calculated by Equation 6; 2) the percentage of positions where the observed amino acids were predicted as the most favorable; and 3) the correlation coefficient between the calculated and experimental unfolding $\Delta\Delta G$ of 103 T4 lysozyme mutations.

non-hydrogen-bonded hydrophilic atoms were quite unfavorable. Some amino acids were predicted more or less frequently than was expected from the composition of the training proteins even when we included the reference value of each amino acid in Eq. 7. For example, glutamine was predicted as the most favorable in 135 searched positions, whereas the 28 training proteins consisted of 230 glutamines totally. We refined the reference values of the 20 amino acids slightly by using Monte Carlo methods so that the sum of the following formula over 20 amino acids was minimized:

$$|N_{\text{predicted}} - N_{\text{native}}| \quad (8)$$

in which N_{native} was the count of a particular amino acid in 28 training proteins and $N_{\text{predicted}}$ was the number of positions where this amino acid was predicted as the most favorable. The weight of each energy term was fixed as obtained from the previous procedure by minimization of Eq. 6. The reference value of each amino acid was initialized with the previous optimized value. The value for Gly remained constant. The sum of Eq. 8 over 20 amino acids was minimized from 786 to 8. The reference values of Cys, Met, His, and Gln were changed most significantly from −3.31 to −1.78, from −4.81 to −3.88, from −3.74 to −3.00, and from −2.78 to −2.12, respectively. Because Eq. 6 could

TABLE IV. Contribution of Each Energy Term to the Calculated Unfolding $\Delta\Delta G$ for 103 T4 Lysozyme Mutations

	RA[a]	$S_{contact}$	$V_{overlap}$	E_{hbond}	E_{elec}	ΔS_{pho}	ΔS_{phi}	$\Delta(F_{phi})^{30}$	$V_{exclusion}$	$\ln(f_1 \times f_2)$	ΔG_{ref}	$\Delta\Delta G_{cal}$	$\Delta\Delta G_{exp}$
I3P	c	−2.92	0.32	−0.00	−0.01	−1.18	0.00	0.05	0.00	1.73	0.61	−1.29	−2.80
I3V	c	−1.32	0.39	−0.00	−0.00	−0.63	0.01	0.02	0.00	1.32	0.40	0.21	−0.40
13Y	c	3.24	−9.02	−0.00	0.02	0.08	−0.07	−0.92	−0.01	0.60	0.23	−5.36	−2.30
M6I	c	−0.39	−2.08	−0.00	−0.02	−0.55	0.00	0.09	0.00	−0.15	−0.25	−3.11	−1.40
K16E	s	−0.08	−0.00	−0.00	0.16	−0.13	−0.00	−0.00	−0.07	−0.19	0.43	0.28	0.50
S38D	s	0.86	−1.56	−0.33	0.95	0.12	−0.03	−0.01	−0.50	0.84	−0.25	−0.07	0.60
N55G	s	−0.18	−0.00	−0.00	0.10	−0.23	0.05	0.04	0.04	0.22	0.72	0.25	−0.60
K60P	s	−0.26	−0.45	−0.30	−1.64	0.01	0.08	0.01	0.53	0.08	0.02	−2.12	0.00
G77A	c	1.93	−0.91	−0.00	−0.00	0.62	−0.01	−0.00	−0.12	0.53	−0.56	1.48	0.40
A82P	s	0.74	−0.13	−0.00	−0.04	0.50	−0.01	0.00	−0.02	−0.29	−0.69	0.21	0.80
R96H	s	1.39	−3.96	−0.79	−0.26	−0.14	0.02	−0.02	0.39	−1.16	0.15	−4.19	−3.20
A98V	c	4.08	−8.50	−0.00	−0.00	0.88	0.01	0.00	−0.01	−0.21	−0.90	−4.56	−4.90
Q105A	s	−2.31	0.58	−0.54	−0.25	−0.74	0.11	0.64	0.13	1.36	0.32	−0.74	−0.60
Q105E	s	−0.10	−0.26	−0.00	−0.11	−0.34	−0.01	0.49	−0.68	0.67	0.04	−0.30	−1.10
Q105G	s	−3.56	0.71	−0.54	−0.25	−1.47	0.11	0.65	0.20	0.79	0.88	−2.49	−1.50
G113A	s	0.77	−0.32	−0.00	−0.00	0.24	−0.00	0.00	−0.00	0.38	−0.56	0.51	0.30
T115E	s	0.24	−0.00	−0.00	0.55	0.16	0.02	0.04	−0.09	−0.15	−0.18	0.77	0.30
N116D	s	0.05	0.32	−0.36	0.62	−0.03	−0.00	−0.17	−0.20	0.82	0.17	1.17	0.60
R119E	s	−0.48	0.13	−0.00	−0.01	−0.62	−0.02	0.01	−0.12	0.45	0.56	0.03	−0.04
Q123E	s	−0.31	−0.00	−0.35	0.69	−0.20	0.00	0.06	−0.38	0.21	0.04	−0.12	0.40
K124G	s	−1.19	0.39	−0.00	−0.18	−0.69	0.04	0.00	0.09	1.43	1.27	0.61	−0.10
V131A	s	−0.59	0.06	−0.00	−0.00	−0.40	0.01	0.03	0.08	0.41	0.90	0.49	0.26
V131D	s	−0.24	0.06	0.32	0.68	−0.40	−0.03	0.03	−0.18	−0.76	0.91	0.52	0.08
V131E	s	−0.27	0.06	−0.00	0.37	−0.17	−0.02	0.00	−0.13	−0.46	0.62	0.55	0.20
V131G	s	−1.04	0.06	−0.00	−0.00	−0.65	0.02	0.03	0.11	−0.14	1.46	−0.17	−0.68
V131I	s	0.17	−0.00	−0.00	−0.00	0.19	0.00	0.00	−0.01	−0.12	−0.40	−0.11	0.16
V131L	s	−0.24	−0.00	−0.00	−0.00	−0.17	0.01	0.03	0.04	0.08	−0.31	−0.31	0.09
V131M	s	0.66	−0.71	−0.00	0.03	0.39	0.01	0.04	0.04	−1.05	−0.15	−0.36	0.12
V131S	s	−0.50	0.06	0.46	−0.09	−0.44	−0.01	0.04	0.08	−0.78	1.16	0.42	−0.05
L133A	c	−3.80	2.21	−0.00	−0.00	−2.66	−0.00	−0.00	0.00	0.28	1.21	−2.77	−3.60
K135E	s	−0.33	−0.00	−0.00	0.12	−0.40	−0.02	−0.00	−0.06	−0.28	0.43	−0.54	−1.00
N144D	s	0.08	−0.00	−0.00	0.60	0.01	0.00	0.01	−0.14	0.52	0.17	1.05	0.50
K147E	s	−0.69	0.71	−0.00	−0.22	−0.55	−0.02	0.01	−0.15	0.14	0.43	−0.22	−0.70
V149C	c	−2.04	0.45	−0.00	−0.04	−0.50	−0.01	−0.00	0.02	−0.71	0.72	−1.89	−2.20
T152S	c	−1.35	0.78	−0.02	−0.05	−0.80	−0.00	0.05	0.01	−0.19	0.36	−1.17	−2.60
R154E	s	−0.64	−0.00	−0.72	−1.62	−0.32	0.02	0.11	−0.06	0.61	0.56	−1.93	−1.10
G156D	c	2.67	−3.05	0.81	0.53	0.27	−0.07	−0.14	−0.62	−1.99	−0.55	−2.13	−2.30
T157A	s	−0.89	0.19	−0.48	0.02	0.20	0.06	0.36	0.06	−0.17	0.10	−0.79	−1.40
T157C	s	−0.55	0.19	−0.48	0.01	0.33	0.05	0.36	0.06	−0.55	−0.08	−0.71	−1.30
T157D	s	−0.53	0.19	−0.48	−0.15	0.17	0.03	0.36	−0.16	−1.00	0.11	−1.50	−1.10
T157E	s	−0.50	0.19	−0.48	0.26	0.31	0.03	0.30	−0.10	−1.13	−0.18	−1.13	−1.50
T157F	s	0.09	−0.45	−0.48	0.02	0.50	0.03	0.33	0.01	−0.45	−1.32	−1.87	−2.40
T157G	s	−1.70	0.19	−0.48	0.02	−0.20	0.08	0.43	0.13	−0.17	0.66	−1.28	−1.10
T157H	s	−0.22	0.06	−0.39	0.29	0.27	0.01	0.36	0.04	−1.23	−0.58	−1.02	−2.10
T157I	s	0.28	−0.39	−0.48	0.02	0.66	0.02	0.00	−0.11	−1.73	−1.20	−2.80	−3.10
T157L	s	−0.23	0.19	−0.48	0.02	0.45	0.03	0.35	−0.03	−0.32	−1.11	−1.37	−1.30
T157N	s	1.28	−1.75	0.04	0.44	0.40	−0.02	0.62	−0.07	−1.18	−0.06	−0.37	−0.45
S44G	s	−0.91	0.06	−0.45	−0.08	−0.19	0.02	−0.01	0.02	0.60	0.30	−0.97	−0.53
S44I	s	0.53	−0.52	−0.45	−0.08	0.69	0.01	−0.00	−0.04	0.85	−1.56	−0.85	0.31
S44K	s	0.51	−0.19	−0.45	0.06	0.59	0.02	−0.00	−0.09	0.14	−0.97	−0.20	0.20
S44L	s	0.42	−0.26	−0.45	−0.08	0.46	0.01	−0.00	−0.03	0.76	−1.47	−0.60	0.39
S44N	s	0.13	−0.00	−0.45	−0.10	0.20	0.00	−0.01	−0.01	−0.08	−0.42	−0.57	−0.14
S44P	s	2.38	−5.19	−1.00	−0.27	0.77	0.03	−0.49	−0.01	−0.04	−0.95	−5.09	−3.03
S44R	s	0.90	−0.13	−0.32	0.14	0.82	−0.06	−0.04	−0.37	−0.24	−1.10	−0.24	0.24
S44T	s	0.05	−0.06	0.04	0.03	0.14	−0.00	0.00	−0.03	0.37	−0.36	−0.02	0.01
S44V	s	0.28	−0.19	−0.45	−0.08	0.41	0.01	−0.00	−0.04	0.99	−1.16	−0.57	0.10
S44W	s	0.26	−0.00	−0.45	−0.07	0.62	0.00	−0.01	−0.06	0.01	−1.83	−1.45	0.05
L46A	c	−4.97	3.37	−0.00	−0.00	−2.20	−0.01	−0.00	0.05	0.24	1.21	−2.30	−2.62
D47A	s	−1.79	0.91	−1.14	−0.86	0.28	0.11	−0.37	0.84	0.57	−0.01	−1.45	−0.28
A49S	s	0.28	−0.00	0.51	0.05	−0.08	−0.04	−0.02	−0.01	−1.11	0.26	0.03	−0.50
T59A	s	−0.57	−0.00	−0.55	0.02	−0.06	0.04	0.01	0.09	−0.07	0.10	−1.11	−1.50

TABLE IV. (Continued)

	RA[a]	$S_{contact}$	$V_{overlap}$	E_{hbond}	E_{elec}	ΔS_{pho}	ΔS_{phi}	$\Delta(F_{phi})^{30}$	$V_{exclusion}$	$\ln(f_1 \times f_2)$	ΔG_{ref}	$\Delta\Delta G_{cal}$	$\Delta\Delta G_{exp}$
T59D	s	−0.37	−0.00	−0.55	−0.23	0.00	0.01	−0.00	−0.15	−0.63	0.11	−1.44	−1.20
T59G	s	−1.04	−0.00	−0.55	0.02	−0.35	0.04	0.09	0.16	0.10	0.66	−0.99	−1.60
T59N	s	1.35	−1.95	−0.04	0.18	0.19	−0.00	0.01	−0.03	−0.74	−0.06	−0.76	−1.10
T59S	s	−0.12	−0.00	0.00	0.05	−0.18	0.00	−0.00	0.07	−0.17	0.36	0.07	−0.20
T59V	s	1.18	−1.82	−0.55	0.02	0.60	0.04	0.02	−0.07	−0.39	−0.80	−1.86	−1.50
D72P	s	1.45	−4.54	−1.48	−1.60	0.84	0.08	−0.01	0.47	−0.13	−0.70	−5.64	−2.70
A73S	s	0.13	−0.13	0.47	0.06	−0.15	−0.04	0.01	−0.01	−1.19	0.26	−0.20	−0.40
V75T	s	−0.92	0.45	0.52	−0.15	−0.62	−0.04	−0.01	0.01	−0.48	0.80	−0.34	−1.30
A82S	s	0.27	−0.00	−0.00	−0.05	−0.00	−0.02	−0.08	−0.00	−0.88	0.26	0.13	−0.30
V87T	c	−1.10	0.32	−0.00	−0.16	−0.72	−0.04	−0.29	0.00	−0.61	0.80	−1.43	−1.60
D92N	s	−0.00	−0.00	−0.31	−1.38	0.23	0.01	−0.76	0.65	−0.51	−0.17	−2.38	−1.40
A93S	s	−0.00	−0.00	−0.00	−0.13	0.01	−0.00	−0.00	−0.00	−1.19	0.26	−0.23	−0.20
A93T	s	−0.00	−0.00	−0.00	−0.10	0.04	−0.00	−0.00	−0.01	−1.01	−0.10	−0.68	0.06
A98S	c	0.74	−1.04	0.08	0.04	−0.39	−0.06	−0.47	−0.00	−1.42	0.26	−1.82	−2.50
L99A	c	−3.61	0.71	−0.00	−0.00	−2.89	−0.01	−0.00	0.00	0.42	1.21	−4.16	−5.00
L99F	c	2.05	−5.97	−0.00	−0.00	0.21	0.00	−0.00	−0.00	−0.39	−0.21	−4.05	−0.40
L99I	c	0.06	−1.10	−0.00	−0.00	−0.31	0.00	0.00	−0.00	−1.07	−0.08	−2.29	−1.40
L99M	c	0.46	−1.30	−0.00	0.01	0.04	0.00	−0.00	0.00	−1.19	0.16	−1.70	−0.70
L99V	c	−1.08	−1.23	−0.00	−0.00	−1.19	0.00	0.00	0.00	0.21	0.31	−2.95	−2.30
M102K	c	0.36	−0.39	−0.00	0.03	−0.24	−0.10	−1.13	−1.08	−0.27	0.34	−2.43	−6.90
M102L	c	0.09	−2.53	−0.00	0.00	−0.57	−0.00	−0.00	−0.01	1.02	−0.16	−2.14	−0.74
T109D	s	0.09	−0.00	−0.00	0.20	0.01	−0.01	−0.01	−0.12	0.15	0.11	0.31	0.60
T109N	s	0.03	−0.00	−0.00	0.13	0.00	−0.01	−0.01	−0.01	−0.45	−0.06	−0.19	0.10
V111I	c	1.73	−3.57	−0.00	−0.00	0.48	−0.00	−0.00	−0.00	−1.19	−0.40	−2.54	−0.69
A130S	c	0.53	−0.45	0.70	−0.20	−0.11	−0.04	−0.03	−0.01	−1.36	0.26	−0.44	−1.00
A134S	s	0.36	−0.19	0.50	−0.12	−0.07	−0.03	0.01	−0.00	−0.51	0.26	0.43	−0.10
V149T	c	−0.96	0.19	0.51	−0.08	−0.79	−0.05	−0.23	0.03	−0.66	0.80	−1.15	−2.80
T151S	s	−0.91	0.32	−0.04	−0.05	−0.58	−0.00	0.04	0.06	0.06	0.36	−0.63	0.39
F153A	c	−5.54	4.80	−0.00	−0.00	−3.30	−0.00	−0.02	0.01	0.89	1.42	−1.74	−3.50
F153I	c	−1.44	0.84	−0.00	−0.00	−0.97	0.01	−0.01	−0.01	0.54	0.13	−0.81	−0.50
F153L	c	−1.58	1.82	−0.00	−0.00	−0.86	−0.00	−0.01	−0.00	0.71	0.21	0.27	0.20
F153M	c	−1.01	2.08	−0.00	−0.02	−0.59	0.00	−0.01	0.01	−0.51	0.38	0.31	−0.80
F153V	c	−2.97	2.27	−0.00	−0.00	−1.78	0.00	−0.01	−0.01	0.68	0.52	−1.27	−1.80

[a]RA, relative accessibility. Residues with <20% solvent accessibility are considered as core residues (c). Other residues are attributed to surface residues (s). $\Delta\Delta G_{cal}$, the calculated unfolding $\Delta\Delta G$; $\Delta\Delta G_{exp}$, the experimental unfolding $\Delta\Delta G$. The definition of each energy term is the same as in Eq. 7. Disulfide bond energy is excluded because all the 103 mutations do not involve in disulfide bond. The difference between the lowest energy rotamers of native and mutated residues is calculated by using a single component of Eq. 7 and considered as the contribution of the corresponding energy term to $\Delta\Delta G_{cal}$. Because the free energy of a particular amino acid on the modeled position is calculated from all its rotamers in a nonlinear way as expressed in Eq. 5, the sum of the contributions of all energy terms is not equal to the $\Delta\Delta G_{cal}$. All the calculated values are adjusted by dividing 2.41, the slope of the regression line between $\Delta\Delta G_{cal}$ and $\Delta\Delta G_{exp}$ [Fig. 1(B)].

bias toward statistically common amino acids, the reference value of uncommon residues such as Cys, Met, and His was overestimated in the previous optimization procedure. Refinement of the reference values was helpful to correct amino acid composition bias. The percentage of positions, in which the native residues were predicted as the most favorable, slightly dropped from 59.1 to 58.9% when the refined references were used. The adjustment of reference values seemed to affect mainly those positions, in which the native residues did not have an exceptionally low energy. With the refined reference values, we expected the designed protein sequences to be more native-like. The ΔG_{ref} of glycine was set to zero, and the values of other amino acids were all negative (Table I). Because hydrophobic interactions were always favorable, residues with large hydrophobic group tended to have low reference values. Table II listed the prediction results for the 28 training proteins using the refined reference values. For positions where the native residues were predicted as the most favorable, the side-chain conformations were also predicted with very high accuracy (χ_1 was correctly predicted at 95.3% positions and χ_{1+2} was correctly predicted at 88.3% positions) even though we did not consider side-chain conformation in the optimization procedure.

Estimation of Mutational Energy Changes

We tested the scoring function by estimating the relative stability of a mutant protein to the wild type. The difference between the energy of the mutated and native residues was calculated ($-(E_{mutated} - E_{native})$) and compared to the difference between experimentally determined unfolding ΔG of the two proteins ($\Delta\Delta G = \Delta G_{mutant} - \Delta G_{native}$). We assumed the backbone conformation and side-chain conformations of the surrounding residues were not changed because of mutation. The energy of the mutated residue was calculated by using the crystal structure of wild type with the side-chain of a native residue deleted at the modeled position. Fifty point mutations of T4 lysozyme wild

type (PDB code 3lzm) and 53 point mutations of its C54T and C97A mutant (PDB code 1l63), which had been used by Ota et al.[29] to test their scoring function, were used in this study. Although different crystal structures were used for the two sets of mutation data, the 103 mutations were combined for the regression analysis; 28 of them were identified as core mutations. The correlation coefficients were 0.77, 0.46, and 0.71 for surface, core, and total mutations, respectively. The contribution of each energy term to the calculated unfolding $\Delta\Delta G$ for the 103 mutations was shown in Table IV. The low correlation coefficient for core mutations was mainly due to the assumption of a fixed backbone. In real proteins, when small residues were mutated to large residues, such as I7Y, L99F, and V111I, backbone shifts occurred and the calculated $\Delta\Delta G$ values were much lower than the experimental data. In other studies, significant backbone relaxation has been clearly demonstrated in a number of core variants, and sequences found to be experimentally stable were sometimes predicted as unallowable by using a fixed backbone model.[30] We adopted the assumption of fixed backbone despite this deficiency because it reduced complexity and computation time dramatically. For the mutated large residues on the surface, the situation was different because they could protrude out into solvent and avoid changing backbone conformation. Therefore, the calculated unfolding $\Delta\Delta G$ for surface residues correlated strongly with the experimental values, and the regression line passed exactly through the origin with a slope of 2.41 [Fig. 1(B)]. According to our visual analysis of the mutant crystal structures, neighboring residues usually did not change their rotamer states even for core mutations. Thus, it was not necessary to repack the neighboring side-chains in this study. In addition, the discrete errors of rotamers could reduce the correlation coefficients if none of the rotamers in the library closely resembled the real side-chain conformation for the native or mutated residue. Ota et al.[29] used the crystal structure of the mutant protein to calculate unfolding ΔG_{mutant}. They got higher correlation coefficients for the core mutations than for total mutations (0.76 and 0.69, respectively). Nonetheless, our scoring function predicted more accurately the behavior of the total mutations without knowing the crystal structures of mutant proteins. Ota et al. obtained the correlation coefficient of 0.58 for another testing protein, human lysozyme (PDB code 1rex). We also obtained a similar coefficient of 0.60 for that protein (Fig. 2). The lower correlation was partially due to A96M mutation in the hydrophobic core. Backbone atoms at the mutated position were shifted by 0.8 Å to accommodate larger methionine in crystal structure of the mutant protein. Excluding this mutation, we got a correlation coefficient of 0.67. In addition, the stability change of surface mutations was very small, which also resulted in a lower correlation coefficient [Fig. 2(B)].

The electrostatic interactions for core and surface residues were treated equally in our scoring function. Therefore, the interaction energy of a buried salt bridge could be underestimated. As a result, the scoring function undervalued the desolvation energy of charged atoms in a compensative way. For example, the experimental unfolding $\Delta\Delta G$ of T4 lysozyme mutant M102K was much lower than the calculated one [Fig. 1(c)]. Excluding the M102K mutation, the correlation coefficients for core and total mutations increased to 0.58 and 0.74, respectively. The desolvation energy of the mutated lysine, which was located in the hydrophobic core and did not form any salt bridge or hydrogen bond, was obviously underestimated. Fortunately, this seemed not to affect the prediction of favorable sequences on a fixed backbone. We found no buried charged residues not involved in a salt bridge in our sequence design experiments.

The prediction results significantly depended on which solvation energy model was adopted (Table III). The scoring function performed much better when atomic solvation parameters were derived in the optimization procedure together with the weights of other energy terms. If solvation energy was calculated by the Wesson–Eisenberg or Lazaridis–Karplus models[18,23] as one term and balanced with other energy terms, we only obtained a small improvement over using no solvation terms. Nonlinear expression of the fraction of buried surface of non-hydrogen-bonded hydrophilic atoms was superior to linear expression even though the formula (F_{phi})[30] was a bit arbitrary. It is of interest that when the objective function (Eq. 6) was minimized to a low value on different solvation models, the scoring function also frequently predicted the observed residue as the most favorable and the calculated unfolding $\Delta\Delta G$ strongly correlated with experimental values (Table III). The cooperative behavior of the three criteria indicated that our evaluation methods were reasonable. To our surprise, addition of solvation terms showed significant improvement only for core residues in estimating mutational energy changes. We obtained a correlation coefficient of 0.81 for T4 lysozyme surface residues when no solvation terms were used. Chakravarty and Varadarajan[31] argued that atoms just below the protein surface could undergo large fluctuations and transiently come into contact with solvent. In previous studies, we demonstrated that addition of solvation terms showed little improvement for side-chain modeling.[22] The solvation energy of core residues was not sensitive to conformation changes, whereas the solvation energy of surface residues might be difficult to evaluate.

More recently, Guerois et al.[32] developed a computer algorithm to estimate the importance of the interactions contributing to the stability of proteins. The weights of different energy terms were fitted by using the experimental $\Delta\Delta G$ values of a training set comprising 339 single-point mutants in 9 different proteins. The crystal structure of native protein was used to calculate energy. The problems related to the modeling of the mutated side-chain were avoided by considering only mutations involving the deletion of groups in the side-chain and the substitution of groups such as E→Q, D→N, T→V, or the reverse of these. The correlation coefficient between the predicted and experimental data was 0.7 for the training set. The predictive power of the methods was then tested by using a blind test database of 625 mutants in 27 proteins, and a similar coefficient 0.73 was obtained. For the same training and testing sets, we obtained correlation of 0.67 and 0.7,

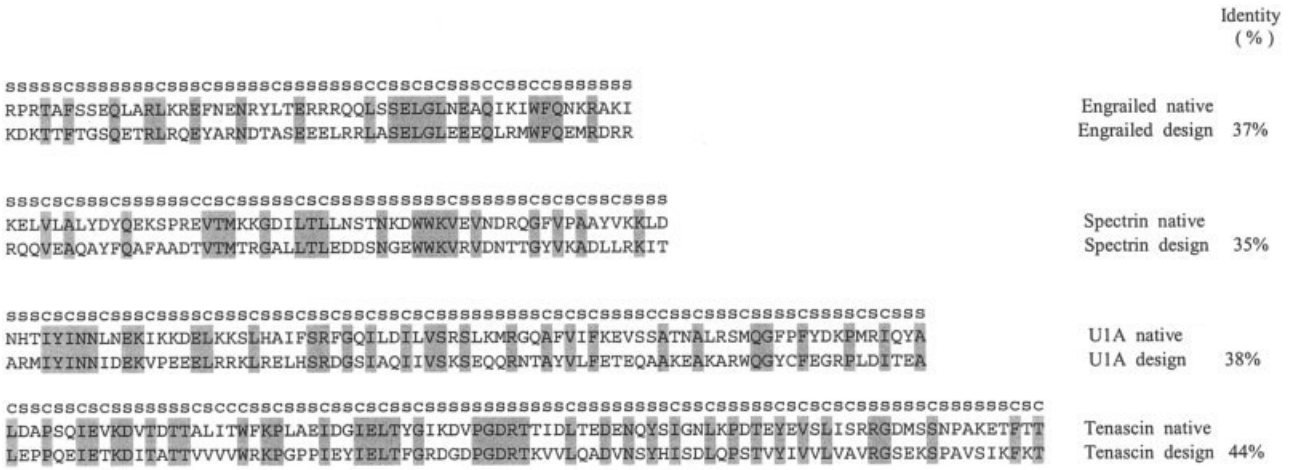

Fig. 3. Comparison of designed and native sequences. For each motif, we repeated 20 calculations. The sequence with the highest identity to its corresponding native sequence was listed. The residues of the native sequence were divided into two groups: surface residues (s) and core residues (c). Residues exposing <20% of their surface area to solvent were considered core residues.

respectively. For core residues, the correlation coefficients were 0.67 and 0.64, respectively. The training and testing sets did not contain mutations of type X→Y (where Y was an amino acid larger than X). Thus, backbone and surrounding side-chain relaxations could be reduced, and the prediction ability was improved for core residues. Because our scoring function was neither designed nor trained to achieve a high correlation coefficient on a mutation database, it was not surprising that we obtained a slightly lower correlation. Furthermore, we used rotamers to calculate the energy of native and mutated residues, which made it more practicable to use our algorithm for all kinds of mutations. The use of rotamers also made it easy to use our scoring function for protein design.

Sequence Design on a Fixed Protein Backbone

The derived scoring function, combined with Monte Carlo simulation methods, was used to predict energetically favorable protein sequences given a fixed backbone. Only the backbone structure was used as input, and the program was allowed to choose any natural amino acid for each modeled position. The quality of the designed proteins was difficult to evaluate without experimental characterization. Identity percentage between a predicted and native sequence could be used as a simple assessment for the predictive ability of the design algorithm.[20] The design experiments were conducted on the representative structures of four distinct folds, engrailed (PDB code 1enh), spectrin (1shg), U1A (1urn), and tenascin (1ten), which have been used by Raha et al. We repeated 20 calculations for each protein backbone. On average, the identity between designed and native sequences was 26%, 27%, 33%, and 37%, respectively. Figure 3 listed the native sequence and the most similar designed one. The extent of similarity was remarkable. In contrast, Raha et al. obtained sequence identities ranging from 24% to 28%. Kuhlman and Baker[10] tested their program on 108 proteins; 27% of all of the designed residues were identical to the native amino acid, and 51% of the core residues in the design sequences were identical to the naturally occurring residues. We obtained the mean identity of 59% for the core residues of the four tested proteins.

In addition, Raha et al.[20] evaluated their designed sequences by using a profile derived from a multiple-sequence alignment of the family to which the template structure belonged. They also determined the solvation parameters in their scoring function by a coarse search for the combination of parameters that gave the best overall profile score for the four motifs. Here, we calculated the profile score with a position-specific score matrix produced by PSI-BLAST search of Jul 2002 nr database.[33] The native sequence of the template structure was used as query, and the search was repeated until convergence. For the template structures of engrailed, spectrin, and U1A, most of our designed sequences had a similar or slightly higher profile score than those designed by Raha et al. (Fig. 4). For the tenascin motif, the profile scores of our designed sequences were significantly higher. In all cases, the profile scores of the designed sequences were lower than their native sequences. The profile scores of the designed sequences were then compared with those of natural sequences that belonged to the same family as the template structure. The members of each family and the sequence alignment were downloaded from the Pfam web site http://pfam.wustl.edu/index.html.[34] Redundant sequences and sequences with long gaps were excluded. For each motif, >300 sequences remained. Only positions that aligned to the native sequence of the design template were counted in the profile score, and the scores of each family fell in a wide range, which overlapped the scope of the profile scores of designed sequences. For tenascin, the profile scores of the designed sequences were higher than most of the natural sequences. This might be due to the large size of tenascin and more core residues that tended to be predicted identical to native residues. Furthermore, the interaction sites of the tenascin family were variable, whereas other families had specific ligand-binding sites

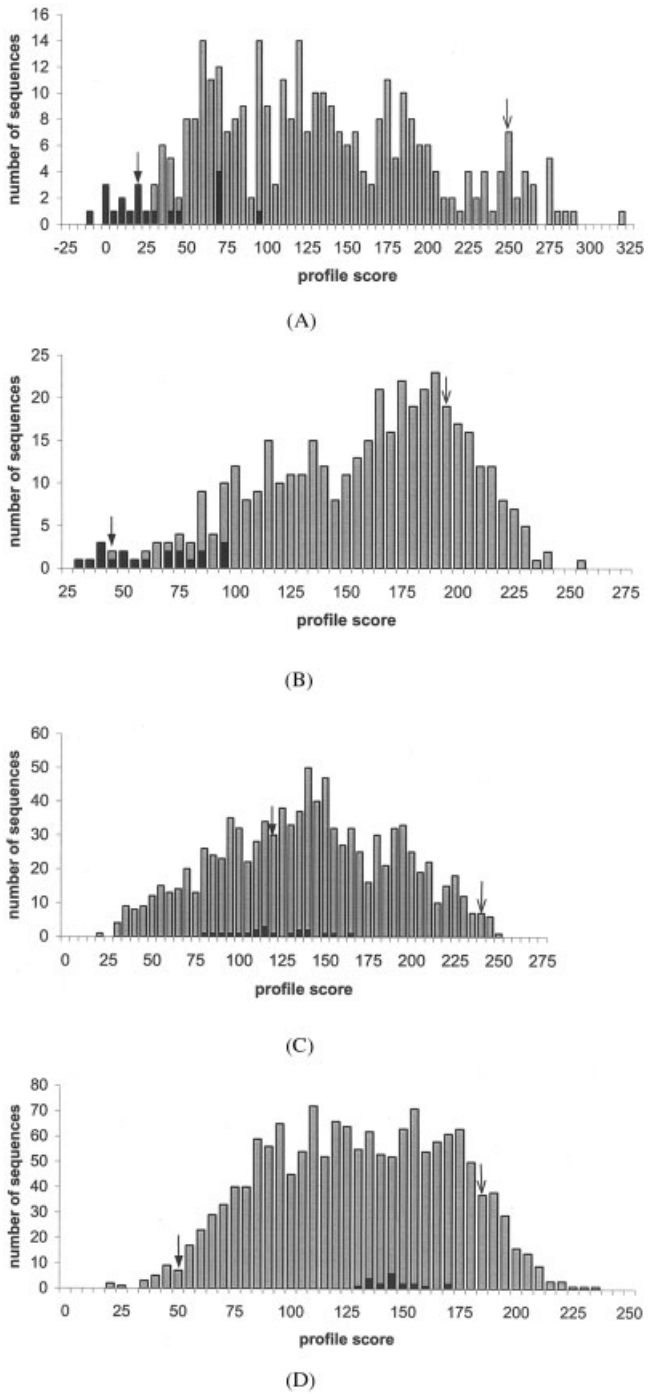

Fig. 4. Distributions of profile scores for designed and natural sequences. The sequences designed on the template backbone were shown in dark. The natural sequences of the protein family to which the template structure belongs were shown in gray. The line arrowhead indicated the profile score of the native sequence of the template structure. The solid arrowhead indicated the profile score of the sequence designed by Raha et al.[20] **A:** Engrailed and homeobox family. **B:** Spectrin and SH3 family. **C:** U1A and RNA recognition motif. **D:** Tenascin and fibronectin type III family.

and conserved functional residues, which could possibly lower the profile score of the sequences designed by using structural considerations only.

We then investigated if the profile of the designed sequences was helpful for homology detection; 100 sequences were designed on the backbone of tenascin. To make the designed sequences more diverse and improve the quality of the profile, we perturbed the backbone conformation before the sequence prediction. The ϕ or ψ angle was rotated $<1°$ at a randomly chosen position. If the position of any backbone atom was shifted for >0.3 Å, the rotation was rejected. The small perturbation of the backbone allowed us not to consider the backbone potential. This procedure was repeated 30,000 times. Again, we used the PSI-BLAST program to identify homologous sequences. The native sequence of tenascin was used as the query, and the designed sequences were used to generate position-specific score matrix. The expectation value was set to 0.1. The search retrieved 392 sequences in nr database. In comparison, we ran PSI-BLAST in the standard way. The close homologue sequences found in each search round were used to generate a score matrix for the next round. The procedure was repeated until converged and 734 sequences were retrieved; 326 sequences found by the two searches were the same. Some sequences found exclusively by the profile of designed sequences were confirmed to be the remote homologues of tenascin, such as chitinase B. Both tenascin and chitinase A N-terminal domain belonged to the immunoglobulin-like β-sandwich fold in SCOP database.[35] If we ran PSI-BLAST for one round and the standard scores were used for each position, only 118 sequences were found. This definitely demonstrated that the profile of the designed sequences was useful for homology detection, especially for protein families that did not contain many sequences and had the crystal structures available for some members.

CONCLUSIONS

We have developed a scoring function for protein design. The formula of each energy term was carefully designed, and the weight was optimized so that the native residue was predicted energetically favorable at each position of the training proteins. The success of our scoring function was demonstrated by predicting mutant changes in the stability for testing proteins. The correlation coefficient between the calculated and experimentally determined unfolding $\Delta\Delta G$ for 103 T4 lysozyme mutants was 0.71. When the scoring function was used for sequence design on a fixed backbone, the designed sequences were similar to the natural sequences of the family to which the template structure belonged. We also found that calculating solvation energy was important for protein design. Atomic solvation parameters should be derived together with the weights of other energy terms. Solvation energy calculated by solvent-accessible surface model may not be suitable for atoms just below the protein surface. New models to calculate solvation energy quickly and accurately were necessary for protein design.

REFERENCES

1. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779–819.

2. Pokala N, Handel TM. Review: protein design—where we were, where we are, where we're going. J Struct Biol 2001;134:269–281.
3. Kono H, Doi J. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. Proteins 1994;19:244–255.
4. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Sci 1995;4:2006–2018.
5. Dahiyat BI, Mayo SL. Protein design automation. Protein Sci 1996;5:895–903.
6. Lazar GA, Desjarlais JR, Handel TM. De novo design of the hydrophobic core of ubiquitin. Protein Sci 1997;6:1167–1178.
7. Johnson EC, Lazar GA, Desjarlais JR, Handel TM. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. Struct Fold Des 1999;7:967–976.
8. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Curr Opin Struct Biol 1999;9:509–513.
9. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
10. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97: 10383–10388.
11. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. J Mol Biol 2000;301:713–736.
12. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
13. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
14. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
15. DeBolt SE, Skolnick J. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng 1996;9:637–655.
16. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
17. Schiffer CA, Caldwell JW, Stroud RM, Kollman PA. Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case. Protein Sci 1992;1:396–400.
18. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
19. Wilson C, Mace JE, Agard DA. Computational method for the design of enzymes with altered substrate specificity. J Mol Biol 1991;220:495–506.
20. Raha K, Wollacott AM, Italia MJ, Desjarlais JR. Prediction of amino acid sequence from structure. Protein Sci 2000;9:1106–1119.
21. Das B, Meirovitch H. Optimization of solvation models for predicting the structure of surface loops in proteins. Proteins 2001;43:303–314.
22. Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci 2002;11:322–331.
23. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
24. Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
25. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculation. J Comput Chem 1983;4: 187–217.
26. Zou X, Sun Y, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the Generalized-Born model. J Am Chem Soc 1999;121:8033–8043.
27. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
28. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys 1953;21:1087–1092.
29. Ota M, Isogai Y, Nishikawa K. Knowledge-based potential defined for a rotamer library to design protein sequences. Protein Eng 2001;14:557–564.
30. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. J Mol Biol 1999;290:305–318.
31. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Struct Fold Des 1999;7:723–732.
32. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
33. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
34. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–266.
35. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.