Designing Protein β-Sheet Surfaces by Z-Score Optimization

Arthur G. Street
Division of Physics, Mathematics and Astronomy, California Institute of Technology, MC 147-75, Pasadena, California 91125

Deepshikha Datta
Division of Biology, California Institute of Technology, MC 147-75, Pasadena, California 91125

D. Benjamin Gordon
Division of Chemistry and Chemical Engineering, California Institute of Technology, MC 147-75, Pasadena, California 91125

Stephen L. Mayo
Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, MC 147-75, Pasadena, California 91125

(Received 27 September 1999)

Studies of lattice models of proteins have suggested that the appropriate energy expression for protein design may include nonthermodynamic terms to accommodate negative design concerns. One method, developed in lattice model studies, maximizes a quantity known as the "Z-score," which compares the lowest energy sequence whose ground state structure is the target structure to an ensemble of random sequences. Here we show that, in certain circumstances, the technique can be applied to real proteins. The resulting energy expression is used to design the β-sheet surfaces of two real proteins. We find experimentally that the designed proteins are stable and well folded, and in one case is even more thermostable than the wild type.

PACS numbers: 87.15.Aa, 87.10.+e, 87.14.Ee, 87.15.Cc

Much effort in the field of computational protein design is directed towards developing a potential function to rank the compatibility of amino acid rotamer sequences with a target structure [1]. In a "protein design cycle" [2,3], the potential function is developed by cycling between experiment and simulation, so that the computational potential ideally approaches nature's "true" potential. This technique has had some remarkable recent successes [4,5].

The approach nevertheless rests on a controversial assumption. Rotamer sequences are threaded onto the target structure, and the sequence with the lowest energy (as determined by the potential function) is reported as the best sequence for that structure. It is conceivable, though, that in some circumstances this sequence will not adopt the desired ground state structure. An extreme example is provided by imagining that the true potential function is one that benefits only hydrophobic contacts (and hydrophobic-polar and polar-polar interactions contribute zero energy) [6]. Then, for any target structure, an all-hydrophobic sequence must be one of the best sequences. This sequence, of course, is not likely to fold specifically to the target structure—some polar residues ought to be included to characterize the surface of the molecule. Overcoming this problem involves introducing nonthermodynamic considerations to the design procedure, collectively known as "negative design" [7].

There are a number of schemes proposed to implement negative design, often specifically to solve the problem of the example in the last paragraph (or variations on it based on the Ising model of ferromagnetism). Perhaps the simplest is to use a fixed sequence composition, that is, to hold the total number of hydrophobic and polar residues constant [8]. Even with this constraint, however, designed sequences are frequently found to fold to alternative structures of lower energy than the target structure [9,10]. Alternatively, instead of minimizing the potential function, it is possible to choose a sequence to maximize the occupation probability of the sequence on the target structure [11,12].

Other approaches employed in lattice model studies involve adding nonthermodynamic terms to the potential function. One method is to introduce a "clamping potential" to force the molecule into the target structure, and then to minimize the difference between the clamping potential and the true potential [13,14]. Another approach involves the addition of a penalty for exposing hydrophobic surface area [15].

Negative design is thus important, at least in lattice model studies with simple potential functions and a limited set of amino acids [16,17]. For real proteins and more physical potential functions, negative design can be necessary to guarantee the correct multimeric state of designed proteins [18]. A penalty for exposing hydrophobic surface area has also been shown to improve the designability of real proteins [5,19].

In this Letter we take another approach to determining the optimal potential function for protein design, in which we maximize the energy gap between a low energy sequence known to fold to the target structure, and the average energy of an ensemble of random sequences threaded onto the target structure [20]. In a lattice simulation, the

 0031-9007/00/84(21)/5010(4)$15.00

desired true potential can be selected manually and the protein folding problem can be solved. Thus a sequence S, whose ground state structure is the target structure, can be determined and its energy calculated. If the distribution of energies of the random sequences is assumed to be Gaussian, the success of the test potential for protein design is measured by the energy gap between the mean of the distribution and the energy of sequence S, normalized by the standard deviation of the distribution. This quantity is known as the Z-score of the sequence S on the target structure. The test potential is then adjusted to maximize the Z-score.

Chiu and Goldstein applied the method to a $3 \times 3 \times 3$ lattice model, using statistically derived pair potentials as the true potential. They found that the potential generated by maximizing the Z-score across many structures led to significantly better success at solving the protein design problem than the true potential. Here we show that the technique does not transfer readily to real proteins in their entirety. Nevertheless, we show that the technique can be applied to certain subsections of proteins. In particular, we use it to design the β-sheet surfaces of the $\beta 1$ immunoglobulin-binding domain of streptococcal protein G (GB1) and of a variant of poplar apoplastocyanin (PCV) with the metal-binding site removed through the mutations His37 to Val and Cys84 to Ala.

One of the key assumptions of the lattice model method of Chiu and Goldstein [20] is that the energies of random sequences threaded onto the target structure form a Gaussian distribution. It would be surprising if this assumption were to hold for real proteins. In particular, one would expect that placing random amino acid side chains in the core of a protein would typically lead to unresolvable steric clashes, especially since the modeled backbone of the target structure is held rigid. Indeed, Fig. 1a shows the distribution of potential energies of random sequences threaded onto the core of GB1. The distribution is clearly not Gaussian, with most sequences yielding enormous energies. A Gaussian distribution may be achievable by using a statistically derived pair potential instead of an atomistic van der Waals potential, but designs using pair potentials have not yielded uniquely characterizable folded states [21].

When only surface residues are considered, the situation is improved. For α-helix and β-sheet surface residues of GB1, the distribution of energies of random sequences is close to Gaussian (Fig. 1b). Thus it appears that on the surface, even randomly selected amino acids are always able to find suitable rotamers that avoid severe steric interference. The Z-score analysis may therefore provide some insight into the appropriate potential function for α-helix and β-sheet surface design, provided one can find an appropriate sequence with which to calculate the Z-score. In lattice models, one knows the true potential function and can exhaustively search all conformations to solve the protein folding problem [8]. Hence the Z-score of a structure could be calculated using the lowest energy sequence whose ground state structure is the target structure.

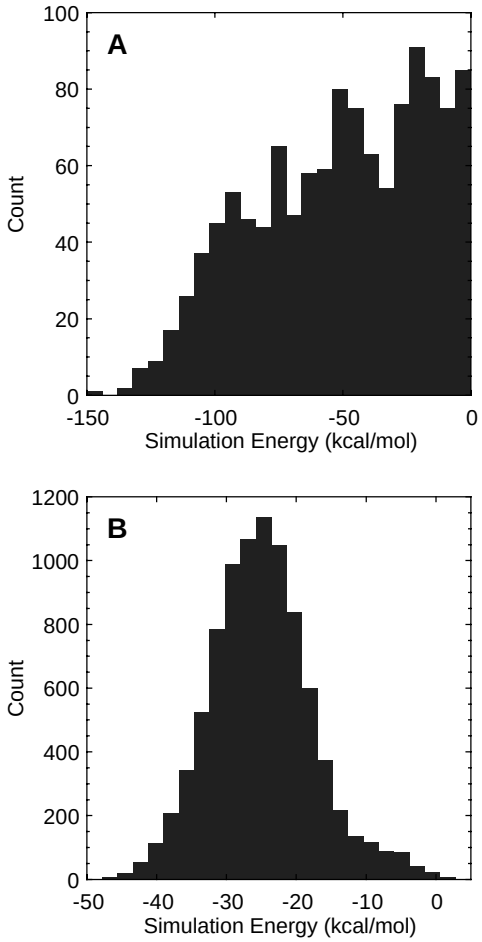

FIG. 1. The actual distribution of energies of various subsets of the real protein GB1, using the potential function derived from the protein design cycle (Table I). Side chain conformations are optimized using the dead-end elimination theorem [26–28]. (a) The core (only the 2.5% lowest energy sequences are shown), and (b) the β-sheet surface. The core residues of GB1 are positions 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54.

In our application of the theory to real proteins, we do not have this luxury. Instead, given an experimentally determined structure, we use the protein's wild-type sequence to calculate its Z-score. In essence, the method then chooses the potential function which locates the protein's wild-type sequence as far as possible down the tail of the distribution of energies.

Since a number of successful computational redesigns of α-helical surfaces have been reported [22], we chose to examine the Z-score technique on the β-sheet surface, where there have been few successful computational protein design efforts. Negative design issues are also expected to play a larger role in β-sheet design [23]. In particular, we chose to apply the technique to the eight β-sheet surface positions (4, 6, 15, 17, 42, 44, 53, and 55) of GB1 which are not involved in stabilizing interactions with neighboring turns, and to the seven β-sheet surface positions (18, 20, 79, 81, 93, 95, and 97) on one face of PCV.

The computational potential function, E, included van der Waals interactions, E_{vdW} [19,24], electrostatics, E_{elec},

a hydrogen bonding potential, E_{HB} [22], a bias for secondary structure propensity, E_{SS} [22], and solvation energies. The solvation energies were a benefit for burial of hydrophobic surface area, $A_{\mathrm{np}}^{\mathrm{buried}}$, a penalty for burial of polar surface area, $A_{\mathrm{polar}}^{\mathrm{buried}}$, a penalty for exposure of hydrophobic surface area, $A_{\mathrm{np}}^{\mathrm{exposed}}$ [25], and a further penalty for polar hydrogen burial, E_{phb} [22].

$$E = \upsilon E_{\mathrm{vdW}} - \sigma_{\mathrm{np}} A_{\mathrm{np}}^{\mathrm{buried}} + \xi_{\mathrm{np}} A_{\mathrm{np}}^{\mathrm{exposed}} + \sigma_{\mathrm{p}} A_{\mathrm{polar}}^{\mathrm{buried}} + \frac{1}{\epsilon} E_{\mathrm{elec}} + D E_{\mathrm{HB}} + P E_{\mathrm{phb}} + E_{\mathrm{SS}}(N). \quad (1)$$

The magnitude of the van der Waals interactions, υ, was held fixed and the relative magnitudes of the other seven energy terms (σ_{np}, ξ_{np}, σ_{p}, ϵ, D, P, and N as shown, where E_{SS} is an exponential function of N) were chosen to maximize the Z-score.

The Z-score was calculated using 4000 random sequences to determine the energy distribution of the potential function on the structure, resulting in an uncertainty in the Z-score of ± 0.04. The random sequences were composed of the polar amino acids Ser, Thr, Asp, Asn, Glu, Gln, Lys, and Arg, as well as the hydrophobic amino acids Ala, Val, and Ile. The results were surprisingly robust to changes in the set of amino acids considered.

In contrast to the case in lattice models, real amino acid side chains may adopt many different conformations or rotamers. The energy of a given amino acid sequence on a structure is thus calculated by minimizing the energy across all possible rotamer configurations, using the dead-end elimination theorem [26–28]. For this procedure a backbone-dependent rotamer library was used [29], in which the first dihedral angle of each hydrophobic amino acid rotamer was expanded ± 1 standard deviation about the mean value [22].

The resulting potential functions are shown in Table I. For GB1, the maximum Z-score is 2.6, i.e., the wild-type sequence is assigned an energy lower than 99.5% of possible sequences. For PCV, the maximum Z-score is 2.2. Also shown in Table I is the potential function built up over many experiments using the protein design cycle, which has been successful for core design and α-helix surface design [2]. The Z-score optimized potential functions exhibit some interesting common features. The hydrophobic burial benefit, which is the main embodiment of the hydrophobic effect, has disappeared. This reflects the relative lack of importance of hydrophobic burial on the surface of proteins.

The most dramatic difference from the protein design cycle potential is the increased importance of electrostatic interactions. The value of the dielectric constant used in the protein design cycle is similar to that of water, and leads to electrostatic interactions being deemphasized. Although salt bridges are not encouraged, the hydrogen bonding potential from the protein design cycle is quite strong (an ideal hydrogen bond receives a benefit of 8.0 kcal/mol). The Z-score optimized dielectric constant is an order of magnitude smaller, closer to unity. This is justifiable because we are considering effects at the molecular level, where the assumptions behind the use of the dielectric constant break down. The screening effect of the solvent is also approximated by using a distance attenuated Coulomb potential [24].

We then applied this potential function towards protein design, using a combination of dead-end elimination and branch and terminate [30] to find the lowest energy sequence for each β-sheet surface. The resulting GB1 variant, GB1-Z1, is a fivefold mutant of the wild-type protein. A cluster of theonines and Ile6 have been replaced by cross-strand salt bridge networks, Asp42 to Arg55, and Arg6 to Glu53 to Lys44. Thr17 and the wild-type salt bridge formed by Lys4 and Glu15 are left unchanged. Cross-strand salt bridges might be expected to contribute to β-sheet formation and stability, and surface networks of salt bridges are postulated to be a stabilizing factor in hyperthermophilic proteins [31]. The resulting PCV variant, PCV-Z1, is a threefold mutant of the wild-type protein, Ser81 to Ile, Val93 to Lys, and Thr97 to Lys. The modeled side chain configurations are shown in Fig. 2. Again, the impact of the electrostatic term is clear, with a salt bridge network formed by Glu18, Lys95, Lys97, and Glu79.

TABLE I. Potential functions determined through different methods. The energy terms considered are shown in Eq. (1). The van der Waals energy scale factor υ was held fixed. A potential function has been developed using the protein design cycle [2] and has been successful for core and α-helix surface design, in particular. The Z-score method applied to the β-sheet surface of PCV and of GB1 yield new potential functions. Also shown are the ranges over which each parameter may be changed while keeping the Z-score within 5% of its maximum (when the other parameters are kept fixed).

Energy term	Design cycle	PCV	Range	GB1	Range
van der Waals υ	1.0	1.0	n.a.	1.0	n.a.
np burial σ_{np} (kcal/mol/Å^2)	0.05	0.0	0.0–0.01	0.0	0.0–0.02
np exposure ξ_{np} (kcal/mol/Å^2)	0.05	0.10	0.04–0.16	0.06	0.02–0.08
Polar burial σ_{p} (kcal/mol/Å^2)	0.0	0.0	0.0–0.04	0.03	0.01–0.06
Dielectric ϵ	40.0	4.0	2.0–6.0	4.0	2.0–6.0
H-bond D (kcal/mol)	8.0	1.0	1.0–8.0	6.0	1.0–8.0
Polar H burial P (kcal/mol)	2.0	9.0	6.0–15.0	3.0	1.0–7.0
Secondary structure bias N	n.a.	1.0	0.0–1.4	1.4	0.8–1.6

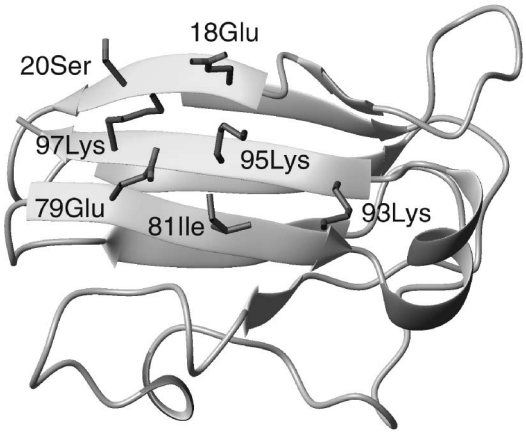

FIG. 2. View of the seven designed positions on the β-sheet surface of PCV-Z1.

The designed proteins were constructed experimentally by standard molecular biological techniques. Their far UV circular dichroism spectra overlay those of the wild-type proteins. The melting temperature of GB1-Z1 was determined to be 71 °C. The melting temperature of GB1 is 87 °C. The designed protein is almost as stable as the wild-type protein and appears to fold to the correct structure. Although the literature contains many examples of alterations to the β-sheet surface of GB1, we know of no instances resulting in greater than wild-type stability. This is the first example of a well-formed, many-stranded β sheet designed through purely computational means.

The results for PCV-Z1 were even more impressive. The melting temperature of PCV-Z1 was determined to be 64 °C, compared to the melting temperature of PCV of 56 °C. The designed protein is thus even more stable than the natural one. To our knowledge, this is the first time a natural protein's stability has been increased by computationally redesigning its β-sheet surface.

We have designed two stable protein β-sheet surfaces using different potential functions. Indeed, further application of the technique to other proteins suggests yet different potentials may be appropriate. This supports the belief that there may be alternative routes taken by nature to stabilize protein surfaces, and which may also be taken in *de novo* design [32]. Of course, one test of this proposal is to use the potential derived from one protein to design the β-sheet surface of another, and preliminary results in this regard appear promising (unpublished data). A further advantage of the approach outlined in this Letter is that it could lead to a faster turnaround time for protein design, since it optimizes the potential function with less frequent recourse to experiment.

[1] D. B. Gordon, S. A. Marshall, and S. L. Mayo, Curr. Opin. Struct. Biol. **9**, 509 (1999).
[2] A. G. Street and S. L. Mayo, Structure **7**, R105 (1999).
[3] B. I. Dahiyat and S. L. Mayo, Protein Sci. **5**, 895 (1996).
[4] B. I. Dahiyat and S. L. Mayo, Science **278**, 82 (1997).
[5] S. M. Malakauskas and S. L. Mayo, Nat. Struct. Biol. **5**, 470 (1998).
[6] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[7] H. W. Hellinga, Proc. Natl. Acad. Sci. U.S.A. **94**, 10 015 (1997).
[8] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[9] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[10] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[11] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **80**, 2237 (1998).
[12] F. Seno, C. Micheletti, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).
[13] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).
[14] T. Kurosky and J. M. Deutsch, J. Phys. A **27**, L387 (1995).
[15] S. Sun, R. Brem, H. S. Chan, and K. A. Dill, Protein Eng. **8**, 1205 (1995).
[16] G. M. Crippen, Proteins **26**, 167 (1996).
[17] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Proteins **32**, 80 (1998).
[18] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber, Science **262**, 1401 (1993).
[19] B. I. Dahiyat and S. L. Mayo, Proc. Natl. Acad. Sci. U.S.A. **94**, 10 172 (1997).
[20] T. L. Chiu and R. A. Goldstein, Protein Eng. **11**, 749 (1998).
[21] Y. Isogai, M. Ota, T. Fujisawa, H. Izuno, M. Mukai, H. Nakamura, T. Iizuka, and K. Nishikawa, Biochemistry **38**, 7431 (1999).
[22] B. I. Dahiyat, D. B. Gordon, and S. L. Mayo, Protein Sci. **6**, 1333 (1997).
[23] M. H. Hecht, Proc. Natl. Acad. Sci. U.S.A. **91**, 8729 (1994).
[24] S. L. Mayo, B. D. Olafson, and W. A. Goddard III, J. Phys. Chem. **94**, 8897 (1990).
[25] A. G. Street and S. L. Mayo, Folding & Des. **3**, 253 (1998).
[26] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters, Nature (London) **356**, 539 (1992).
[27] R. F. Goldstein, Biophys. J. **66**, 1335 (1994).
[28] D. B. Gordon and S. L. Mayo, J. Comput. Chem. **19**, 1505 (1998).
[29] R. L. Dunbrack and M. Karplus, J. Mol. Biol. **230**, 543 (1993).
[30] D. B. Gordon and S. L. Mayo, Structure **7**, 1089 (1999).
[31] A. H. Elcock, J. Mol. Biol. **284**, 489 (1998).
[32] M. H. J. Cordes, A. R. Davidson, and R. T. Sauer, Curr. Opin. Struct. Biol. **6**, 3 (1996).