Protein Science (2000), *9*:403–416. Cambridge University Press. Printed in the USA.

A new approach to the design of uniquely folded thermally stable proteins

XIN JIANG,[1] HANY FARID,[2] ERNIE PISTOR,[1] AND RAMY S. FARID[1]

[1]Department of Chemistry, Rutgers, The State University of New Jersey, Newark, New Jersey 07102
[2]Department of Computer Science, Dartmouth College, Hanover, New Hampshire 03755

(RECEIVED September 9, 1999; FINAL REVISION December 7, 1999; ACCEPTED December 8, 1999)

Abstract

A new computer program (CORE) is described that predicts core hydrophobic sequences of predetermined target protein structures. A novel scoring function is employed, which for the first time incorporates parameters directly correlated to free energies of unfolding (ΔG_u), melting temperatures (T_m), and cooperativity. Metropolis-driven simulated annealing and low-temperature Monte Carlo sampling are used to optimize this score, generating sequences predicted to yield uniquely folded, stable proteins with cooperative unfolding transitions. The hydrophobic core residues of four natural proteins were predicted using CORE with the backbone structure and solvent exposed residues as input. In the two smaller proteins tested (Gβ1, 11 core amino acids; 434 cro, 10 core amino acids), the native sequence was regenerated as well as the sequence of known thermally stable variants that exhibit cooperative denaturation transitions. Previously designed sequences of variants with lower thermal stability and weaker cooperativity were not predicted. In the two larger proteins tested (myoglobin, 32 core amino acids; methionine aminopeptidase, 63 core amino acids), sequences with corresponding side-chain conformations remarkably similar to that of native were predicted.

Keywords: computational; conformational entropy; convergence temperature; heat capacity; Metropolis Monte Carlo; protein design; simulated annealing

Considerable attention has recently been directed at the design of protein sequences that fold into predetermined structures (Tuchscherer et al., 1998). Target structures include functional de novo designed proteins, thermally or chemically stable variants of natural enzymes, and enzymes with altered functionality. Methods for producing proteins with targeted structure and function include iterative design and characterization (DeGrado et al., 1989), combinatorial synthesis (Kamtekar et al., 1993), and computational approaches (Street & Mayo, 1999). With recent dramatic increases in computing speeds, computational approaches aimed at generating sequences that stabilize desired target structures have become more feasible as evidenced by several noteworthy successes (Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Dahiyat & Mayo, 1996, 1997a; Su & Mayo, 1997; Harbury et al., 1998; Kono et al., 1998). Computer programs have been written that "score" sequences by utilizing one or a combination of van der Waals potential energy, solvation energy, amino acid propensities for secondary structure, electrostatic energy, and hydrogen bond energy. Various optimization algorithms, such as simulated annealing (Hellinga & Richards, 1994), dead-end elimination (Desmet et al., 1992), Metropolis Monte Carlo sampling (Holm & Sander, 1991; Lee & Levitt, 1991), and genetic algorithms (Desjarlais & Handel, 1995), are then employed to optimize the score. However, despite tremendous efforts, there are only a few examples of designed proteins that fold into predetermined target structures with properties that characterize well-ordered stable proteins. These properties include (1) structural uniqueness; (2) maximum stability at room and physiological temperatures as indicated by large free energies of unfolding (ΔG_u) at these temperatures; (3) optimal thermal stability as indicated by high values for T_m, the temperature at which $\Delta G_u = 0$ (for monomeric proteins); and (4) highly cooperative unfolding transitions, indicating that the protein folds into a highly ordered structure. One of the primary barriers to producing proteins with these key properties is the difficulty in deconvoluting all of the terms that contribute to protein thermodynamics and structural uniqueness. In addition, it is necessary to calculate these energy terms for both the folded and unfolded states, a task that clearly presents a significant challenge. The use of terms such as amino acid secondary structure propensities or solvation energy avoids the need to explicitly consider the unfolded state; however, it is not clear whether consideration of these terms alone can lead to accurate sequence predictions. Another significant complication is that individual terms that contribute to protein thermodynamics are correlated to one another in a complex nonlinear way, making it difficult to determine whether a particular term should be increased or decreased to produce stable pro-

Reprint requests: Ramy S. Farid, Department of Chemistry, Rutgers University, 73 Warren Street, Newark, New Jersey 07102-1811; e-mail: rfarid@newark.rutgers.edu.

teins. The goal is, therefore, to define criteria that are directly correlated to experimentally measurable thermodynamic parameters that define protein stability, to rapidly calculate these criteria, and then to efficiently optimize the criteria. To this end, a new computer program (CORE) has been developed that predicts sequences and side-chain conformations of hydrophobic core residues by scoring sequences with criteria that are shown to directly correlate to the free energy of unfolding (ΔG_u), melting temperature (T_m), cooperativity, and structural uniqueness. We recently reported on the de novo design of a thermally stable synthetic protein using CORE (Jiang et al., 1997). This paper describes CORE in more detail and presents sequence prediction results of native proteins that validate the scoring function and the algorithms employed to optimize the score.

Results

Free energy of unfolding and thermal stability

A primary goal in protein design is the generation of proteins that exhibit large free energies of unfolding (ΔG_u) and high melting temperatures (T_m). The analysis presented below reveals that design of proteins exhibiting maximal ΔG_u and T_m values can be accomplished by maximizing the heat capacity change of unfolding (ΔC_p), as long as hydrophobic amino acids are modified exclusively, and the protein backbone structure is fixed. It is demonstrated below that only under these conditions is it possible to quantitatively express ΔG_u and T_m as a function of ΔC_p.

The temperature dependence of ΔG_u can be obtained by considering the temperature dependence of the enthalpy (ΔH_u) and entropy (ΔS_u) of unfolding, expressed in Equations 1 and 2, respectively, where T is temperature, T_R and T_R' are reference temperatures, and ΔC_p is assumed to be temperature independent. It has been experimentally demonstrated that ΔC_p exhibits little dependence on temperature from 20 to 80 °C (Privalov & Gill, 1988). Above 80 °C, ΔC_p decreases, approaching zero at ~130 °C.

$$\Delta H_u(T) = \Delta H(T_R) + \Delta C_p \times (T - T_p) \quad (1)$$

$$\Delta S_u(T) = \Delta S(T_R') + \Delta C_p \times \ln(T/T_R'). \quad (2)$$

When normalized to the number of residues (N_{res}), the value of ΔH_u for proteins converges to approximately the same value ΔH_u^* at a common temperature T_H^* (Privalov & Khechinashvili, 1974). The same is true for ΔS_u; at a common temperature T_S^*, proteins have approximately the same entropy of unfolding per residue ΔS_u^*. These convergence temperatures are the temperatures at which the apolar contributions to both $\Delta H°$ and $\Delta S°$ are zero (see below for more details). The convergence parameters can be incorporated into Equations 1 and 2 to yield Equations 3 and 4, respectively.

$$\Delta H_u(T) = N_{res}\Delta H_u^* + \Delta C_p \times (T - T_H^*) \quad (3)$$

$$\Delta S_u(T) = N_{res}\Delta S_u^* + \Delta C_p \times \ln(T/T_S^*). \quad (4)$$

The free energy of unfolding can therefore be expressed by Equation 5, in which the heat capacity is now expressed as the average per residue value, $\Delta\bar{C}_p$.

$$\Delta G_u(T) = N_{res}\{\Delta H_u^* - T\Delta S_u^* + \Delta\bar{C}_p[(T - T_H^*) - T\ln(T/T_S^*)]\}. \quad (5)$$

Model compound and protein studies have yielded the following values for the convergence parameters (Murphy & Freire, 1992): $\Delta H_u^* = 1{,}350 \pm 0.11$ cal·mol^{-1}·res^{-1}, $\Delta S_u^* = 4.30 \pm 0.1$ cal·mol^{-1}·K^{-1}·res^{-1}, $T_H^* = 373.5 \pm 6$ K, $T_S^* = 385 \pm 1$ K. Therefore, Equation 6 expresses the free energy of unfolding as a function of temperature, the number of residues, and the heat capacity change of unfolding per residue.

$$\Delta G_u(T) = N_{res}\{1.35 - 0.0043\cdot T + \Delta\bar{C}_p[(T - 373.5) - T\ln(T/385)]\}. \quad (6)$$

Figure 1 shows the dependence of ΔG_u on temperature for a hypothetical 100 amino acid protein calculated using Equation 6. Three curves are presented for different values of $\Delta\bar{C}_p$ that bracket the range of values found for natural proteins, 10–20 cal·mol^{-1}·K^{-1}·res^{-1} (Murphy et al., 1990). The plot clearly shows that T_m increases monotonically with increasing $\Delta\bar{C}_p$. ΔG_u above 22 °C (T_i) also increases with increasing $\Delta\bar{C}_p$. T_i depends on the value of the convergence temperatures T_H^* and T_S^*. Given the range in values for T_H^* and T_S^*, the highest value for T_i is ~54 °C, occurring when T_S^* is at the lower limit (384 K) and T_H^* is at the upper limit (379.5 K). Therefore, it is conceivable (but less likely) that in-

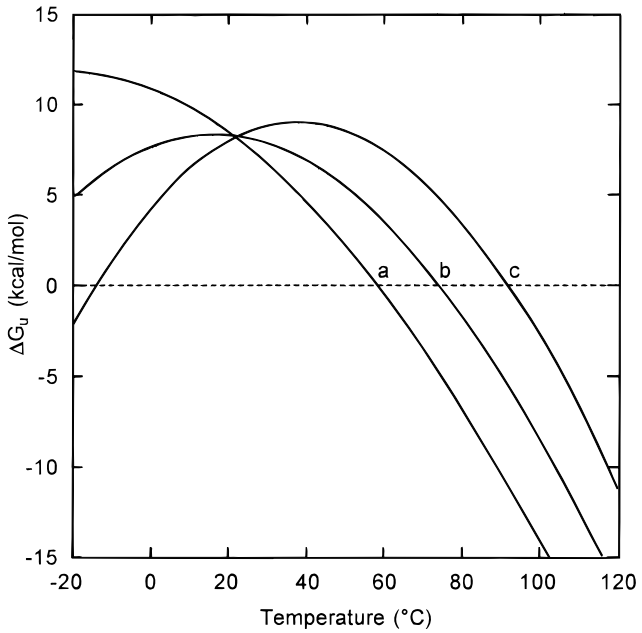

Fig. 1. Plot of free energy of unfolding (ΔG_u) vs. temperature generated using Equation 6 for three hypothetical 100 amino acid monomeric proteins that have the same backbone structure, but different values of the heat capacity change per residue ($\Delta\bar{C}_p$). Equation 6 is valid if $\Delta\bar{C}_p$ is modulated by altering hydrophobic amino acids without concomitant changes in the backbone structure or burial of polar groups. This condition assures that enthalpy and entropy convergence will occur (see text for details). T_m values (the temperature at which $\Delta G_u = 0$) increase with increasing $\Delta\bar{C}_p$. ΔG_u above 22 °C also increases with increasing ΔC_p. (**a**) $\Delta\bar{C}_p = 0.01$ kcal/mol/K/res, (**b**) $\Delta\bar{C}_p = 0.015$ kcal/mol/K/res, (c) $\Delta\bar{C}_p = 0.02$ kcal/mol/K/res.

creases in $\Delta\bar{C}_p$ would result in increases in ΔG_u only at temperatures higher than 54 °C rather then 22 °C, as indicated in Figure 1.

It is particularly useful to express the difference in $T_m(\Delta T_m)$ for two proteins at a given set of values of the convergence parameters as a function of the change in $\Delta\bar{C}_p(\Delta\Delta\bar{C}_p)$. However, it is not possible to express the true relationship between ΔT_m and $\Delta\Delta\bar{C}_p$ because of the transcendental form of Equation 5. Derivation of an empirical relationship reveals that ΔT_m is linearly dependent on $\Delta\Delta\bar{C}_p$ and on a weighted linear combination of the convergence parameters as indicated by Equation 7, where $a = -0.0127$, $b = 3.03$, $c = -0.114$, $d = 0.0925$, $e = 14.7$.

$$\Delta T_m = (a\Delta H_u^* + b\Delta S_u^* + cT_H^* + dT_S^* + e) \times \Delta\Delta\bar{C}_p\ (\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}). \quad (7)$$

Given the range in values of the convergence parameters, Equation 7 can be simplified to Equation 8, which expresses a simple linear dependence of ΔT_m on $\Delta\Delta\bar{C}_p$.

$$\Delta T_m = (3.6 \pm 2.5) \times \Delta\Delta\bar{C}_p\ (\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}). \quad (8)$$

Equation 8 reveals that a ΔT_m value as large as 6.1 times the difference in heat capacity change per residue (expressed in $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}\cdot\text{res}^{-1}$) can be observed. This is put into perspective by considering that a mutation of a single buried Ala residue to a Phe in a hypothetical 50 amino acid protein [$\Delta\Delta\bar{C}_p = 1$ cal$\cdot$mol$^{-1}\cdot$K$^{-1}\cdot$res^{-1} (see below)] would result in an increase in T_m of as much as 6 °C, provided that the mutation does not induce a change in the backbone structure.

ΔC_p can be expressed as a function of the change of buried apolar and polar surface area upon unfolding as shown in Equation 9 (Murphy & Gill, 1991).

$$\Delta C_p = 0.45\cdot\Delta ASA_{ap} - 0.26\cdot\Delta ASA_{pol} \quad (9)$$

where ΔC_p is expressed in units of $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$, and ΔASA_{ap} and ΔASA_{pol} are the changes in solvent accessible surface area in Å^2 upon protein denaturation for apolar and polar areas, respectively. The constants are taken from solid model compound studies (Murphy & Gill, 1991) and have been shown to be reasonably accurate in estimating values of ΔC_p for proteins (Murphy et al., 1992). If amino acid side chains are fully buried, and the protein backbone is fixed (such that solvent exposure of main-chain atoms remains constant), ΔC_p for each amino acid is constant; therefore, the ΔC_p for core residues can simply be calculated from a lookup table of individual ΔC_p values. Using values for ΔASA obtained from Privalov and Makhatadze (1990), the following ΔC_p values for buried hydrophobic amino acids were calculated: Ala = 30.2, Met = 41.5, Val = 52.6, Tyr = 53.6, Leu = 61.6, Ile = 63.0, Phe = 78.7, and Trp = 80.7 $\text{cal}\cdot\text{mol}^{-1}\cdot\text{K}^{-1}$.

The quantitative expressions presented above, relating ΔC_p to ΔG_u and T_m, are valid only if enthalpy and entropy convergence actually occurs. The convergence of thermodynamic quantities at some temperature will occur when there are two dominant interactions (e.g., apolar and polar) that independently contribute to the thermodynamics, and when one of these interactions is constant (Murphy & Freire, 1992; Murphy & Gill, 1990, 1991). Because only the hydrophobic (apolar) contribution is modified by mutation of core residues without significant effect on the polar contribution (such as hydrogen bonds), convergence will be observed (Murphy & Freire, 1992). Under these conditions, the apolar contribution to ΔH° is zero at T_H^* and the apolar contribution to ΔS° is zero such that ΔH_u^* and ΔS_u^* represent the polar contribution to the thermodynamics at T_H^* and T_S^*, respectively. It is expected, therefore, that within a series of proteins in which the backbone structure is maintained and only hydrophobic core residues are modified, the convergence parameters will remain constant such that Equation 5 is valid and an increase in ΔC_p will, in fact, be associated with an increase in T_m as well as an increase in ΔG_u above 20 to 50 °C.

Protein structural uniqueness and cooperativity

Native proteins generally adopt highly ordered, densely packed, unique structures in solution. This is evident from the cooperative unfolding transitions observed for natural proteins. It has been shown previously that the primary source of cooperativity in proteins is extensive side-chain contacts that give rise to expansive networks of interacting amino acids (Freire & Murphy, 1991; Murphy et al., 1992). Extensive side-chain contacts lead to highly ordered structures with constrained side-chain mobility and thus low side-chain conformational entropy. Therefore, side-chain conformational entropy was chosen as a criteria by which sequences of uniquely folded proteins with cooperative unfolding transitions can be predicted.

The conformational entropy change upon folding of a single amino acid (ΔS_{conf}) can be calculated using the Boltzmann equation given by Equation 10 (Shenkin et al., 1996).

$$\Delta S_{conf} = R(\ln W^* - \ln W) \quad (10)$$

where R is the gas constant and W is the number of conformations adopted in the unfolded state, taken here as the total number of possible rotamers (n_{rot}) derived from a suitable rotamer library (see Methods). Because of the lack of information for the unfolded state, all rotamers in the unfolded state are treated as energetically equivalent. W^* is the number of allowable rotamers in the folded state weighted by the probability of each rotamer existing in the folded structure. It is convenient to think of W^* as the effective number of rotamers in the folded sate. Equation 11 expresses W^* as a function of p_i and the fractional population of each rotamer state i in the folded state.

$$W^* = \exp\left(-\sum_i^{n_{rot}} p_i \ln(p_i)\right). \quad (11)$$

The conformational entropy change for the hydrophobic core of a protein is the average of ΔS_{conf} for each core residue. ΔS_{conf} for Ala is zero, because the conformational entropy in the folded and unfolded state are the same.

Bumps

Steric compatibility is undoubtedly the most important criteria necessary to stabilize a target structure. Amino acid side chains that exhibit unfavorable van der Waals interactions will most likely induce changes in the backbone structure to relieve steric crowding. Most commonly a van der Waals energy calculation is employed in an attempt to accurately represent steric compatibility

with a target structure. This calculation is necessarily conducted on a structure generated using a so-called rotamer library that defines discrete allowable side-chain rotamers. However, if the structure is not first energy minimized, the van der Waals energy is subject to potentially large errors, and its value may not accurately represent steric compatibility. A detailed discussion of the use of van der Waals to define steric compatibility is presented in Discussion. For computational efficiency, a hard-sphere model (Ponder & Richards, 1987; Shenkin et al., 1996) is preferred to represent side-chain contacts. In this model, an unfavorable interatomic contact (bump, B) occurs when the distance between a side-chain atom and any other atom in the protein is shorter than a given allowed value. Allowable interatomic distances are obtained by summing appropriate combinations of the following distances obtained from the Tripos 5.2 Force Field (Clark et al., 1989): C = 1.34, H = 0.95, O = 1.2, N = 1.3, S = 1.5 Å. For example, an H···H nonbonding contact of less than 1.90 Å is considered a bump.

Description of the program CORE

The three criteria described above, heat capacity change of unfolding (ΔC_p), conformational entropy change of folding (ΔS_{conf}), and bumps, are incorporated into a protein design program called CORE. CORE is designed to predict sequences and corresponding side-chain conformations of hydrophobic core residues yielding thermally stable proteins with high cooperativity. This is accomplished by selecting sequences with zero hard-sphere bumps, maximum ΔC_p, and minimum ΔS_{conf}. As described in detail above, zero bumps assures steric compatibility with a target structure, maximizing ΔC_p leads to proteins with maximal T_m and ΔG_u above T_i, and minimizing ΔS_{conf} gives rise to uniquely folded proteins with optimal cooperativity. The number of bumps (B) and the value of ΔS_{conf} for a given sequence are calculated using a previously written program (Shenkin et al., 1996) that utilizes Metropolis-driven simulated annealing (Metropolis et al., 1953) to determine side-chain conformations associated with a minimum number of bumps in a fixed backbone structure. Subsequent low-temperature Monte Carlo sampling of side-chain rotamers, at the final simulated annealing Metropolis temperature, yields p_i values (see Equation 11). As mentioned above, ΔC_p for a sequence is obtained from a lookup table of individual ΔC_p values. The values for B, ΔS_{conf}, and ΔC_p are incorporated into a *Score* for the sequence, determined from a linear combination of these three quantities (Equation 12).

$$Score = \alpha B + \beta \Delta S_{conf} - \gamma \Delta C_p. \quad (12)$$

To assure that only sequences with $B = 0$ are selected, α is set to an arbitrarily large number. Since sequences with large values for ΔC_p are necessarily associated with large buried hydrophobic surface areas and therefore large side-chain groups, it is expected that large amino acids will exist in a smaller number of rotamers, thus exhibiting low values for ΔS_{conf}. This should lead to a correlation between ΔS_{conf} and ΔC_p, making the relative *Scores* for sequences somewhat insensitive to the values of β and γ. Therefore, β and γ are both initially set to unity; however, if cooperativity is a dominant design goal, β is set to a higher value, and if high T_m values are desired, γ is set higher relative to β. To predict sequences with optimized *Score*, Metropolis-driven simulated annealing is used again, this time to produce a single protein structure (i.e., sequence and corresponding side-chain conformations). Low-temperature Monte Carlo sampling, initiated from the simulated annealing sequence, is then used to search for sequences near the global minimum. Monte Carlo sampling is terminated when no new sequences are predicted. Sequences predicted during Monte Carlo sampling are ranked by *Score* such that the top sequences are predicted to exhibit optimized thermal stability and cooperativity.

Sequence prediction

A good test of a protein design program is its ability to accurately predict the sequence of naturally occurring proteins. Four proteins, ranging in size from 56 to 259 amino acids, were chosen to validate the underlying principles guiding CORE, including the scoring function and the optimization algorithms described above.

Protein Gβ1 domain

The B1 domain (IgG binding domain) of protein G (Gβ1) is composed of 56 amino acids and contains no disulfide bonds or bound cofactors. Native Gβ1 forms a well-packed structure in solution containing β-sheet, α-helix, and turn. Stability studies of both wild-type Gβ1 and engineered mutants have provided reliable thermodynamic parameters (Alexander et al., 1992; Dahiyat & Mayo, 1997b), making this protein well-suited for theoretical investigation. Eleven core residues were defined that exhibit little or no solvent exposure. Trp at position 43, although partially solvent exposed, was selected as a core residue because of multiple contacts with other buried hydrophobic core residues.

The simulated annealing procedure from one run of CORE produced a single sequence with a near minimized *Score* (i.e., zero bumps, near maximum ΔC_p, and near minimum ΔS_{conf}). During the subsequent low-temperature Monte Carlo sampling, 417 unique sequences were generated of which nearly 75% exhibited better (lower) *Score*s than the simulated annealing sequence. Among these 417 predicted sequences was the native (WT) sequence that exhibited an intermediate value for ΔC_p and a large value for ΔS_{conf}. In addition to regenerating the WT sequence, the native side-chain conformations of core residues was also reproduced. With respect to *Score*, the WT sequence is ranked in the top 60% of sequences predicted from the Monte Carlo sampling. A higher rank is not expected because there is little natural evolutionary pressure to produce proteins with maximum thermal stability and cooperativity; undoubtedly, evolutionary pressure to optimize function is more dominant. Data on the WT sequence and the 10 sequences with lowest *Score* are presented in Table 1. The top 10 sequences have values for ΔC_p higher than that of the native, suggesting that these designed variants would exhibit higher T_m values. Indeed, one of these sequences is the thermally stable engineered variant, α90 (Dahiyat & Mayo, 1997b). The predicted increase in T_m compared to the WT value (ΔT_m), presented in Table 1, is calculated using Equation 8 and $\Delta\Delta C_p$ values. The α90 variant has a predicted ΔT_m of at most 4.1 °C, while the measured ΔT_m value is 5 °C. The similarity in these values is an indication that enthalpy and entropy convergence occurs for α90 and WT, and that the backbone structures of these two proteins are very similar. This latter conclusion is supported by experimental evidence indicating similar structures for α90 and WT (Dahiyat & Mayo, 1997b).

Only 13,716 sequences out of the possible 2 billion (7^{11}: 11 core positions, 7 hydrophobic amino acids) were sampled during the sequence prediction run. For each of the sequences sampled, bumps

Table 1. *Sequence prediction of core hydrophobic residues of Gβ1 from a single CORE run*

Protein[a]	Rank (/417)	Bumps	ΔS_{conf}[b]	ΔC_p[c]	*Score*[d]	Core sequence position											$\Delta\Delta C_p$/res[e]	Exp. ΔT_m[f]	Calc. ΔT_m[g]
						3	5	7	20	26	30	34	39	43	52	54			
WT	247	0	−3.66	55.54	−59.20	Y	L	L	A	A	F	A	V	W	F	V	—	—	—
	1	0	−4.98	59.65	−64.63	F	.	I	.	.	.	.	I	F	.	I	0.81		5.1
	2	0	−4.57	59.52	−64.09	F	.	I	.	.	.	L	I	A	.	F	0.78		4.9
	3	0	−4.75	58.70	−63.45	F	.	I	.	.	.	.	I	V	.	F	0.62		3.9
	4	0	−4.64	58.70	−63.34	F	.	V	.	.	.	.	I	F	.	I	0.62		3.9
	5	0	−5.28	57.97	−63.24	L	.	I	.	.	.	L	I	A	.	F	0.48		3.0
α90	6	0	−4.30	58.88	−63.18	F	.	I	.	.	.	.	I	.	.	.	0.66	5	4.1
	7	0	−5.03	58.09	−63.13	L	.	I	.	.	.	.	I	F	.	I	0.50		3.1
	8	0	−4.95	58.09	−63.04	F	.	I	.	.	.	.	I	L	.	I	0.50		3.1
	9	0	−4.25	58.76	−63.01	F	.	.	.	.	.	.	I	.	.	.	0.63		4.0
	10	0	−4.26	58.64	−62.90	F	.	.	.	.	.	.	L	.	.	.	0.61		3.8

[a] α90 was computer designed, synthesized, and characterized by Mayo and co-workers (see text).
[b] Average per residue conformational entropy change of folding for core residues in units of $cal \cdot mol^{-1} \cdot K^{-1} \cdot res^{-1}$.
[c] Average per residue heat capacity change of unfolding for core residues in units of $cal \cdot mol^{-1} \cdot K^{-1} \cdot res^{-1}$.
[d] Calculated using Equation 12 with $\beta = 1$ and $\gamma = 1$.
[e] Per residue change in ΔC_p for the entire protein relative to WT.
[f] The T_m of WT is 87 °C.
[g] Calculated using the upper limit value in Equation 8.

were calculated for 2 million structures in which amino acid side chains populate different rotamers derived from a rotamer library file (see Methods), such that a total of nearly 3×10^{10} structures were sampled. The output of this single run of CORE is represented graphically in Figure 2 that plots the ΔC_p and ΔS_{conf} for sequences with zero bumps generated during both simulated annealing and Monte Carlo sampling. The plot reveals the expected correlation between ΔC_p and ΔS_{conf} mentioned above. In addition, the plot shows that sequences sampled during simulated annealing span a wide range of *Score*s ($\Delta S_{conf} - \Delta C_p$), while sequences accepted during the low-temperature Monte Carlo sampling span a much narrower range of ΔC_p and ΔS_{conf} values. It is also noteworthy that sequences with the lowest *Score*s are generated from the Monte Carlo sampling, not the simulated annealing procedure, justifying the use of low-temperature Monte Carlo sampling.

Not surprisingly, subsequent sequence prediction runs starting from different random sequences do not produce the same 417 sequences; however, it is striking that from three separate runs of CORE, the same 10 sequences presented in Table 1 are predicted (data not shown). This strongly suggests that the sequence with the best *Score* (see Table 1) is indeed at the global minimum.

434 cro

The cro protein from bacteriophage 434 (434 cro) is a small 64 amino acid protein that does not contain disulfide bonds or metal binding sites. The hydrophobic core of 434 cro has previously been redesigned by Handel and co-workers using their protein design program that scores sequences by employing both van der Waals energy and changes in buried volume calculations (Desjarlais & Handel, 1995). Based on their computational results, several variants were synthesized and carefully characterized. In this previous study, 12 hydrophobic core residues were targeted for redesign; however, two of these residues (Leu13 and Trp58) exhibit significant solvent exposure. Therefore, only 10 buried hydrophobic amino acids were identified as core residues in the current study. Exclusion of the residues at positions 13 and 58 did not preclude direct comparison of the available experimental data with results generated by CORE, because the native amino acids at these positions were retained for all predicted sequences in the previous design. As was done in the previous study (Desjarlais & Handel, 1995), a Cys residue at position 54 was mutated to Val, facilitating direct comparison between *Score*s generated by CORE and available thermodynamic data.

The low-temperature Monte Carlo sampling procedure from a single run of CORE generated 151 sequences predicted to stabilize the native structure of 434 cro. It was encouraging to find the "WT" (C54V) sequence among these predicted sequences. Once again, it should not be surprising that this sequence is not found among the top ranked sequences, because there is little or no natural evolutionary pressure to produce a highly thermally stable 434 cro. Table 2 presents data on the top 10 predicted sequences. Among these sequences is the only variant designed by Handel and co-workers with a higher T_m than the native. In addition, previously designed variants with lower T_m values were not predicted because, as it turns out, these sequences are associated with nonzero bumps (D-7 and D-8 in Table 2).

Myoglobin

One of the principal difficulties in protein design is dealing with the enormous number of combinations of all possible amino acid sequences and side-chain conformations. This becomes particularly challenging as the size of the target protein increases. However, the efficiency by which the *Score* is calculated in CORE may allow for sequence prediction of the hydrophobic cores of large proteins. Horse heart myoglobin, which consists of 153 amino acids including 34 residues that define an extensive hydrophobic core, was chosen as the first test. The number of possible structures (sequences and rotamers) is a staggering 10^{83} for this protein,

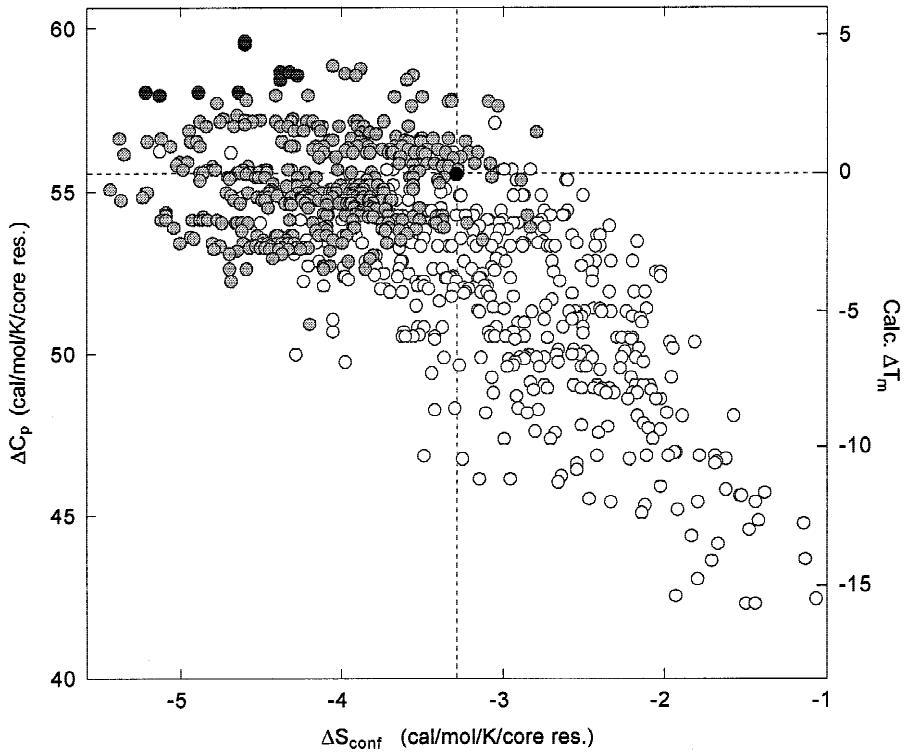

Fig. 2. Plot of ΔC_p vs. ΔS_{conf} for Gβ1 sequences with zero bumps sampled during the simulated annealing portion (white), and accepted during the Monte Carlo portion (light gray) of a single run of CORE. The top 10 sequences with respect to *Score* ($-\Delta C_p + \Delta S_{conf}$) are highlighted (dark gray) in the top left corner of the plot. The native sequence that was predicted during the Monte Carlo run is also highlighted (black). The dashed lines represent the WT ΔC_p and ΔS_{conf} values.

Table 2. *Sequence prediction of core hydrophobic residues of 434 cro from a single CORE run*

Protein[a]	Rank (/151)	Bumps	ΔS_{conf}[b]	ΔC_p[c]	*Score*[d]	Core sequence position										$\Delta\Delta C_p$/res[e]	Exp. ΔT_m[f]	Calc. ΔT_m[g]
						2	6	20	26	31	34	45	48	52	59			
"WT"	78	0	−5.42	61.15	−36.00	L	L	L	V	I	I	L	I	L	L	—	—	—
	1	0	−6.08	62.46	−37.31	I	I	.	I	.	I	.	.	.	.	0.21		1.3
D-5	2	0	−6.13	62.32	−37.29	I	I	.	I	.	L	.	.	.	.	0.19	4	1.2
	3	0	−6.12	62.32	−37.28	I	I	.	I	.	.	.	L	.	.	0.19		1.2
	4	0	−6.14	62.19	−37.24	I	I	.	I	.	L	.	L	.	.	0.17		1.1
	5	0	−6.12	62.05	−37.15	.	.	.	I	.	.	.	L	.	.	0.15		0.94
	6	0	−6.57	61.02	−37.08	I	I	.	L	V	L	.	L	.	.	−0.02		−0.13
	7	0	−6.47	61.15	−37.05	I	I	.	L	V	.	.	L	.	.	0.00		0
	8	0	−6.47	61.15	−37.05	I	I	V	L	.	L	.	L	.	.	0.00		0
	9	0	−5.91	62.19	−37.01	I	.	.	I	.	.	.	L	.	.	0.17		1.1
	10	0	−6.35	61.29	−37.00	I	I	V	L	.	.	.	L	.	.	0.03		0.19
D-7	—	3	—	—		I	F	V	L	V	.	.	L	.	.	—	−39	—
D-8	—	10	—	—		F	I	.	L	V	L	.	L	.	.	—	−6	—
M-5	—	5	—	—		.	.	.	L	L	L	.	L	.	.	—	−23	—

[a]D-5, D-7, and D-8 were computer designed, synthesized, and characterized by Handel and co-workers (see text). M-5 has all core residues mutated to leucine. The bottom three sequences were not predicted by CORE due to nonzero bumps and are shown only for comparison.

[b]Average per residue conformational entropy change of folding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.

[c]Average per residue heat capacity change of unfolding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.

[d]Calculated using Equation 12 with $\beta = 1$ and $\gamma = 0.5$.

[e]Per residue change in ΔC_p for the entire protein relative to WT.

[f]The T_m of C54V ("WT") is 56 °C.

[g]Calculated using the upper limit value in Equation 8.

resulting from the possible seven hydrophobic residues and an average of 40 rotamers per residue position.

In a single run of CORE, only 3,000 sequences were sampled during the simulated annealing procedure to yield one sequence near the global minimum with respect to *Score*; this step took 86 h on an SGI Onyx Workstation (MIPS R10000, 180 MHz processor). The subsequent low-temperature Monte Carlo sampling procedure, which was manually terminated after 7 days, produced 202 unique sequences with remarkable sequence homology to WT. One of the predicted sequences, ranked in the top 10%, has nearly 75% (25 of 34) core residue identity with WT. Data on this sequence and the top five predicted sequences are presented in Table 3. All six sequences in Table 3 are associated with larger ΔC_p and lower ΔS_{conf} values compared to WT, suggesting that these variants would exhibit higher T_m values and cooperativity. The predicted structure of the top ranked sequence is strikingly similar to that of the crystal structure of WT. The identity and side-chain conformation of all seven WT aromatic core residues (W14, F23, F43, F46, F123, F138, Y146) are duplicated in the predicted structure with the exception of a minor Y146F mutation. One hundred twenty-nine (87%) of the 149 nonhydrogen side-chain atoms from the 34 core amino acids in the WT structure are duplicated in the predicted structure, with an RMS deviation of only 0.4 Å.

A plot of ΔC_p vs. ΔS_{conf} for sequences with zero bumps sampled during the simulated annealing and Monte Carlo procedures is presented in Figure 3. The plot highlights the ΔC_p and ΔS_{conf} values for WT, which are on the edge of the values generated during Monte Carlo sampling. This may explain why the WT sequence is not predicted, despite high homology with the predicted sequences.

A random subset of the predicted myoglobin sequences was analyzed to quantitatively demonstrate the relationship discussed above between the extent of networking of side-chain contacts (i.e., cooperativity) and conformational entropy. Figure 4 presents the results in which a "contact index" is plotted as a function of conformational entropy. The contact index is calculated using Equation 13 and corresponds to the extent of networking within the protein interior.

$$\text{Contact Index} = \frac{\sum_{i}^{i=N_g} [m_i(m_i - 1)]}{2N_{core}(2N_{core} - 1)} \tag{13}$$

where N_{core} is the total number of core residues, m_i is the number of amino acids in an island of side chains within van der Waals contact, and N_g is the number of islands. The factor of 2 in the denominator assumes that core residues make on average one contact with a noncore residue such that the contact index ranges from 0–1. The contact index favors long-range side-chain interactions that lead to extensive networks; for example, two proteins each with $N_{core} = 10$ and two clusters of contacting side chains ($N_g = 2$) exhibit very different contact indices if one of the proteins has 2 and 18 amino acid clusters and the other has 10 and 10 amino acid clusters. In the former the contact index is 0.8, while in the latter it is 0.5.

Methionine aminopeptidase

Final validation of the scoring function and optimization algorithm employed by CORE was accomplished through prediction of the core sequence of a protein nearly twice the size of myoglobin; methionine aminopeptidase from the hyperthermophilic organism *Pyrococcus furiosus*. This 259 amino acid protein contains an extensive hydrophobic core consisting of 63 amino acids.

In the three previous runs, simulated annealing was initiated at a high Metropolis temperature starting with a random sequence. A significant portion of the simulated annealing run is therefore devoted to predicting sequences with zero bumps. It is only in the latter portion of the simulating annealing procedure that the ΔC_p is maximized and the ΔS_{conf} is minimized. For this much larger protein, simulating annealing was initiated at a lower Metropolis temperature starting from a sequence in which all 63 positions were mutated to Ala. This step, while still avoiding input of a sequence bias, significantly shortens the run time because sequences that exhibit zero bumps are predicted from the beginning and the whole simulated annealing procedure is devoted to maximizing ΔC_p and minimizing ΔS_{conf}.

A single run of CORE produced 330 unique sequences; data on the top five sequences as well as the sequence with the highest homology to WT are presented in Table 4. The top five sequences have better (lower) *Score*s than WT, indicating that these protein variants would exhibit enhanced thermal stability and cooperativity compared to WT. The sequence with highest homology is ranked in the top 10% and has 45 out of 63 positions that match the sequence of WT. The striking similarity to WT is made even more remarkable when considering that the probability of obtaining 45 matches through random selection is 10^{-24}. The native sequence is not generated, despite having a *Score* within the range of predicted sequences. This may in part be due to the fact that there is an enormous number of possible sequences (10^{53}) for the 63 amino acid core.

The efficiency by which simulated annealing reaches a sequence near the global minimum is shown in Figure 5 that plots the ΔC_p, ΔS_{conf}, and the *Score* for all sequences sampled during the run. The plot reveals that at early times during simulated annealing when the Metropolis temperature is relatively high the algorithm allows for escape from local minima. It is clear from the plot that the subsequent low-temperature Monte Carlo sampling procedure produces sequences with *Score*s lower than that generated during simulated annealing.

Discussion

We have demonstrated that for a series of proteins with fixed backbone structure and constant polar contribution to the thermodynamics of folding, increases in ΔC_p (associated with increases in buried hydrophobic surface area) correlate to increases in T_m and increases in ΔG_u above a fixed temperature between 20 and 50 °C defined by the convergence parameters. Metropolis-driven simulated annealing and low-temperature Monte Carlo sampling are effectively utilized to sample the enormous number of possible sequences and side-chain conformations to predict hydrophobic core sequences and structures of native proteins. Results are presented for hydrophobic core sequence prediction of four proteins ranging in size from 56 to 259 amino acids. These results clearly validate the use of hard sphere bumps to accurately represent steric compatibility with a target structure. Also validated is the use of heat capacity change and conformational entropy as design criteria to predict sequences of thermally stable proteins with high cooperativity.

Table 3. *Sequence prediction of core hydrophobic residues of myoglobin from a single CORE run*

Protein	Rank (/202)	Bumps	ΔS_{conf}[a]	ΔC_p[b]	*Score*[c]	Core sequence position																																								
						2	10	13	14	17	28	29	30	32	33	43	46	49	61	68	69	72	75	76	86	89	99	104	107	110	111	114	115	123	134	135	138	142	146							
WT	—	0	−5.15	61.29	−66.44	L	V	V	W	V	V	L	I	L	F	F	F	L	L	V	L	L	I	L	L	L	I	L	I	A	I	V	L	F	A	L	F	I	Y							
	1	0	−5.63	63.30	−68.93	.	L	.	.	L	I	.	.	.	.	.	.	.	.	I	.	.	.	F	I	F	V	I	V	.	V	.	.	.	.	.	.	L	F							
	2	0	−5.89	62.86	−68.75	.	L	.	.	L	L	.	.	.	.	.	.	.	.	I	.	.	.	F	I	F	A	I	V	.	L	.	.	.	.	.	.	L	F							
	3	0	−5.93	62.79	−68.72	.	L	.	.	L	I	.	.	.	.	.	.	.	.	I	.	.	.	F	I	F	V	I	V	.	V	.	.	.	.	.	L	L	F							
	4	0	−6.29	62.38	−68.67	.	L	.	I	L	L	.	.	.	.	.	.	.	V	I	I	.	.	F	V	Y	.	I	.	.	L	.	.	.	.	.	.	.	F							
	5	0	−6.00	62.55	−68.55	.	L	.	.	L	L	.	.	.	.	.	.	.	.	I	.	.	.	F	I	.	V	I	V	.	.	.	.	.	.	.	L	L	F							
	23	0	−5.62	62.47	−68.09	.	.	.	.	L	L	.	.	.	.	.	.	.	.	I	I	.	.	.	.	F	.	I	.	.	L	.	A	.	.	.	.	.	F							

[a]Average per residue conformational entropy change of folding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.
[b]Average per residue heat capacity change of unfolding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.
[c]Calculated using Equation 12 with $\beta = 1$ and $\gamma = 1$.

Table 4. *Sequence prediction of core hydrophobic residues of methionine aminopeptidase from a single CORE run*

Protein	Rank (/330)	Bumps	ΔS_{conf}[a]	ΔC_p[b]	Score[c]	Core sequence[d]
WT	—	0	−3.26	50.82	−54.08	L A A V A A L L A I I A F V L I A A L L I V V I A A V V V A A A L A A A I L A I I I L I L A I L V F A I F A V L V A F I I V V
	1	0	−3.77	51.20	−54.97	. . . . . A . . . V . . L I L I . F V V . . F . . I L I . . . . . F . L W I I . . I . . . A . . V . . .
	2	0	−3.98	50.97	−54.95	. . . . . A . . . L . A L F . I . F V V . . F . . I L I . . . . . L L . I . F . L W I I . . I . . . A . . . V L .
	3	0	−3.78	51.15	−54.93	. . . . . A F . . L . . . V . I . Y V . I . F . . I L I . . . . . V F . V V . I . F . . . . I . . V L . . . A . . . . L .
	4	0	−3.75	51.16	−54.92	. . . . . A . . . . . I . F V L L . . . . I L I . . . . . I . . L V . L . . L . I . . F I . . . A . . V . I .
	5	0	−3.79	51.13	−54.91	. . . . . A F . . L . . . V . I . Y V . I . F . . I L I . . . . . V F . V V . . . F . . . I . . V L . . . A . . . . L .
	36	0	−3.68	50.83	−54.51	. . . . . A . . . . . I . F V L L . . . . I L I . . . . . I . . L . . . . . L . I . . F A . . . A . . V . I .

[a]Average per residue conformational entropy change of folding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.
[c]Average per residue heat capacity change of unfolding for core residues in units of cal·mol^{-1}·K^{-1}·res^{-1}.
[c]Calculated using Equation 12 with $\beta = 1$ and $\gamma = 1$.
[d]Core residue positions: 6, 9, 13, 16, 20, 24, 30, 33, 34, 37, 41, 49, 50, 52, 54, 56, 60, 61, 73, 79, 81, 83, 85, 91, 92, 95, 96, 98, 100, 110, 111, 114, 115, 117, 118, 122, 128, 131, 134, 135, 139, 147, 150, 155, 160, 162, 166, 178, 183, 184, 185, 186, 189, 190, 197, 268, 276, 277, 279, 283, 284, 285, 290.

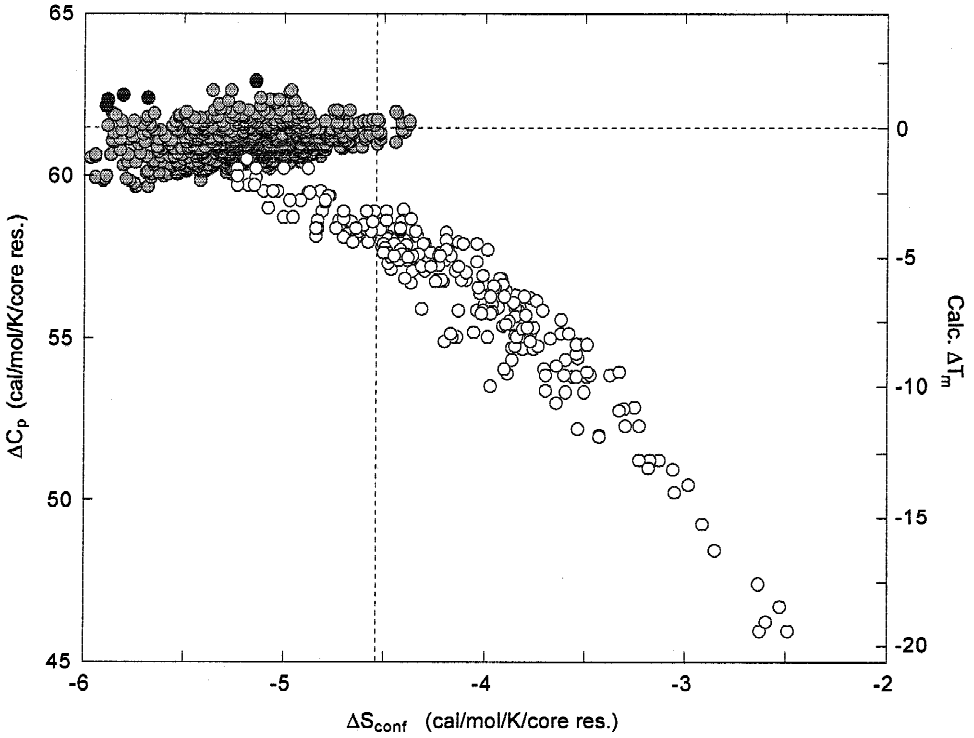

Fig. 3. Plot of ΔC_p vs. ΔS_{conf} for myoglobin sequences with zero bumps sampled during the simulated annealing portion (white), and accepted during the Monte Carlo portion (light gray) of a single run of CORE. The top five sequences with respect to *Score* ($\Delta S_{conf} - \Delta C_p$) are highlighted (dark gray) in the top left corner of the plot. The ΔC_p and ΔS_{conf} values of the native sequence that was not predicted are highlighted with dashed lines. The plot clearly shows the expected correlation between ΔC_p and ΔS_{conf} mentioned in the text.

Conformational entropy and free energy of unfolding

Side-chain conformational entropy of folding (ΔS_{conf}) is minimized using CORE to predict sequences with maximum cooperativity. At first glance, this might appear to have the inadvertent effect of decreasing ΔG_u, because conformational entropy is a major portion of the entropic component of the free energy of protein folding, contributing as much as -0.5 kcal·mol^{-1} per rotamer at room temperature (Doig & Sternberg, 1995). However, decreases in ΔS of folding are compensated by a similar decrease in ΔH of folding near room temperature. This is often referred to as enthalpy-entropy compensation of protein thermodynamics (Lumry & Rajender, 1970; Dunitz, 1995). The result of the nearly 1:1 compensation between enthalpy and entropy leads to small and unpredictable modulation of ΔG_u upon changes in ΔS_{conf}.

Enthalpy–entropy compensation is a general feature of many chemical reactions and processes in biological systems (Lumry & Rajender, 1970; Dunitz, 1995). Enthalpy–entropy compensation of folding has also been reported for natural proteins and their mutants (Hawkes et al., 1984; Shortle et al., 1988). The slope of the linear plot of ΔH° vs. ΔS° is the compensation temperature T_c. The value of T_c is close to room temperature (Lumry & Rajender, 1970); therefore, the wide range of ΔH_u and ΔS_u values for native proteins are adjusted so that ΔG_u remains nearly constant at room temperature. A consequence of enthalpy–entropy compensation with respect to protein design is that enthalpy and entropy terms should not be optimized separately if protein thermodynamics is a design goal; it is the combination of enthalpy and entropy that yields stability at all temperatures. ΔS_{conf} is minimized in CORE only to produce sequences of proteins with high cooperativity and is not intended to be used as a criteria to optimize thermal or chemical stability. As indicated by Equation 5, it is only through maximizing ΔC_p that individual contributions from ΔH and ΔS are jointly accounted for to produce sequences with optimal thermal stability.

Use of van der Waals energy as a design criteria

A simple and computationally efficient hard-sphere bump calculation is employed in CORE to define steric compatibility with an input target structure. However, most other protein design programs utilize E_{vdw} instead to define steric compatibility. Furthermore, low E_{vdw} is often assumed to be associated with well-ordered, thermally stable proteins. Therefore, programs that utilize E_{vdw} generate sequences predicted to stabilize an input target structure by searching for sequences associated with minimal E_{vdw} of the folded state of a protein. Although implementing E_{vdw} in this way may seem logical, it is, in fact, very unlikely that any correlation between E_{vdw} and thermal stability exists for the following reasons: (1) the unfolded state or states of a protein cannot be ignored when enhanced thermodynamic stability is the design goal; (2) the existence of complex correlations between E_{vdw} and other parameters that contribute to ΔG, such as hydrophobic surface area and side-chain conformational entropy, makes the magnitude and sign of the contribution of E_{vdw} to ΔG_u and T_m impossible to

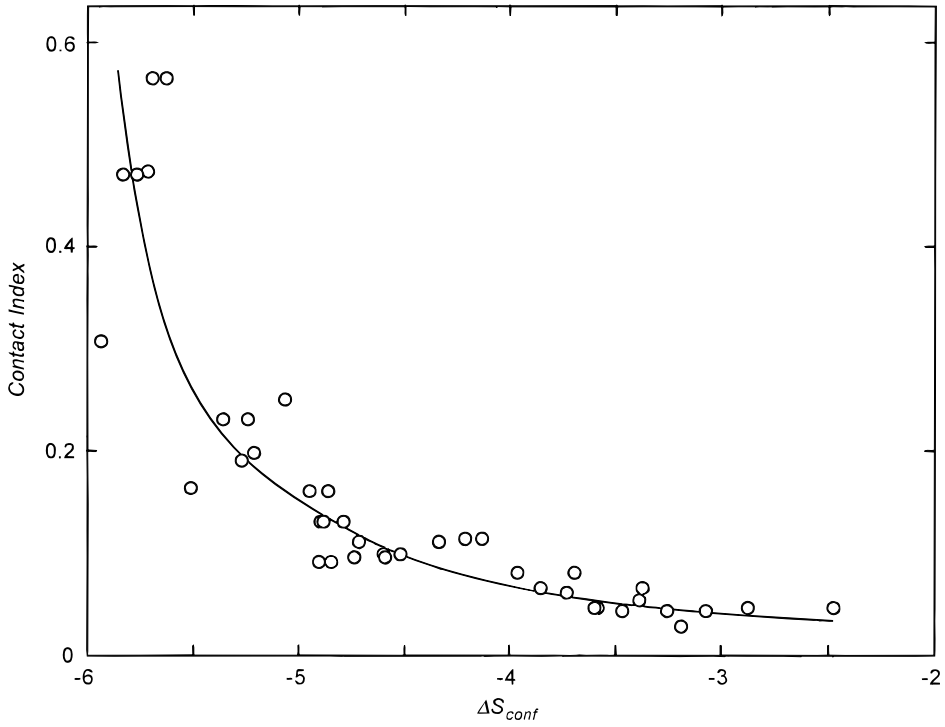

Fig. 4. Plot of contact index vs. ΔS_{conf} for a random sampling of 16 predicted myoglobin sequences. The contact index represents the extent to which side chains form long-range contacts giving rise to extended side-chain networks, a key feature of uniquely folded natural proteins with cooperative unfolding transitions. The plot demonstrates that minimizing ΔS_{conf} is an effective means by which protein cooperativity can be maximized.

determine; and (3) the value of E_{vdw} is subject to potentially large errors, because it is necessary in protein design programs to define side-chain conformations using a discrete rotamer library. This last reason presents the most significant problem with respect to using E_{vdw} as a design criteria, because slight fluctuations in side-chain torsion angles can lead to dramatic changes in E_{vdw}. For example, changes in the side-chain torsion angles χ_1 and χ_2 within a small 5° window result in side-chain hydrogen atoms shifting position by as much as 0.5 Å for Phe and 0.3 Å for Leu. Considering only one nonbonding H···H interaction, a change in H···H distance of 0.5 Å, from a reasonable distance of 2.5 to 2.0 Å, increases E_{vdw} (12,6 Leonard–Jones potential) by about 4 kcal/mol. A change of 0.5 Å, from the reasonable distance of 2.2 to 1.7 Å, is associated with an increase in E_{vdw} of ~40 kcal/mol. Clearly, E_{vdw} of a structure generated using a discrete rotamer library may not accurately represent steric compatibility. To demonstrate this, the van der Waals energies of 16 random myoglobin structures with zero bumps sampled during a CORE run (see Results) were calculated using the Tripos Force Field (in Sybyl 6.3). These protein structures were then energy minimized using 10 iterations, holding the backbone atoms and heme group fixed. A plot of the resulting E_{vdw}^{min} vs. the initial E_{vdw} (Fig. 6A) shows only weak correlation between the two energies (slope = 0.47) with significant noise (correlation coefficient $R = 0.36$). This plot demonstrates the potential errors associated with using E_{vdw} of structures generated using discrete rotamer libraries and idealized side chains. Subsequent analysis of the minimized structures reveals that indeed the 10-iteration minimization led to χ_2 angle changes of as large as 6° for some of the amino acids. One way to minimize these errors associated with using E_{vdw} as a criterion to represent steric compatibility is to define side-chain torsion angles using a small enough increment; however, the increment must clearly be less than 5°, a condition that is not computationally feasible. For side chains with two torsion angles (Leu, Ile, Phe, Tyr, Trp), the number of rotamers would approach 500 if a 5° incremented rotamer library were used. Significantly longer time would then be required to effectively sample the enormous number of possible conformations. The error associated with using E_{vdw} can be partially overcome by decreasing the van der Waals radius scaling factor (Dahiyat & Mayo, 1997b); however, this may lead to selection of sequences with side chains that in reality are sterically clashing.

The use of hard-sphere bumps to select sequences sterically compatible with target structures does not, of course, avoid any of the potential errors associated with using E_{vdw}. In fact, the way in which bumps are employed as a design criteria in CORE introduces an additional potential error; a contact that is only 0.001 Å shorter than the predefined cutoff distance would be associated with a nonzero bump and would, therefore, be rejected. The "softness" of E_{vdw} avoids this problem to some degree. Although there will be significant overlap between sequences predicted using E_{vdw} and bumps, there will undoubtedly be sequences that one criteria rejects that the other accepts and vice versa. To demonstrate this point, E_{vdw} and E_{vdw}^{min} of myoglobin sequences presented in Figure 6A are plotted as a function of ΔC_p (Fig. 6B,C). The shaded region in Figure 6B indicates the range of ΔC_p values for CORE-predicted sequences and the dark circles represent actual sequences predicted by CORE. About half of these sequences are associated with van der Waals energies high enough (100 kcal/mol

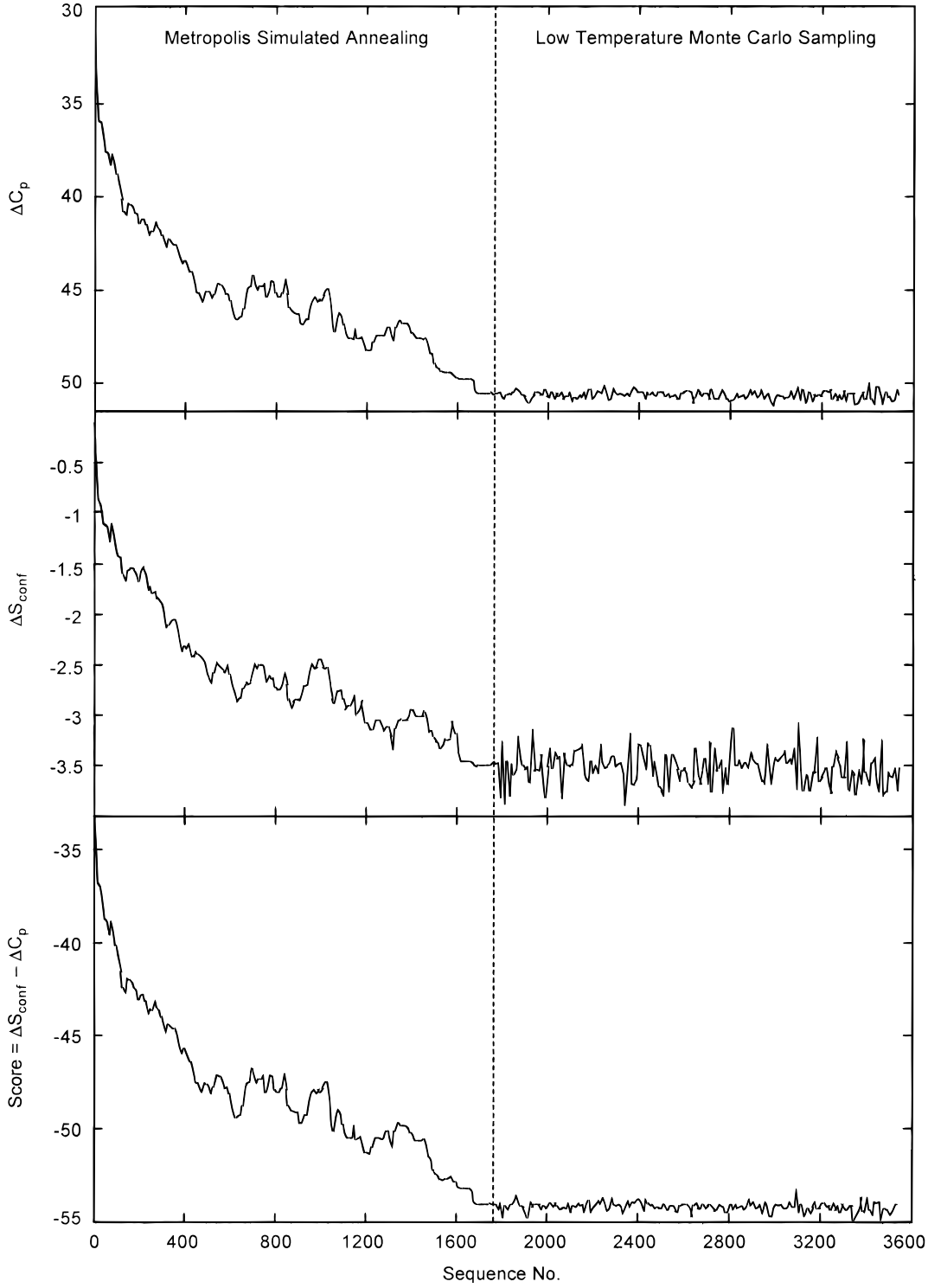

Fig. 5. Plots showing the optimization of ΔC_p, ΔS_{conf}, and *Score* as a function of sequence number for methionine aminopeptidase sequence prediction using CORE. Because the starting point is the sequence in which all positions are mutated to Ala, the initial ΔC_p value is 30.15 cal·mol^{-1}·res^{-1} and the initial ΔS_{conf} is 0 cal·mol^{-1}·res^{-1}. Approximately 1,700 sequences were sampled during the simulated annealing portion of the run. Many of the predicted sequences from the subsequent low-temperature Monte Carlo sampling have lower *Scores* than that of the sequence predicted from simulated annealing.

higher than the minimum) that it appears they would have been rejected if E_{vdw} were used as a design criteria. To demonstrate that these sequences are not, in fact, incompatible with the native structure of myoglobin, E_{vdw}^{min} was plotted as function of ΔC_p. This plot (Fig. 6C) shows that all the predicted sequences generated by CORE are, in fact, sterically compatible with the native structure

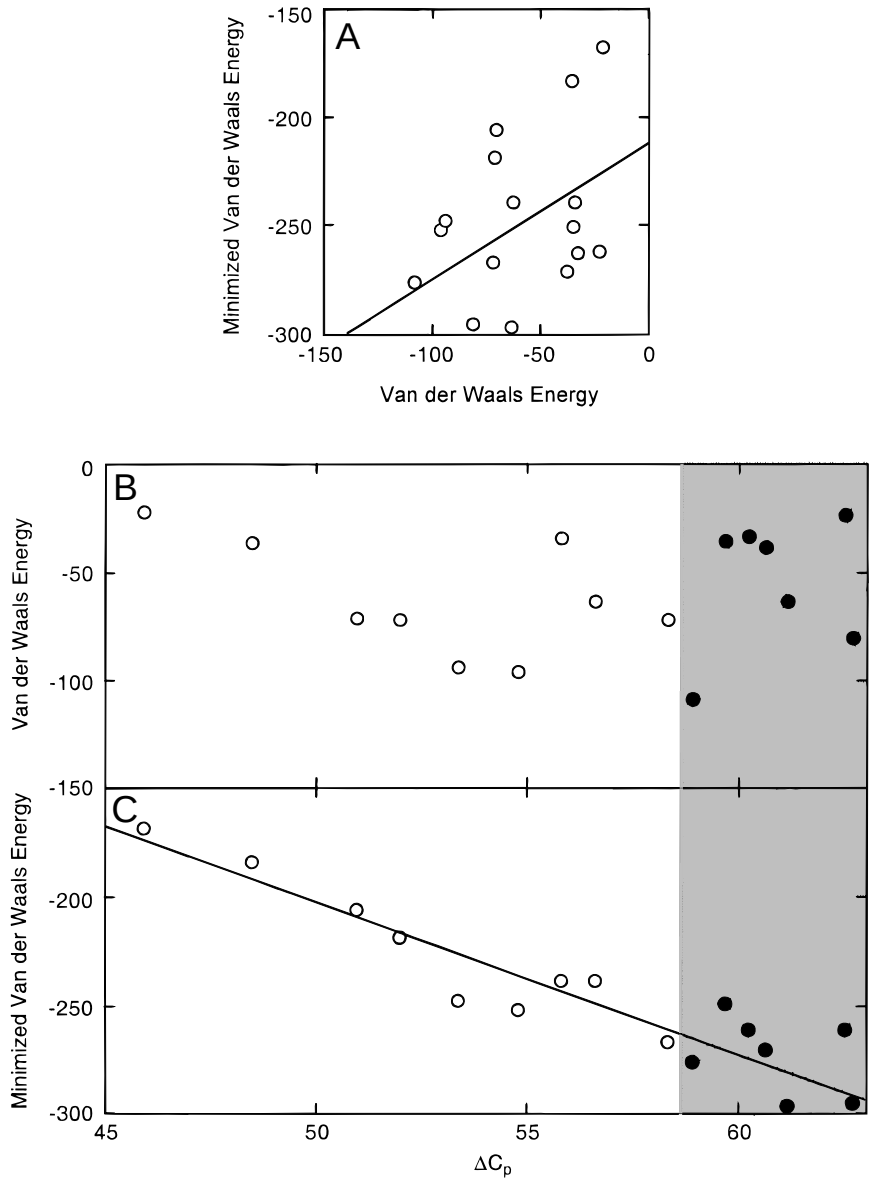

Fig. 6. **A:** Plot of van der Waals energy after a 10-iteration energy minimization vs. van der Waals energy before minimization for 16 random myoglobin structures sampled during a CORE run. **B:** Van der Waals energy plotted as a function of ΔC_p for these same myoglobin structures, showing no obvious correlation. The shaded region represents the range of ΔC_p values for predicted myoglobin sequences, black circles represent predicted myoglobin sequences, and open circles represent sequences with zero bumps sampled, but not predicted. **C:** Van der Waals energy after the 10-iteration energy minimization vs. ΔC_p, showing good correlation between these parameters.

of myoglobin. The plot also shows a striking correlation between ΔC_p and E_{vdw}^{min}, suggesting that sequences with low E_{vdw}^{min} will indeed be associated with stable protein structures possessing high T_m and ΔG_u. The correlation is not surprising, because large ΔC_p of sequences with zero bumps is associated with large buried hydrophobic surface area, which is in turn, most likely associated with low van der Waals energy because of increased favorable hydrophobic interactions. Of course, this correlation does not provide any utility with respect to employing E_{vdw}^{min} as a design criteria, because calculating E_{vdw}^{min} from a 10-iteration energy minimization requires approximately five orders of magnitude more time than a hard-sphere bump calculation. The weak correlation between ΔC_p and E_{vdw}, especially at high energy, provides a possible explanation for some of the observed similarities in sequence prediction between CORE and other design programs (see Results), despite the use of very different criteria to define protein stability. Although the use of E_{vdw} and bumps in defining steric compatibility are both associated with potential errors as a consequence of the necessary implementation of discrete rotamer libraries, a hard-sphere bump calculation is associated with shorter computation times and therefore appears to offer an advantage over the use of E_{vdw} to define steric compatibility.

The appropriate way to implement E_{vdw} in protein design programs is to define a cutoff energy for which sequences with E_{vdw} above this energy are rejected. A second criteria that can be shown to correlate to measurable thermodynamic parameters such as heat

capacity change should then be used to rank the sequences with E_{vdw} below the cutoff energy. The value of the cutoff energy is difficult to determine (as is the cutoff distance in a bump calculation), because it is important that *all* sterically compatible sequences be accepted while *all* sequences with clashing side chains be rejected. Adjusting scale factors that modulate the van der Waals radii in E_{vdw} calculations (Dahiyat & Mayo, 1996) is often necessary to achieve the desired balance between accepting sterically compatible sequences while rejecting sequences with clashing side chains.

The results presented here demonstrate that hard-sphere bumps accurately represent steric compatibility with a target protein structure. The use of van der Waals energy to accomplish the same thing offers no significant advantage and is computationally less efficient. The results also show that side-chain conformational entropy of folding and heat capacity change of unfolding yield a scoring function that can be effectively minimized using Metropolis-driven simulated annealing and low-temperature Monte Carlo sampling, thus allowing sequence prediction of proteins exhibiting optimal cooperativity and high thermal stability. Moreover, the "force field" employed in CORE is simple and highly efficient to calculate, thereby greatly facilitating the design of large synthetic proteins and reengineered natural proteins that exhibit optimized thermal and chemical stability.

Methods

Rotamer library

Accurate side-chain entropy calculations and bump calculations rely on the use of an extensive rotamer library with a large range of allowable torsion angles incremented by relatively small steps. To generate an accurate rotamer library for use in CORE, analysis of side-chain conformations of the hydrophobic amino acids of 44 nonhomologous high resolution protein crystal structures using Iditis (Oxford Molecular, Oxford, United Kingdom) was conducted (the list of proteins was obtained from Williams et al., 1994). This revealed that rotamers were scattered about a mean value within a range of about 40° for each χ angle. Therefore, to generate the rotamer library used to define side-chain conformations in CORE, each average χ value of a rotamer was expanded by $\pm 20°$ such that each Val rotamer was expanded by an additional two rotamers, each Ile, Leu, Phe, Tyr, and Trp rotamer was expanded by an additional eight rotamers, and each Met rotamer was expanded by an additional 26 rotamers. The resulting rotamer library contains a total of 657 rotamers (Ile = 54, Leu = 63, Met = 378, Phe = 45, Trp = 63, Tyr = 45, Val = 9).

Protein input

WT structures were obtained from the Brookhaven Protein Data Bank (Gβ1, 1PGA; 434 cro, 2CRO; myoglobin, 1WLA; methionine aminopeptidase, 1XGO). To more accurately describe side-chain contacts, explicit hydrogen atoms were added using the Biopolymer module in Sybyl 6.3 (Tripos, St. Louis, Missouri). Potential strain introduced by addition of hydrogen atoms was relaxed by a 100-iteration energy minimization using the Kollman all-atom force field and Kollman charges, distance dependent dielectric constant of 6, nonbonding cutoff of 8.0 Å, and aggregated backbone atoms. To assure that core residues remain buried during side-chain mutations and rotations, minimized proteins were solvated with one layer of H_2O using the droplet method in Sybyl and a van der Waals bump factor of 1.0. Core residues were initially identified using the program DSSP (Kabsch & Sander, 1983) and then confirmed and modified if needed by visual inspection.

Acknowledgments

This work is partially supported by the Johnson & Johnson Discovery Research Fund and a Rutgers-Busch Biomedical Award. Computational resources were provided by the Rutgers-Newark Center for Computational Neuroscience. X.J. is grateful to the Department of Chemistry for fellowship support.

References

Alexander P, Fahnestock S, Lee T, Orban J, Bryan P. 1992. Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: Why small proteins tend to have high denaturation temperatures. *Biochemistry 31*:3597–3603.

Clark M, Cramer RD, Van Opdenbosch N. 1989. Validation of the general-purpose Tripos 5.2 Force-field. *J Comp Chem 10*:982–1012.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.

Dahiyat BI, Mayo SL. 1997a. De novo protein design: Fully automated sequence selection. *Science 278*:82–87.

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*:10172–10177.

DeGrado WF, Wasserman ZR, Lear JD. 1989. Protein design, a minimalist approach. *Science 243*:622–628.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

Desmet J, Maeyer MD, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*:539–542.

Doig AJ, Sternberg MJE. 1995. Side-chain conformational entropy in protein folding. *Protein Sci 4*:2247–2251.

Dunitz JD. 1995. Win some, lose some: Enthalpy-entropy compensation in weak intermolecular interactions. *Curr Biol 2*:709–712.

Freire E, Murphy KP. 1991. Molecular basis of co-operativity in protein folding. *J Mol Biol 222*:687–698.

Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. 1998. High-resolution protein design with backbone freedom. *Science 282*:1462–1467.

Hawkes R, Grutter MG, Schellman J. 1984. Thermodynamic stability and point mutations of bacteriophage T4 lysozyme. *J Mol Biol 175*:195–212.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.

Holm L, Sander C. 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from a C(alpha) trace application to model building and detection of co-ordinate errors. *J Mol Biol 218*:183–194.

Jiang X, Bishop EJ, Farid RS. 1997. A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins. *J Am Chem Soc 119*: 838–839.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*: 2577–2637.

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*:1680–1685.

Kono H, Nishiyama M, Tanokura M, Doi J. 1998. Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on side-chain packing. *Protein Eng 11*:47–52.

Lee C, Levitt M. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature 352*:448–451.

Lumry R, Rajender S. 1970. Enthalpy-entropy compensation phenomena in water solutions of proteins and small molecules: A ubiquitous property of water. *Biopolymers 9*:1125–1134.

Metropolis N, Rosenbluth M, Rosenbluth A, Teller E, Teller J. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*:1087–1092.

Murphy KP, Bhakuni V, Xie D, Freire E. 1992. Molecular basis of co-operativity in protein folding III. Structural identification of cooperative folding units and folding intermediates. *J Mol Biol 227*:293–306.

Murphy KP, Freire E. 1992. Thermodynamics of structural stability and cooperative folding behavior in protein. In: Anfinsen CB, Edsall JT, Richards FM, Eisenberg DS, eds. *Advances in protein chemistry*. San Diego: Academic Press. pp 313–361.

Murphy KP, Gill SJ. 1990. Group additivity thermodynamics for dissolutions of solid cyclic dipeptides into water. *Thermochim Acta 172*:11–20.
Murphy KP, Gill SJ. 1991. Solid model compounds and the thermodynamics of protein unfolding. *J Mol Biol 222*:699–709.
Murphy KP, Privalov PL, Gill SJ. 1990. Common features of protein unfolding and dissolution of hydrophobic compounds. *Science 247*:559–561.
Ponder JW, Richards FM. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.
Privalov PL, Gill SJ. 1988. Stability of protein structure and hydrophobic interaction. *Adv Protein Chem 39*:191–234.
Privalov PL, Khechinashvili NN. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure. *J Mol Biol 86*:665–684.
Privalov PL, Makhatadze GI. 1990. Heat capacity of proteins II. Partial molar heat capacity of the unfolded polypeptide chain of proteins: Protein unfolding effects. *J Mol Biol 213*:385–391.
Shenkin PS, Farid H, Fetrow JS. 1996. Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins Struct Funct Genet 26*:323–352.
Shortle D, Meeker AK, Freire E. 1988. Stability mutants of staphylococcal nuclease: Large compensating enthalpy-entropy changes for the reversible denaturation reaction. *Biochemistry 27*:4761–4768.
Street AG, Mayo SL. 1999. Computational protein design. *Structure 7*:R105–R109.
Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6*:1701–1707.
Tuchscherer G, Scheibler L, Dumy P, Mutter M. 1998. Protein design: On the threshold of functional properties. *Biopolymers 47*:63–73.
Williams MA, Goodfellow JM, Thornton JM. 1994. Buried waters and internal cavities in monomeric proteins. *Protein Sci 3*:1224–1235.