Modeling the effects of mutations on the denatured states of proteins

D. SHORTLE, H. S. CHAN and K. A. DILL

Protein Sci. 1992 1: 201-215

Protein Science (1992), *1*, 201-215. Cambridge University Press. Printed in the USA.
Copyright © 1992 The Protein Society 0961-8368/92 $5.00 + .00

Modeling the effects of mutations on the denatured states of proteins

DAVID SHORTLE,[1] HUE SUN CHAN, AND KEN A. DILL
Department of Pharmaceutical Chemistry, Box 1204, University of California at San Francisco,
San Francisco, California 94143

(RECEIVED August 1, 1991; ACCEPTED September 11, 1991)

Abstract

We develop a model for the reversible denaturation of proteins and for the effects of single-site mutations on the denatured states. The model is based on short chains of sequences of H (hydrophobic) and P (other) monomers configured as self-avoiding walks on the two-dimensional square lattice. The N (native) state is defined as the unique conformation of lowest contact energy, whereas the D (denatured) state is defined as the collection of all other conformations. With this model we are able to determine the exact partition function, and thus the exact native-denatured equilibrium for various solvent conditions, using the computer to exhaustively enumerate every possible configuration. Previous studies confirm that this model shows many aspects of protein-like behavior. The present study attempts to model how the denatured state (1) depends on the amino acid sequence, and (2) is changed by single-site mutations. The model accounts for two puzzling experimental results: (1) the replacement of a polar residue by a hydrophobic amino acid on the surface of a protein can destabilize a native protein, and (2) the "denaturant slope," $m = \partial \Delta G / \partial c$ (where c is the concentration of denaturant – urea, guanidine hydrochloride), can sometimes change by as much as 30% due to a single mutation. The principal conclusion of the present study is that, under strong folding conditions, the denatured conformations that are in equilibrium with the native state are not open random configurations. Instead, they are an ensemble of highly compact conformations with a distribution that depends on the residue sequence and that can be substantially altered by single mutations. Most importantly, we conclude that mutations can exert their dominant effects on protein stability by changing the entropy of folding.

Keywords: denatured state; protein denaturation; solvent denaturation; statistical mechanical model

The stability of a protein is defined as the difference in free energy between its native and denatured states. Mutational changes in a protein affect its stability only through this free energy difference; there is no direct experiment to determine whether a mutation has more effect on one or the other of the native or denatured states. However, one way to learn about the relative importance of native and denatured states in mutational processes is through the use of theoretical models. The purpose of this paper is to describe an elementary model of protein stability that can be used to assess how mutations can affect the native and denatured state contributions to stability and the denaturation processes. The model shows that mutations can have large effects on the denatured states, and it provides a conceptual basis for understanding two puzzling experimental observations: (1) that replacing a polar residue by a hydrophobic residue on the surface of a protein can *destabilize* the molecule, and (2) that a single site mutation can affect the "denaturation slope" (i.e., the m value; see below) by as much as 30%; previous theory has not been able to account for this (Alonso & Dill, 1991; Dill & Shortle, 1991).

The lattice model

The philosophy underlying our choice of a model for protein denaturation is as follows: it should be based on the dominant physical driving forces – the hydrophobic interactions, conformational freedom of the chain, and the steric restrictions imposed by the excluded volume of the chain. It should provide results from which we can draw rigorous conclusions, without adjustable parameters or arbitrary assumptions, about native and denatured states

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California at San Francisco, San Francisco, California 94143.

[1] Present address: Department of Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205.

for different sequences of monomers. These requirements preclude the use of atomic resolution models or the use of mean-field or long-chain lattice models. These requirements are satisfied, however, by an elementary model of short chains configured on two-dimensional square lattices that are sequences of H (hydrophobic) and P (other) monomers (Lau & Dill, 1989, 1990; Chan & Dill, 1991b). In this model, a contact between two H monomers is favorable by a free energy, ϵ ($\epsilon < 0$), envisioned to arise mainly from hydrophobic interactions (Dill, 1990; Chan & Dill, 1991a). The contact interaction free energy is a potential of mean force (Chan & Dill, 1991a) and implicitly accounts for the desolvation of the H monomers that must occur prior to formation of HH contacts. Because the chains are sufficiently short ($n = 16$ monomers in the present study), exhaustive computer enumeration of all the possible self-avoiding conformations permits unequivocal identification of the native conformations, i.e., those at the global minimum of free energy and therefore also of all other conformations that by definition constitute the denatured state. In this model, native conformations are those that have the greatest number of HH contacts. Because the chains are sufficiently short, we can then identify the native structure(s) for every possible sequence of H and P monomers, i.e., for the full sequence space.

Despite the obvious simplifications—that the chains are short, two-dimensional, and low-resolution—the model displays the following general properties of real proteins. (1) For a substantial fraction of sequences, small changes in the solvent character or temperature near a point of marginal stability lead to a relatively sharp transition from a large ensemble of denatured conformations to a relatively small ensemble of highly compact native conformations with a hydrophobic core; for a fraction of these folding sequences the ensemble of native structures is very small—on the order of one or a few configurations (Lau & Dill, 1989, 1990; Chan & Dill, 1991b). (2) The two-dimensional compact configurations have the same distribution of helix and parallel and antiparallel sheet topologies as in the known proteins (Chan & Dill, 1989b, 1990b). (3) The mutational properties of the model are consistent with those of real proteins: (a) most mutations among sequences with unique native structures are neutral in that they do not change the native states (see below), (b) the surface residues are more mutable than core sites, (c) there are second-site revertants, and (d) there is much sequence convergence; i.e., a given native structure will be encoded often in many different monomer sequences (Lau & Dill, 1990; Chan & Dill, 1991b). (4) The kinetics of folding to the native state resembles that of proteins in the following respects: (a) there are favored folding pathways that are sequence-dependent (Miller et al., 1992), and (b) the kinetic intermediates are highly compact states (Chan & Dill, unpubl.). We believe the shortness and the two-dimensionality of the chains are not as severe a limitation as it might seem, at least for questions of principle, because a most important quantity relevant to the driving forces is the ratio of surface-to-interior sites in the compact structure. For real three-dimensional proteins of chain length equal to 100 amino acids, the ratio of the number of surface/interior sites is about 2.3–4.0; for two-dimensional model chains, the same value of this ratio is obtained for chain lengths of $n = 16$, the chain length we explore here.

The purpose of the present paper is to apply this model to problems of the reversible denaturation of proteins and of the effects of mutations on stability. Our study here of H/P two-dimensional square lattice chains of length $n = 16$ monomers does not include the full sequence space; we focus only on the subset of sequences that fold to unique native structures. Whereas there are $2^{16} = 65{,}536$ sequences in the $n = 16$ sequence space, only 1,539 sequences fold uniquely to a single native conformation. We believe that this subset of sequences is most representative of real biological proteins. In the present analysis, the number of sequences counts both a sequence and its distinguishable reverse sequence, i.e.,

PHPHHPHHHPHHHHHH and

HHHHHHPHHHPHHPHP,

as distinct. For any $n = 16$ unique sequence there are $\Omega_0(16) = 802{,}075$ conformations on the square lattice (Chan & Dill, 1989b), one of which is the unique native (N) state, and the ensemble of all the rest of the conformations are defined to be the denatured (D) state. Each one of the $\Omega_0(16) - 1 = 802{,}074$ conformations in the D state will be referred to as a D conformation (see Fig. 1).

For any reversible process between two states, the equilibrium constant K_{eq} is the ratio of the probabilities that the system will be in one or the other state. Hence for the reversible folding–denaturation process, D $\rightleftharpoons$ N,

$$K_{eq} = \frac{P_N}{P_D}, \tag{1}$$

where P_N and P_D are the equilibrium probabilities of the N and D states, respectively. The free energy of folding ΔG, is given by

$$\Delta G = -kT \ln K_{eq}, \tag{2}$$

where k is the Boltzmann constant and T is the absolute temperature. If h_N is the number of HH contacts in the N state, standard statistical mechanical considerations imply that

$$\begin{aligned} P_N &= Q^{-1} \exp\left[-\frac{h_N \epsilon}{kT}\right], \\ P_D &= Q^{-1} \sum_{h=0}^{h_N - 1} g(h) \exp\left[-\frac{h\epsilon}{kT}\right], \end{aligned} \tag{3}$$

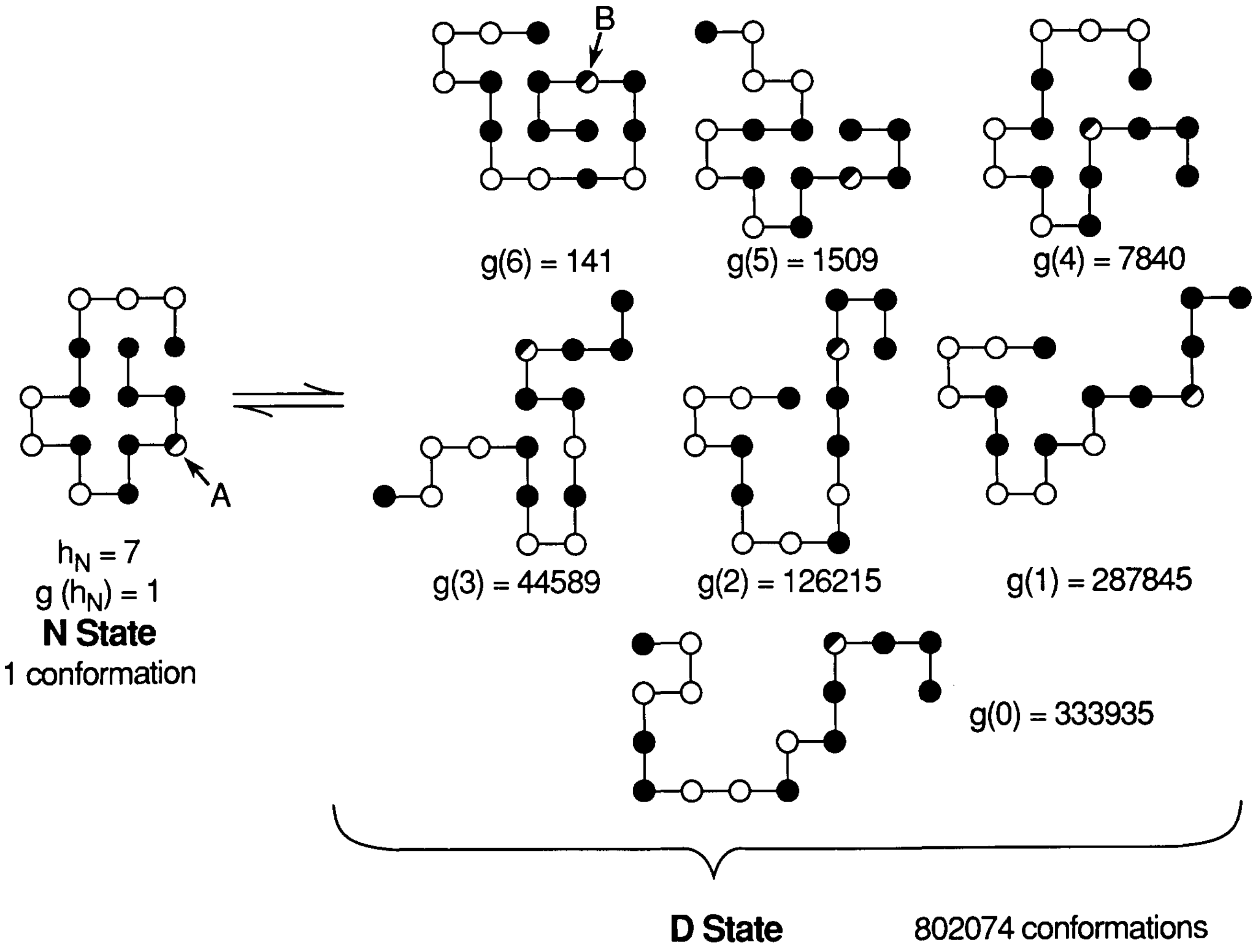

Fig. 1. The model N and D conformations for a particular sequence. Open, filled, and half-filled circles represent P and H monomers and the mutation site, respectively. The number of conformations with h HH contacts is $g(h)$. One example D conformation is shown for each value of h. Site A is a position that has no H interaction in the N state (a corner site); site B shows that the same monomer does have one H interaction (an edge site) in some of the dominant D conformations (the $g(6) = 141$ conformations are dominant under strong folding conditions). Thus H at that position *destabilizes* the N state. Note that the H/P sequence determines the number and the structure of these $h = 6$ HH-contact D conformations.

where $g(h)$ is the number of conformations with h HH contacts and Q is the partition function defined in the Appendix. For $n = 16$ square-lattice chains, h ranges from 0 to the maximum (Chan & Dill, 1989b) of 9, depending on the monomer sequence. It follows from Equations 1–3 that three factors are responsible for determining the equilibrium between the N and D states: (1) ϵ, the strength of the HH potential of mean force in the solvent, (2) h_N, the maximum number of HH contacts that are achievable in the unique N state of a given sequence, and (3) $g(h)$, the sequence-dependent D state distribution of the number of conformations as a function of the number of HH contacts. The statistical mechanics given in Equation 3 and in the Appendix is general for H/P copolymers, and is not restricted to lattice models or to two-dimensional models.

We focus below on two properties of the D ⇌ N equilibrium: (1) the stability, $\Delta G(\epsilon)$, and (2) the rate of change of stability with respect to HH contact free energy,

$$m(\epsilon) = \frac{\partial \Delta G(\epsilon)}{\partial \epsilon}. \tag{4}$$

The rate of change, m_x, of stability with respect to the change in any physical property x, is given by

$$m_x = \frac{\partial \Delta G}{\partial x} = \frac{\partial \Delta G(\epsilon)}{\partial \epsilon}\frac{\partial \epsilon}{\partial x} = m(\epsilon)\frac{\partial \epsilon}{\partial x}, \tag{5}$$

where $\partial\epsilon/\partial x$ is the rate of change of HH contact free energy with respect to the physical property x.

The study of these quantities in the model provides a conceptual framework for understanding the protein-folding equilibrium, especially with regard to native state stability and the experimentally determined "solvent denaturation slope,"

$$m_{\text{exp}} = \frac{\partial \Delta G}{\partial c}, \tag{6}$$

where c is the concentration of some denaturing agent in the solution. Commonly used in experiments are the solutes urea or guanidine hydrochloride (GuHCl) (Tanford, 1968; Shortle & Meeker, 1986; Shortle et al., 1990; Dill & Shortle, 1991). In particular, we consider here only the simplest denaturing agents, for which the hydrophobic interaction decreases essentially linearly with increasing denaturant concentration. Data from experiments on free energies of transfer of hydrophobic residues indicate that this approximation of linear functional dependence of ϵ on c is reasonably good for both urea and GuHCl, whereas the deviation from linearity is slightly larger for GuHCl than for urea (Nozaki & Tanford, 1963, 1970, 1971; Alonso & Dill, 1991). Hence for our purposes, the rate of change $\partial\epsilon/\partial c$ of ϵ with respect to concentration c may be taken to be constant. Therefore because $m_{\text{exp}} = m(\epsilon)\partial\epsilon/\partial c$, with $\partial\epsilon/\partial c$ a constant, it is most convenient to focus here simply on the behavior of $m(\epsilon)$.

Solvent effects on denaturation

How does the denaturation equilibrium depend on the strength of the HH attraction? It is clear from Equation 3 that if the HH attraction is strengthened ($|\epsilon|$ is increased, ϵ more negative), by changing the solvent conditions or other physical parameters, the N state will be favored relative to every D conformation. This is because the number h_N of HH contacts in the N state is by definition greater than the number h of HH contacts in any D conformation. The free energy of folding is shown as a function of ϵ in the inset of Figure 2. It is shown in the Appendix that the slope of this curve is

$$m(\epsilon) = h_N - \langle h \rangle_D(\epsilon) \tag{7}$$

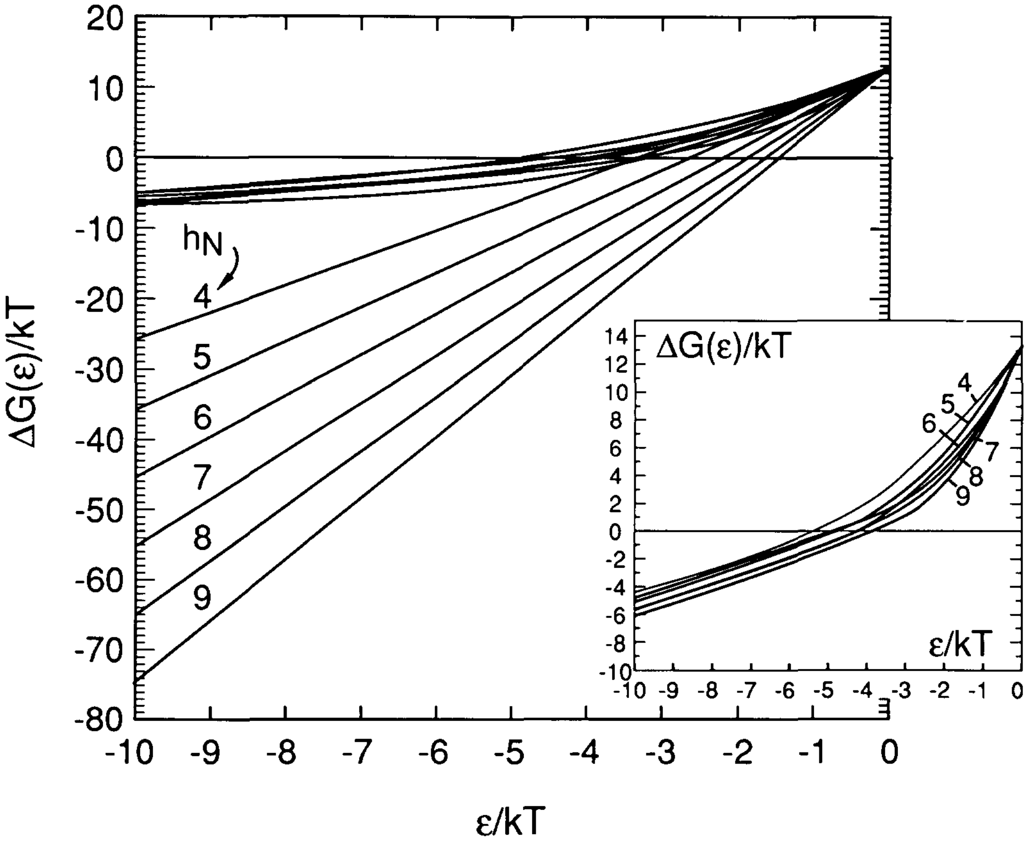

Fig. 2. Stability, $\Delta G(\epsilon) = G_N(\epsilon) - G_D(\epsilon)$, as functions of HH contact free energy ϵ for H/P unique sequences ($n = 16$). Both $\Delta G(\epsilon)$ and ϵ are given in units of kT. More negative values of $\Delta G(\epsilon)$ imply that the N states are more stable. Each curve that is concave upward shows $\Delta G(\epsilon)$ averaged over sequences with a specific number h_N of N state HH contacts, and is labeled by h_N in the inset. These curves are computed from the physically correct model that accounts for HH interactions in both the N and the D states. The straight lines in the main figure (labeled by h_N, which is equal to the slope of these straight lines) follows from the physically incorrect model discussed in the text, which arbitrarily assumes that there is no HH interaction in the D state. The inset shows more clearly the ϵ-dependence of the correct $\Delta G(\epsilon)$ on an expanded vertical scale.

for any H/P sequence, where $\langle h\rangle_D(\epsilon)$ is the ensemble-averaged number of HH contacts in the D state. In strong denaturing conditions ($|\epsilon| \approx 0$), there are on average relatively few HH contacts in the D state (i.e., $\langle h\rangle_D$ is small) and the slope m is mainly determined by h_N, the number of HH contacts, i.e., the amount of hydrophobic burial, in the N state. For some H/P sequences with $n = 16$, h_N is as small as 4 (relatively less stable N states); for other sequences, h_N is as large as 9 (relatively more stable N states). Under strong denaturing conditions ($|\epsilon| \approx 0$), the more hydrophobic contacts that a native protein has, the more strongly it is driven to fold up as the solvent is made to be increasingly poor for the hydrophobic residues, and thus $m(\epsilon)$ increases.

A more interesting result from the inset of Figure 2 is the behavior of chains in solvent conditions for which there is large HH attraction (large $|\epsilon|$). As $|\epsilon|$ increases, the average number h of HH contacts in the D state, $\langle h\rangle_D(\epsilon)$, increases, limiting at the highest h possible for the D conformations as $\epsilon \to -\infty$. For all the $n = 16$ unique sequences, the highest possible h is $h_N - 1$, one fewer than the number of HH contacts in the N state. It follows that in the strong HH attraction limit, the very small class of D conformations that have $h_N - 1$ HH contacts constitute the dominant population of the D state, hence $\langle h\rangle_D$ is approximately given by $h_N - 1$. Thus, by Equation 7 the slopes $m = h_N - \langle h\rangle_D \approx 1$ for all sequences. Independent of their stabilities, the m value becomes identical for all sequences in this limit. In the same large $|\epsilon|$ limit the stability of the N state is approximately given by (see Equation A.11 in the Appendix)

$$\Delta G(\epsilon) \approx \epsilon + kT \ln g(h_N - 1), \tag{8}$$

thus the stability of the N state is principally determined by the number $g(h_N - 1)$ of D conformations that have one fewer HH contact than the N state. In this limit, the denaturation midpoint ($\Delta G = 0$) is the point at which the extra contact gained by the N state is just counterbalanced by the higher degeneracy (conformational freedom) of the compact D conformations with $h = h_N - 1$.

It has been common to view denatured states of proteins as random flight chains that are highly expanded and that are highly exposed to the solvent; such conformations are observed for many different proteins in 6 M GuHCl (Tanford, 1968). In contrast, the results above indicate that the ensemble of denatured conformations depends on (1) ϵ, the solvent-dependent strength of the hydrophobic interaction, and (2) $g(h)$, the property of the monomer sequence that describes how the chain can configure to form all the possible nonnative clusters of H monomers. In solvents that strongly favor folding, the stability of the native state is principally determined by the number of denatured conformations that have the highest number of hydrophobic contacts (corresponding to $g(h_N - 1)$ in the model, see also Equation A.9 in the Appendix). This is a relatively small ensemble of denatured conformations; they have contact energies close to that of the native structure. Although we prefer the term "compact denatured states," similar conformations have also been referred to as the "molten globule" state.

To illustrate how different from this is the view of a fully exposed denatured state, we construct for comparison a hypothetical but physically incorrect model in which all denatured conformations are arbitrarily assumed to have zero contact interaction energy, as if all H residues were fully exposed to the solvent in the D state. According to this incorrect model, HH contact interactions are present only in the N state. The probability P_N for the N state is therefore identical to that of Equation 3,

$$P_N = Q^{-1} \exp\left[-\frac{h_N \epsilon}{kT}\right]. \tag{9}$$

However, because the D state is assumed to have no HH interaction, the $h\epsilon$ term in the second equation of Equation 3 is zero, thus

$$P_D = Q^{-1} \sum_{h=0}^{h_N - 1} g(h) \exp[0] = Q^{-1}[\Omega_0 - 1], \tag{10}$$

resulting in a sequence-independent D state that depends only on the total number of D conformations. This hypothetical model of D ⇌ N equilibrium is incorrect because it assumes that the HH interaction can be turned on or off depending on whether the chain is in the N state or in one of the D conformations. This is clearly physically unrealistic. The consequence of this "wrong" model for the D states is shown in Figure 2, and is compared with the physically correct Boltzmann-weighted D states model of Equation 3. In this "wrong" model, the D state carries no sequence information; only the sequence-dependent number h_N of HH contacts in the N state is important for stability. According to the "wrong" model, the slopes m are always equal to h_N and independent of ϵ. Consequently, the stabilities predicted by assuming that conformations of the D state are always fully exposed to the solvent are always larger than the true stabilities (see Fig. 2).

Single-site mutations

How are the stability and the denaturant slope of a protein affected by changes in its sequence of residues? To address these questions, we have taken as wild type all 1,539 16-mer sequences that encode unique native structures, and we have mutated every possible position, one at a time, from H to P or P to H, whichever is appropriate. Each mutant sequence was analyzed to determine if it encoded the same unique N state as the wild-type sequence by exhaustive conformational search. Of the 1,539 × 16 = 24,624 mutant sequences thus generated, 21,192 (86.06%) did not encode a unique N conformation and 704 (2.86%) encoded a unique N state different than that of the wild type, leaving 2,728 (11.08%) mutants that uniquely fold to the wild-type N structure; it is these mutations we selected for characterization. Thus, most (2,728/3,432 = 79.49%) mutations that generate sequences with unique N states are neutral in that they do not change the wild-type N states. However, in contrast to earlier studies of the maximally compact conformations (Lau & Dill, 1990), a large fraction of all possible mutations either does not retain a single N state or alters the native states.

The effects of mutations depend on ϵ. Most of the results below are presented for $\epsilon = -4kT$. For this value a majority of sequences have a stable N state; i.e., $\Delta G(\epsilon) < 0$. This value of ϵ, however, has no particular significance: qualitatively similar results are obtained for values of ϵ ranging from -1 to $-5kT$.

By symmetry, each P to H substitution that converts a wild-type sequence S to mutant sequence S′ will also be found when H to P substitution in S′ converts it to S. We consider only one direction here. In the present model, substitution mutations are classified into three types:

1) H0 → P0: In this class of substitution, an H residue that contacts no other residue in the N state is replaced by a P. These correspond to the corner positions in the square-lattice conformations. There are 824 such substitutions (and 824 P0 → H0 substitutions in the reverse direction). Two of the 824 substitutions do not have any effect on stability because the mutated monomer cannot form a contact with any H residue along the sequence and therefore they do not change the D state distribution $g(h)$. Because this is an artifact of the even–odd effect of the square lattice (Chan & Dill, 1989a, 1990a), we only consider below the other 822 H0 → P0 substitutions that lead to changes in $g(h)$.

2) HP → PP: In this class of substitution, an H residue that forms a topological contact with a P residue in the N state is replaced by a P. Because these residues are involved in contacts, they cannot be at lattice corners; rather they occur along outside edges of the lattice configuration of the N state. There are 104 such substitutions.

3) PH → HH: In this class of substitution, a P residue that forms a topological contact with an H in the N state is replaced by an H. These also correspond to edge positions in the square-lattice conformations. There are 434 such substitutions.

Only two of the 2,728/2 = 1,364 wild-type/mutant pairs involve a position that forms two contacts in the N state (in both cases with two P monomers). Presumably, interior mutations of monomers that have two or more N state contacts with other monomers almost invariably lead to multiply-degenerate N states or to an N state with a different structure.

Effects of mutations on stability

H0 → P0 mutations

Without exception (822/822), the replacement of a noncontacting surface H monomer with a P monomer *increases* the stability of the N state ($\Delta\Delta G < 0$). The stability increase is found to be quite substantial on average (top panel of Fig. 3A). The average gain in stability is approximately $0.25|\epsilon|$, and the maximum is approximately $0.6|\epsilon|$. This is a significant fraction of $|\epsilon|$, the maximum change possible for forming a single HH contact. These mutants by definition have the same N structure as the wild type, and the mutated monomer is at a position that makes no intrachain contacts in the N conformation. Therefore the free energy of the N state must be unchanged by the mutation, so any change in the D ⇌ N equilibrium *must be* due to changes in the D states. This is illustrated in Figure 1. Although the mutated monomer forms no contacts with other monomers in the N state, the same monomer is involved in intrachain contacts in the D state. It is clear that the number h of HH contacts in some of the D conformations in Figure 1 depends on whether the mutated monomer is an H or a P. It follows that the mutation must alter $g(h)$, the distribution of conformations in the D state. The reason that this type of mutation always stabilizes the N state is because replacing an H monomer by a P necessarily causes fewer HH contacts in the ensemble of D conformations, increasing the free energy of D. Figure 4A(i) shows the distribution $g(h)$ and the N structure of the sequence that undergoes the largest increase in stability after a single mutation of this type. The D conformations shift toward those with fewer HH contacts. The numbers of D conformations with one fewer ($h = h_N - 1$) and two fewer ($h = h_N - 2$) HH contacts than the N state are also shown in the same figure for the wild-type/mutant pair of sequences. These numbers, $g(h_N - 1)$ and $g(h_N - 2)$, are major determinants of stability under strong folding conditions (see Equation 8 and Equation A.11 in the Appendix).

How does the mutation in Figure 4A(i) increase the stability so much? The answer is most clearly seen by considering strongly folding conditions, for which the D ⇌ N equilibrium is determined by the balance between the unique N conformation ($g(h_N) = 1$, zero conformational entropy) with h_N HH contacts (low contact energy), and the larger number $g(h_N - 1)$ of D conformations (higher conformational entropy) that have one fewer HH contacts than the N state (higher contact energy). Figure 4A(i) shows that the wild type is capable of configuring into many D conformations with $h = h_N - 1$ ($g(h_N - 1) = 94$). Thus, the wild-type N state is relatively unstable due to this large entropic favorability of the compact D conformations with $h = h_N - 1$. On the other hand, the mutation leads to a sequence that has far fewer accessible highly compact D conformations ($g(h_N - 1) = 7$). The gain in stability, as calculated by substituting these num-

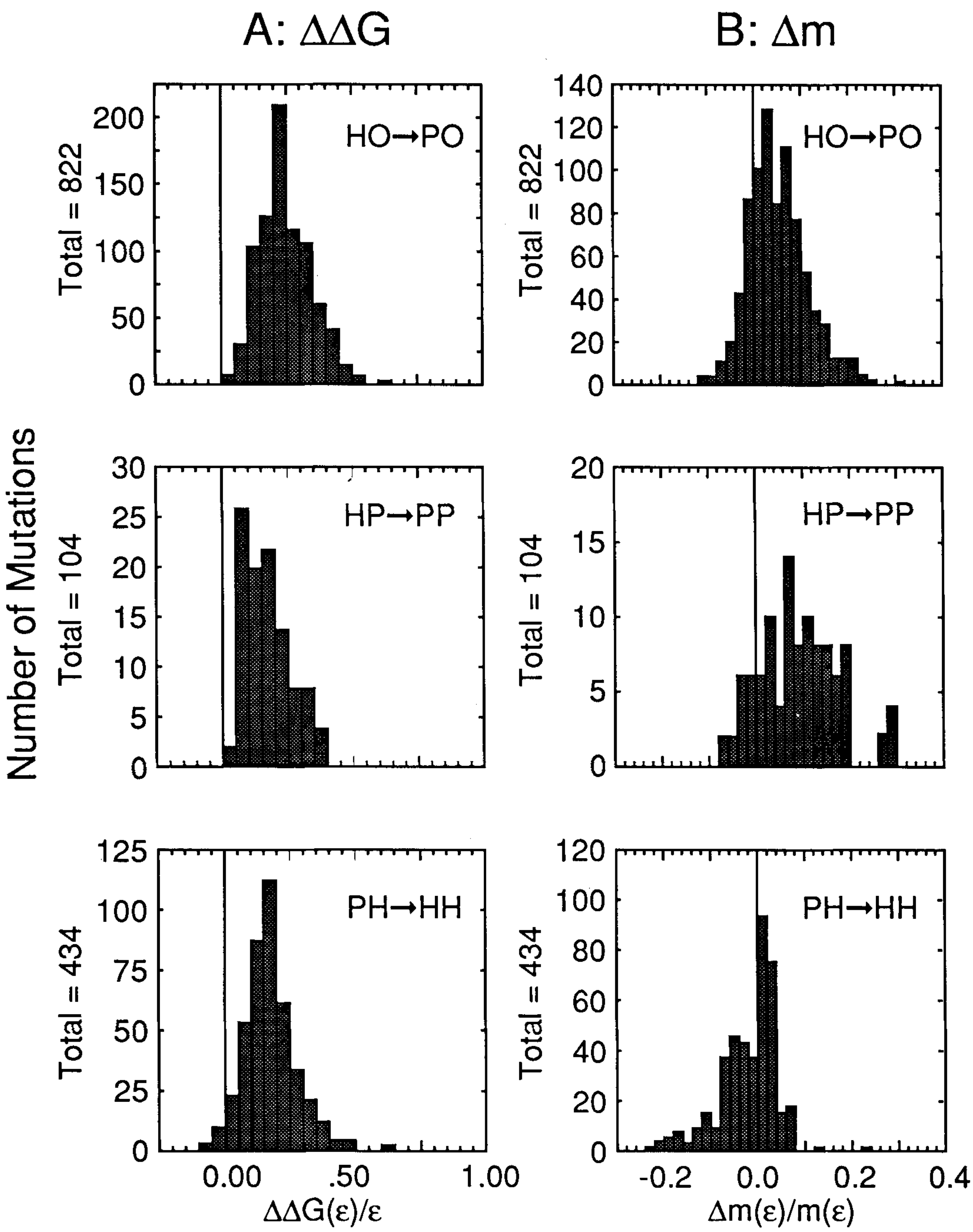

Fig. 3. **A**: Distribution of the change in the free energy of stabilization $\Delta\Delta G(\epsilon)$ due to mutations (given in units of the HH contact free energy ϵ, for $\epsilon = -4kT$). Because $\epsilon < 0$, positive changes in these plots indicate increases in N state stability; negative changes indicate decreases in N state stability. **B**: Fractional change $\Delta m(\epsilon)/m(\epsilon)$ $(\epsilon = -4kT)$.

bers into the approximate relation Equation 8, is accurate to 8% for $\epsilon = -4kT$. Thus, the main effect of the mutation is on the entropy of the highly compact D conformations with $h = h_N - 1$.

HP → PP mutations

The results from this class of mutations are very similar to those from the H0 → P0 class above. Again, the mutated monomers do not contribute to the HH contact free energy of the N states. Consequently, replacement of H by P increases the N state stability for all 104 HP → PP mutations. The distribution of stability gain is given in the middle panel of Figure 3A. The effect of this type of mutation on $g(h)$ is shown in Figure 4B(i), for a mutant that causes a maximal gain in stability. The number of D conformations with $h = h_N - 1$ is 31 before and 4 after the H0 → P0 mutation. This large decrease in $g(h_N - 1)$ accounts for the large gain in stability.

The main difference between the H0 → P0 above and

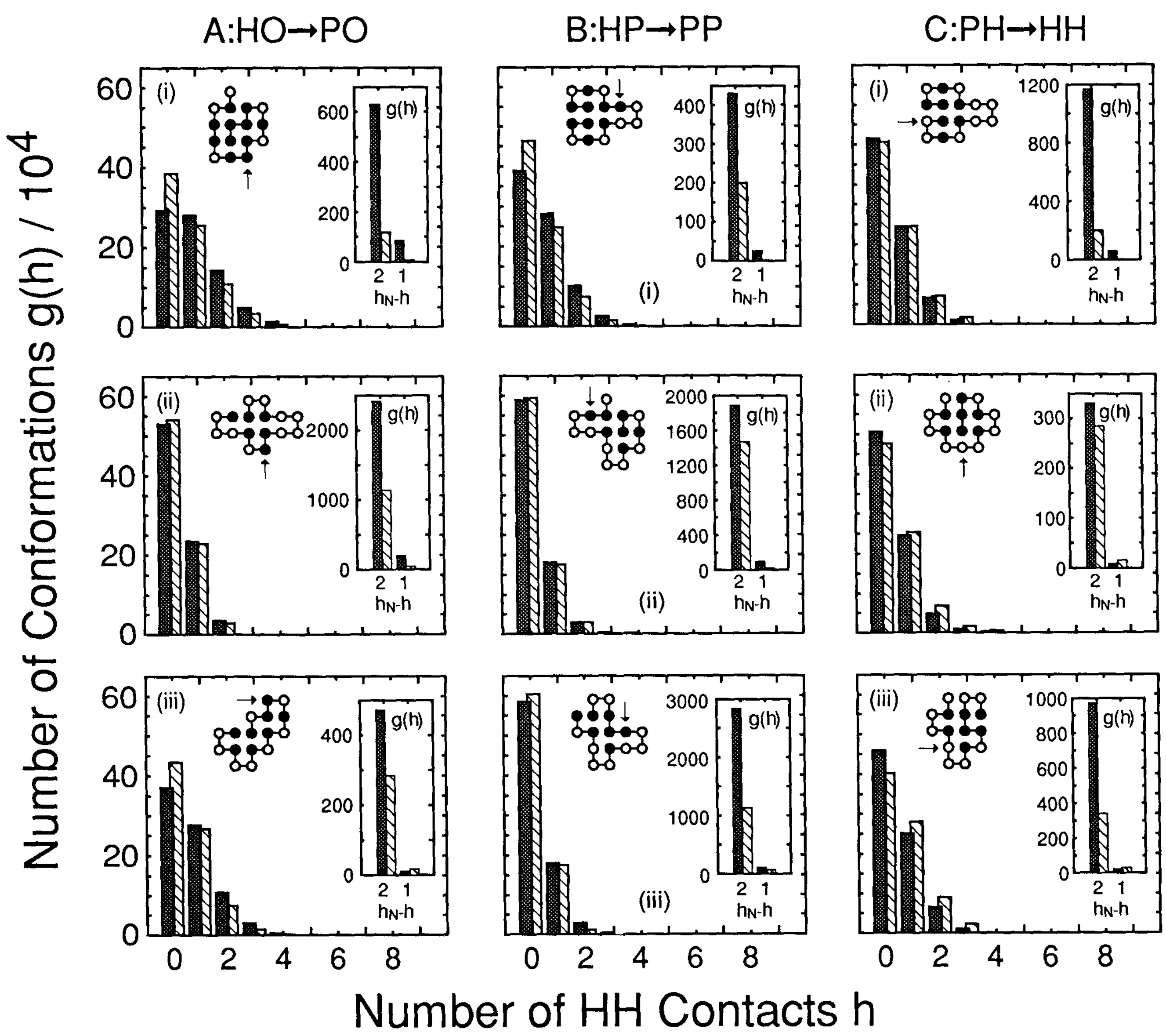

Fig. 4. Changes in the distributions $g(h)$ of the number h of HH contacts in the D state for sequences before (solid bars) and after (striped bars) mutations. Each wild-type sequence is shown in its N structure; mutation sites are denoted by arrows. The insets show $g(h)$ for $h = h_N - 1$ and $h = h_N - 2$, i.e., the number of D conformations with one-fewer and two-fewer HH contacts than the N states, before and after mutations. Note that the number of N-state HH contacts h_N is unchanged by mutations of classes A and B whereas class-C mutations always increase h_N by one.

the HP → PP mutations is that the average $\Delta\Delta G$ is significantly smaller for the latter. For instance the average $\Delta\Delta G(\epsilon)$ over all HP → PP mutations shown in the middle panel of Figure 3A for $\epsilon = -4kT$ is 0.174ϵ, only about 70% of that for the H0 → P0 mutations. Similar shifts in $\Delta\Delta G$ are observed for other ϵ. For ϵ ranging from -1 to $-5kT$, the average $\Delta\Delta G$ over the HP → PP mutations is 44–75% of the average $\Delta\Delta G$ from the H0 → P0 mutations. Because these mutated monomers can contribute to HH contact energy only in the D states, this observation implies that on average an H monomer before an HP → PP mutation is less efficient in forming HH contacts in the D state than an H monomer before an H0 → P0 mutation. This phenomenon can be explained in the context of the model. For any given jth monomer along the sequence, only contacts with monomers numbered $j-3, j-5, \ldots$, and $j+3, j+5, \ldots$ are possible. Each of these potential contacting monomers can either be an H or a P, depending on the sequence. If an H monomer to be mutated is already in contact with a P monomer in the N state (for HP → PP mutations), on average the number of its potential contacting partners that happen to be H's will be reduced relative to an H monomer to be mutated that does not have any intrachain contacts in the N state (in H0 → P0 mutations). This is confirmed by the observation that the average number of H monomers that are able to contact the mutated monomer in the D state for the H0 → P0 and HP → PP mutations is 2.52 (SD 0.79) and 1.77 (SD 0.72), respectively.

PH → HH mutations

This class of mutations differs from the other two in that it alters the energetics of the N state—an additional HH contact is formed in the N state. However, even these mutations do not always lead to a net stabilization of the N state. In 14 out of 434 cases, the outcome is a net loss of stability when $\epsilon = -4kT$. Also noteworthy is the fact that the average increase in stability (bottom panel of Fig. 3A) is only approximately 0.17ϵ for $\epsilon = -4kT$, less than the stability gain for the corresponding H0 → P0 class. Surprisingly, on average, N is stabilized more by removing a noncontacting H at the surface than by adding an additional HH contact.

Obviously, the stabilization that could be added by introducing a new HH contact in the N state is at most ϵ. However, any mutation of P to H will contribute not only to new HH contacts in the N state but also to new HH contacts in the D state, so in general the increased stability will be less than ϵ. How much smaller depends both on the value of ϵ and on the details of the sequence, through the distribution function $g(h)$. Because this type of mutation affects both the N and the D states, the change in stability is more complex than the two types of mutations discussed above. Although the other two types of mutations always stabilize, it is possible, as noted above, for PH → HH mutations to destabilize the N states.

Figure 4C(i) shows a pair of histograms of $g(h)$ for a wild type and its mutant, for the mutation that is maximally stabilizing. At $\epsilon = -4kT$, stability is mainly dependent upon $g(h_N - 1)$, the number of highly compact D conformations with $h = h_N - 1$. The PH → HH mutation changes $g(h_N - 1)$ from 72 to 4, which leads to the large gain in stability. The main effect of this mutation on stability is therefore caused by the decreased entropy of the highly compact D state rather than on the additional HH contact in the N state. Similar histograms for the most destabilizing mutation are shown in Figure 4C(ii). As anticipated from the arguments above, this mutation leads to an increase in $g(h_N - 1)$, in this case from 7 to 15.

Effects of mutations on the slope $m(\epsilon)$

The denaturant slope m_{exp} defined in Equation 6 and the stability ΔG reflect different aspects of the distribution of protein conformations and can be affected differently by mutations. A principal conclusion of this section is that whereas under strong folding conditions $g(h_N - 1)$ is a major determinant of stability, it is the ratio $g(h_N - 2)/g(h_N - 1)$ that is a major determinant of the denaturant slope. It is shown in the Appendix that if a wild-type sequence S is mutated to the mutant sequence S′, the change in $m(\epsilon)$ is

$$\Delta m(\epsilon) = m(\epsilon)' - m(\epsilon) = \Delta h_N + [\langle h\rangle_D(\epsilon)] - [\langle h\rangle_D(\epsilon)]'. \tag{11}$$

Here $\Delta h_N = h_N' - h_N$ is the change in the number h_N of HH contacts in the N state upon mutation, and $\langle h\rangle_D$ is the ensemble-averaged number of HH contacts in the D state.

Under strong folding conditions (large $|\epsilon|$), the mutational change $\Delta m(\epsilon)$ is approximately (see Equation A.17 in the Appendix)

$$\Delta m(\epsilon) \approx \left\{\left[\frac{g(h_N - 2)}{g(h_N - 1)}\right]' - \left[\frac{g(h_N - 2)}{g(h_N - 1)}\right]\right\} e^{\epsilon/kT}. \tag{12}$$

This shows that the behavior of $m(\epsilon)$ is more subtle than $\Delta G(\epsilon)$ at large $|\epsilon|$. Whereas $\Delta\Delta G$ is determined mainly by the change in the number $g(h_N - 1)$ of highly compact D conformations (see above), Δm is determined mainly by the change in $g(h_N - 2)/g(h_N - 1)$, the ratio of the number of D conformations with $h = h_N - 2$ relative to that with $h = h_N - 1$. In other words, the large $|\epsilon|$ limit of the m value of a sequence depends on the shape of the distribution $g(h)$ near $h = h_N$.

H0 → P0 mutations

As noted earlier, all of the mutations in this class increase the stability of the N state. However their effects on m are more subtle and diverse. The top panel of Figure 3B shows that for the 822 mutants in this class, mutation at $\epsilon = -4kT$ leads to a wide distribution of changes in m, some positive and some negative. Under strong folding conditions (large $|\epsilon|$), the change in m mainly depends on $g(h_N - 2)/g(h_N - 1)$. A majority of mutations leads to an increase in m ($\Delta m > 0$), which according to Equation 12 implies that $g(h_N - 2)/g(h_N - 1)$ increases. On the other hand a minority of mutations leads to $\Delta m < 0$, i.e., for these mutations $g(h_N - 2)/g(h_N - 1)$ decreases instead. Although all these mutations shift the D state ensemble to a smaller number of HH contacts, the relative partitioning of these changes among the subpopulations of compact conformations varies widely depending on the monomer sequence. The top panel of Figure 3B shows that a single mutation can change m by as much as 31% relative to the wild type. Because the HH interactions in the N state are not affected by this class of mutations, this remarkably large change in m is solely attributable to the way the mutation affects the distribution of conformations in the D state. Figure 4A(ii) and Figure 4A(iii) show examples of mutations that lead to a large increase and decrease in m, respectively. The ratio $g(h_N - 2)/g(h_N - 1)$ changes from 2,436/201 = 12.12 to 1,116/34 = 32.82 for the mutations with the largest increase ($\Delta m/m = 31\% > 0$) and $g(h_N - 2)/g(h_N - 1)$ changes from $477/15 = 29.8$ to $291/15 = 19.4$ for the largest decrease ($\Delta m/m = -10.4\% < 0$).

HP → PP mutations

This class of mutations shows the same patterns as the H0 → P0 mutations. A majority of mutations increase the value of m, but a significant fraction (16/104) decrease m (see the middle panel of Fig. 3B). Examples of the maximum increase and decrease in m at $\epsilon = -4kT$ are given in Figure 4B(ii, iii, respectively). The $g(h_N - 2)/g(h_N - 1)$ changes in the two cases are $1{,}904/106 = 17.96$ to $1{,}480/25 = 59.2$ ($\Delta m/m = 30\% > 0$), and $2{,}871/146 = 19.66$ to $1{,}183/104 = 11.38$ ($\Delta m/m = -8\% < 0$), respectively.

PH → HH mutations

Although most of the mutations of this class increase the stability of the N state (see above), the effects on m are about evenly divided—about half the mutants have higher m than wild type, and about half have lower m (see the bottom panel of Fig. 3B). Because the mutant chain has an extra HH contact in the N state, Δh_N in Equation 11 is equal to +1, in contrast with the other two classes of mutations, which have $\Delta h_N = 0$. However, the average number of HH contacts $\langle h \rangle_D$ always increases because of the additional H monomer, resulting in a negative contribution to $\Delta m(\epsilon)$. In many cases the latter negative contribution overcompensates the positive $\Delta h_N = +1$ contribution. Consequently, the distribution of Δm for this class of mutations is shifted toward more negative values relative to the distribution in the previous two classes of mutations. An example of this class of mutations, which leads to the largest decrease in m at $\epsilon = -4kT$, is given in Figure 4C(iii), the change in $g(h_N - 2)/g(h_N - 1)$ is from $972/14 = 69.43$ (wild type) to $334/18 = 18.56$ (mutant), with $\Delta m/m = -24\%$.

Summary

We have developed a simple model for the protein-folding equilibrium. It is based on short self-avoiding chains of H/P (hydrophobic/polar) copolymers on a two-dimensional square lattice. We compute the partition function for the D ⇌ N (denatured/native) equilibrium exactly by exhaustive computer enumeration, and we explore all the possible single-site mutations for the sequences that encode unique native conformations. The main results of this model are:

1) Mutations can have large effects on the entropy of the denatured states. By definition, the stability of a protein under native conditions is the free energy difference, $\Delta G = -kT \ln[\mathrm{N}]/[\mathrm{D}_0]$, where [N] is the concentration of the folded (native) species and $[\mathrm{D}_0]$ is the concentration of the denatured species under the same solution conditions (Dill & Shortle, 1991). The present model study indicates that the denatured population D_0 is dominated by a relatively small ensemble of compact conformations with many hydrophobic contacts. Under strong folding conditions, the dominant denatured species will be highly compact and will have the second highest possible number of hydrophobic contacts h. In the two-dimensional lattice model analyzed here, the second highest h is always one fewer than the native number of HH contacts ($h_N - 1$), and there are $g(h_N - 1)$ such conformations. These highly compact denatured conformations are determined by the amino acid sequence, just as the native conformation is. A major conclusion of the present work is that the amino acid sequence also determines *how many* highly compact denatured conformations there are. That is, the entropy of the denatured state of a protein under strong folding conditions is sequence dependent. A single mutation can have a large effect on the stability of the protein by controlling how many highly compact denatured conformations are available (i.e., $g(h_N - 1)$).

2) The model shows that there are upper limits to stabilization that can be expected from addition of hydrophobic monomers, in agreement with an earlier mean-field model (Dill, 1985; Alonso & Dill, 1991). As more H monomers are added to the sequence, more HH contacts become possible in the N state, but more HH contacts also occur in the conformations of the D state.

3) Two puzzling experimental results can be rationalized by the principles that emerge from the model. First, the model shows that a protein can be *destabilized* by replacing a surface residue by a more hydrophobic amino acid. This situation arises if the mutation site is at a position in the structure that makes no hydrophobic contacts in the native state, because the added hydrophobic residue can potentially make hydrophobic contacts in the denatured state, resulting in destabilization. Such a result has been observed experimentally (Pakula & Sauer, 1990), where it has been referred to as the "reverse hydrophobic effect." In our view, it is important not to misconstrue the term "reverse hydrophobic effect" to imply some special solvation process, because, according to the present model, the effect can arise simply from greater burial of nonpolar surface in the denatured than in the native state. Likewise, replacing a buried polar residue by a hydrophobic residue may not stabilize a protein, because the free energy of the denatured state may be lowered by a greater amount than the free energy of the native state.

Second, there is now much experimental evidence for mutational changes in denaturation behavior that cannot be readily accounted for by changes in the native state alone (Shortle & Meeker, 1986; Shortle et al., 1990). One example is the very broad distribution of $m_{\mathrm{exp}} = \partial\Delta G/\partial c$ values for the mutational changes of stabilities of staphylococcal nuclease (see Fig. 5B). Simple thermodynamic models and mean-field statistical mechanical models (Alonso & Dill, 1991; Dill & Shortle, 1991) would predict that if one monomer is altered in a chain of n monomers, then the change in m value should be approximately of magnitude $1/n$, i.e., a few percent at most. However, the present model, which takes into account the effects of amino acid sequence on the denatured states, shows that the effect can be considerably larger than this; Figure 5A shows predicted changes of up to 31%. Consistent with this, the experiments by Shortle and coworkers on staphylococcal nuclease show similar distributions, and with single mutational change on m by up to 30-40%. The reason for such large changes, according to the model, is that the denaturant slope is determined by both the native structure and the small ensemble of highly compact denatured conformations. The distribution of hydrophobic contacts in this small ensemble is highly sensitive to changes in the amino acid sequence and therefore can be changed dramatically by single mutations.

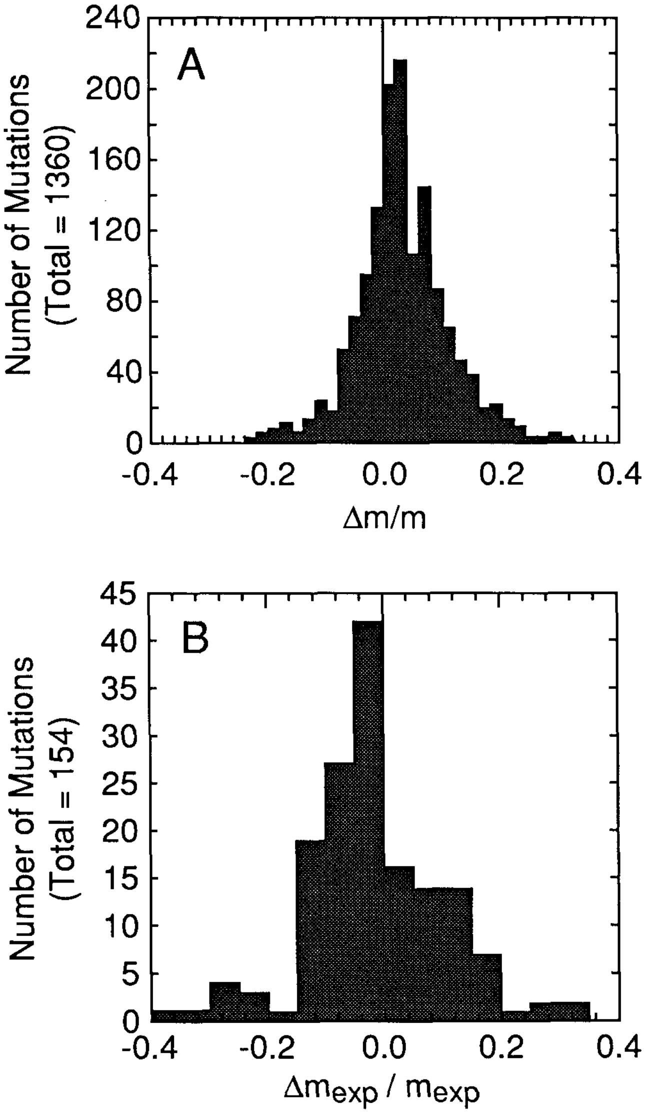

Fig. 5. **A**: Distribution of fractional change $\Delta m(\epsilon)/m(\epsilon)$ $(\epsilon = -4kT)$ for all mutations. **B**: Distribution of fractional change $\Delta m_{\mathrm{exp}}/m_{\mathrm{exp}}$ in denaturant slope for 154 single mutations on staphylococcal nuclease (D. Shortle et al., unpubl.). This includes substitutions of phenylalanine, isoleucine, leucine, methionine, asparagine, proline, glutamine, serine, threonine, valine, and tyrosine residues to both alanine and glycine, as well as substitutions of alanine to glycine and glycine to alanine.

Although the present prediction of large mutational effects is based on results on short chains, we believe the same principles should apply to longer chains. This is because under strong folding conditions the D $\rightleftharpoons$ N equilibrium is mainly determined by the distribution $g(h)$ for h values slightly less than h_{N}. In the compact conformations of copolymers, the clustering of hydrophobic residues should be spatially localized and relatively independent of chain length, in analogy with other conformational properties of polymers that become independent of n at high densities (de Gennes, 1979). However, the validity of this opinion remains to be tested on longer chains.

Acknowledgments

We thank the URI Program of DARPA and NIH (grants GM34171 to D.S. and GM34993 to K.A.D.) for financial support.

References

Alonso, D.O.V. & Dill, K.A. (1991). Solvent denaturation and stabilization of globular proteins. *Biochemistry 30*, 5974-5985.

Chan, H.S. & Dill, K.A. (1989a). Intrachain loops in polymers: Effects of excluded volume. *J. Chem. Phys. 90*, 492-509.

Chan, H.S. & Dill, K.A. (1989b). Compact polymers. *Macromolecules 22*, 4559-4573.

Chan, H.S. & Dill, K.A. (1990a). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys. 92*, 3118-3135.

Chan, H.S. & Dill, K.A. (1990b). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA 87*, 6388-6392.

Chan, H.S. & Dill, K.A. (1991a). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem. 20*, 447-490.

Chan, H.S. & Dill, K.A. (1991b). Sequence space soup of proteins and copolymers. *J. Chem. Phys. 95*, 3775-3787.

de Gennes, P.-G. (1979). *Scaling Concepts in Polymer Physics.* Cornell University Press, Ithaca, New York.

Dill, K.A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry 24*, 1501-1509.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry 29*, 7133-7155.

Dill, K.A. & Shortle, D. (1991). Denatured states of proteins. *Annu. Rev. Biochem. 60*, 795-825.

Lau, K.F. & Dill, K.A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*, 3986-3997.

Lau, K.F. & Dill, K.A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA 87*, 638-642.

Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S., & Dill, K.A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys. 96*, in press.

Nozaki, Y. & Tanford, C. (1963). The solubility of amino acids and related compounds in aqueous urea solutions. *J. Biol. Chem. 238*, 4074-4081.

Nozaki, Y. & Tanford, C. (1970). The solubility of amino acids, diglycine and triglycine in aqueous guanidine hydrochloride solutions. *J. Biol. Chem. 245*, 1648-1652.

Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions—Establishment of a hydrophobicity scale. *J. Biol. Chem. 246*, 2211-2217.

Pakula, A.A. & Sauer, R.T. (1990). Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature 344*, 363-364.

Schellman, J.A. (1978). Solvent denaturation. *Biopolymers 17*, 1305-1322.

Shortle, D. & Meeker, A.K. (1986). Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation. *Proteins Struct. Funct. Genet. 1*, 81-89.

Shortle, D., Stites, W.E., & Meeker, A.K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry 29*, 8033-8041.

Tanford, C. (1968). Protein denaturation. *Adv. Protein Chem. 23*, 121-282.

Appendix: The statistical mechanics of the model

The following is the general theory of conformational change in H/P copolymers and proteins. The theory is not restricted to lattice models and applies to any spatial dimensionality.

Stability and the partition function

The probability of a single conformation with h HH contacts is proportional to

$$\exp\left[-\frac{h\epsilon}{kT}\right], \tag{A.1}$$

where $\epsilon < 0$ is the contact free energy per HH contact, k is the Boltzmann constant, and T is the absolute temperature. Hence the probability that the chain adopts conformation(s) with h HH contacts is

$$P(h) = g(h)e^{-h\epsilon/kT}/Q, \tag{A.2}$$

where $g(h) \geq 0$ is the number of conformations that have h HH contacts, and the summation

$$Q \equiv \sum_{h=0}^{h_N} g(h)e^{-h\epsilon/kT} \tag{A.3}$$

is the partition function; hence $\Sigma_{h=0}^{h_N} P(h) = 1$. Here h_N is the maximum number of HH contacts for the specific sequence. Because we are primarily interested in sequences with unique native structures, i.e., $g(h_N) = 1$, h_N is the number of HH contacts in the unique native structure of a particular sequence. (The quantity $g(h_N)$ is identical to the degeneracy g in Chan and Dill [1991b].) As the denatured (D) state is the ensemble of all conformations except the unique native (N) state, the probability P_N and P_D of the N and the D state are, respectively,

$$P_N(\epsilon) = \frac{e^{-h_N\epsilon/kT}}{Q} \tag{A.4a}$$

and

$$P_D(\epsilon) = \frac{\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}}{Q}. \tag{A.4b}$$

The uniqueness of the N state implies that its free energy $G_N(\epsilon)$ is determined by the sum of HH contact free energies,

$$G_N(\epsilon) = -kT\ln P_N(\epsilon) = h_N\epsilon + kT\ln Q, \tag{A.5}$$

whereas the free energy $G_D(\epsilon)$ of the D state consists of contributions from the ensemble of all other conformations except the N state,

$$\begin{aligned} G_D(\epsilon) &= -kT\ln P_D(\epsilon) \\ &= -kT\ln\left[\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}\right] + kT\ln Q. \end{aligned} \tag{A.6}$$

Thus, the stability of the N state relative to the D state is given by the difference in free energy,

$$\begin{aligned} \Delta G(\epsilon) &\equiv G_N(\epsilon) - G_D(\epsilon) \\ &= h_N\epsilon + kT\ln\left[\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}\right]. \end{aligned} \tag{A.7}$$

Hence the stability of the N conformation is a balance: the N state is stabilized by its h_N HH contacts but is destabilized by the multiplicity $g(h)$ of D conformations. The N state becomes highly populated as the HH attraction becomes strong (large $|\epsilon|/kT$), but under this condition the dominant species of D in the summation of Equation A.7 are the few conformations with the largest h (i.e., the most compact D states). Under these conditions of strong HH attraction, Equation A.7 may be expressed as

$$\begin{aligned} \Delta G(\epsilon) &= h_N\epsilon + kT\ln\Bigg\{e^{-(h_N-1)\epsilon/kT}g(h_N-1) \\ &\quad\times\left[1 + e^{\epsilon/kT}\frac{g(h_N-2)}{g(h_N-1)} + e^{2\epsilon/kT}\frac{g(h_N-3)}{g(h_N-1)} + \ldots\right]\Bigg\} \\ &= \epsilon + kT\ln g(h_N-1) \\ &\quad + kT\ln\left[1 + e^{\epsilon/kT}\frac{g(h_N-2)}{g(h_N-1)} + e^{2\epsilon/kT}\frac{g(h_N-3)}{g(h_N-1)} + \ldots\right], \end{aligned} \tag{A.8}$$

for sequences with $g(h_N - 1) \neq 0$. The generalization of this result is straightforward; if $h_N - j$ is the highest number of HH contacts possible among the D state conformations, the above expression is instead

$$\Delta G(\epsilon) = j\epsilon + kT \ln g(h_N - j) + kT \ln\left[1 + e^{\epsilon/kT}\frac{g(h_N - j - 1)}{g(h_N - j)} + e^{2\epsilon/kT}\frac{g(h_N - j - 2)}{g(h_N - j)} + \ldots\right]. \quad (A.9)$$

Because $j = 1$ for all $n = 16$ sequences considered here, subsequent analysis will only be presented for the case of Equation A.8. It is straightforward to obtain results for the general case by using Equation A.9 instead of Equation A.8 in subsequent derivations.

The last logarithmic term in Equation A.8 can be further expanded using the Taylor expansion,

$$\ln(1 + x) = -\sum_{s=1}^{\infty}\frac{(-x)^s}{s} = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots. \quad (A.10)$$

The stability ΔG can then be expressed as a power series in $e^{\epsilon/kT}$,

$$\Delta G(\epsilon) = \epsilon + kT \ln g(h_N - 1) + kT\left[e^{\epsilon/kT}\frac{g(h_N - 2)}{g(h_N - 1)} + e^{2\epsilon/kT}\left\{\frac{g(h_N - 3)}{g(h_N - 1)} - \frac{1}{2}\left[\frac{g(h_N - 2)}{g(h_N - 1)}\right]^2\right\} + O(e^{3\epsilon/kT})\right], \quad (A.11)$$

where $O(e^{3\epsilon/kT})$ stands for all terms with $e^{\epsilon/kT}$ raised to the third and higher powers. When the HH attraction is very strong, $\epsilon/kT \to -\infty$, all powers of $e^{\epsilon/kT}$ tend to zero. Consequently, the stability ΔG of the N state is primarily determined by the relative magnitude of ϵ and the entropic free energy contributed by the conformational freedom of the $h_N - 1$ ("first excited" state) conformations. Only when the HH attraction is not too strong can conformations that have fewer than $h_N - 1$ HH contacts have more significant effects on stability.

Rate of change of stability with respect to HH contact energy

The rate of change of ΔG with respect to the contact free energy ϵ is given by the derivative of Equation A.7,

$$m(\epsilon) \equiv \frac{\partial \Delta G(\epsilon)}{\partial \epsilon} = h_N - \frac{\sum_{h=0}^{h_N-1} hg(h)e^{-h\epsilon/kT}}{\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}} = h_N - \langle h \rangle_D(\epsilon), \quad (A.12)$$

where $\langle h \rangle_D(\epsilon)$ is the average number of HH contacts in the denatured (D) state. Thus $m(\epsilon)$ is the difference between the number of HH contacts in the N state and the average number of HH contacts in the D state. If the contacting area per HH contact is given by C_A, and the exposed area of an H monomer is defined as a quantity proportional to the number of contacts it makes with P monomers or solvent, i.e., all contacts except those with other H monomers, then the exposed H area A_N in the N state is

$$A_N = A_0 - 2C_A h_N, \quad (A.13)$$

where A_0 is the exposed H area when all H monomers are fully exposed. This is because two units of exposed area are buried per contact pair. Similarly, the average exposed H area in the D state is

$$\langle A \rangle_D(\epsilon) = A_0 - 2C_A\langle h \rangle_D(\epsilon). \quad (A.14)$$

Hence it follows from Equation A.12 that

$$m(\epsilon) = \frac{1}{2C_A}[\langle A \rangle_D(\epsilon) - A_N], \quad (A.15)$$

i.e., $m(\epsilon)$ is proportional to the difference between the average exposed area of H monomers in the D state and the exposed area of H monomers in the N state, as proposed by Schellman (1978).

The second derivative of ΔG yields

$$\frac{\partial m(\epsilon)}{\partial \epsilon} \equiv \frac{\partial^2 \Delta G(\epsilon)}{\partial \epsilon^2} = \frac{1}{kT}\left\{\frac{\sum_{h=0}^{h_N-1} h^2 g(h)e^{-h\epsilon/kT}}{\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}} - \left[\frac{\sum_{h=0}^{h_N-1} hg(h)e^{-h\epsilon/kT}}{\sum_{h=0}^{h_N-1} g(h)e^{-h\epsilon/kT}}\right]^2\right\} = \frac{1}{kT}\{\langle h^2 \rangle_D(\epsilon) - [\langle h \rangle_D(\epsilon)]^2\}, \quad (A.16)$$

where $\langle h^2 \rangle_D$ is the mean square number of HH contacts in the D state. The last line in Equation A.16 indicates that the second derivative of ΔG is equal to the variance of the number of HH contacts in the D state divided by kT.

It is obvious from Equations A.12 and A.16 that both $m(\epsilon)$ and $\partial m(\epsilon)/\partial\epsilon$ are nonnegative. Hence all ΔG vs. ϵ curves have positive slopes and are concave upward (Fig. 2). When there is little or no attraction between H monomers ($\epsilon \rightarrow 0$), the favored D conformations are open and have small h. On the other hand, when the HH attraction is very strong ($\epsilon \rightarrow -\infty$), the favored D conformations are the most compact, with large $h \approx h_N - 1$. A broader distribution, i.e., a larger variance, of h is only possible at intermediate values of ϵ. Hence by Equation A.16, ΔG vs. ϵ has its largest curvature at intermediate ϵ (see Fig. 2).

It is useful to express $m(\epsilon)$ as a power series in $e^{\epsilon/kT}$. Differentiation of Equation A.11 yields

$$m(\epsilon) = 1 + e^{\epsilon/kT}\frac{g(h_N-2)}{g(h_N-1)} + e^{2\epsilon/kT}\left\{2\frac{g(h_N-3)}{g(h_N-1)} - \left[\frac{g(h_N-2)}{g(h_N-1)}\right]^2\right\} + O(e^{3\epsilon/kT}). \quad \text{(A.17)}$$

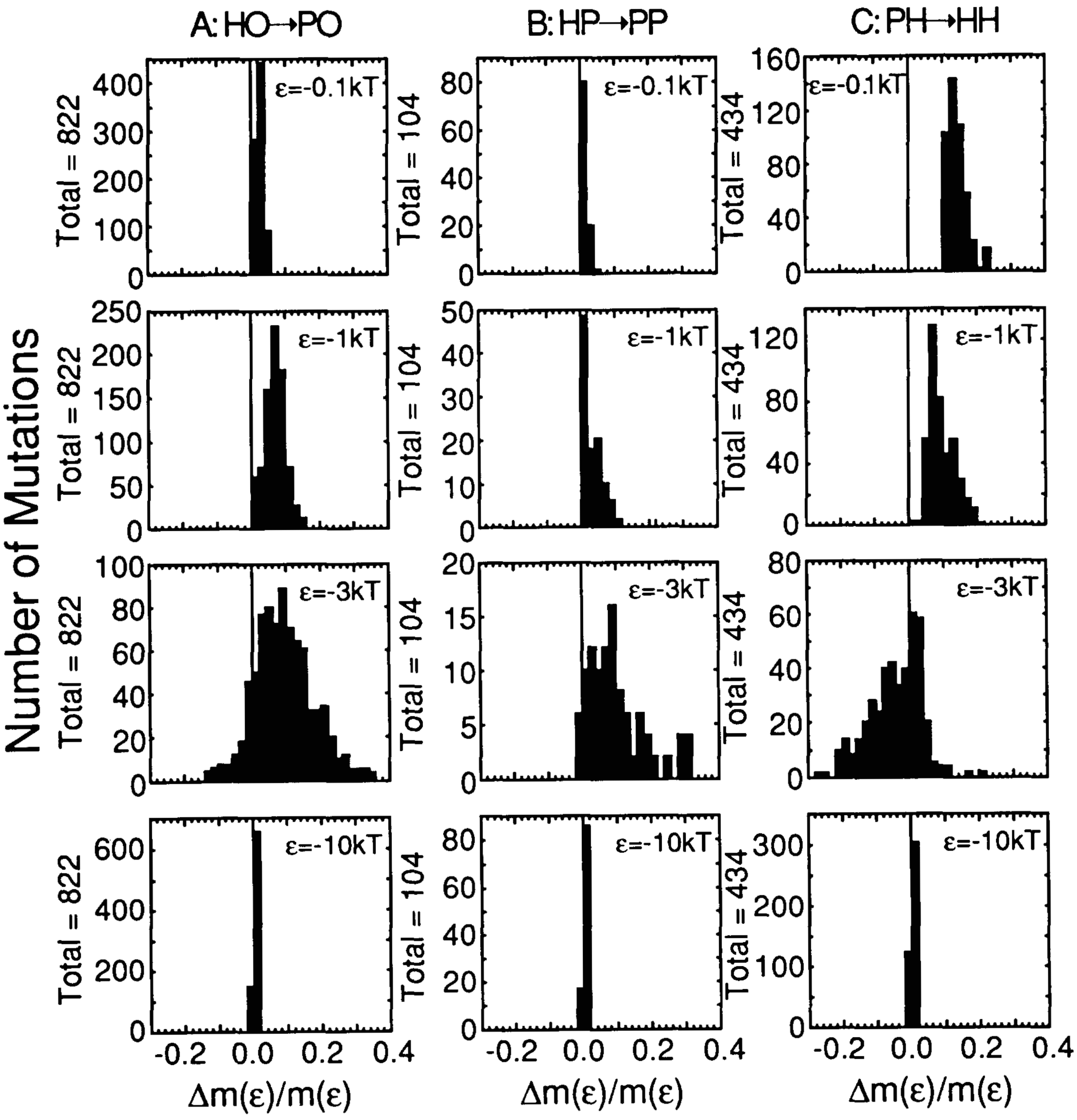

Fig. 6. Distribution of fractional change $\Delta m(\epsilon)/m(\epsilon)$ as a function of HH contact free energy ϵ (for $\epsilon = -0.1, -1, -3$, and $-10kT$).

When the HH attraction is very strong ($\epsilon \rightarrow -\infty$), the m value for all sequences with $g(h_N - 1) > 0$ reduces to unity. When the HH attraction is sufficiently strong, Equation A.17 shows that the m value is essentially determined by the ratio of the number of conformations at the $h = h_N - 1$ and $h = h_N - 2$ levels. This feature is illustrated by the examples discussed in the text (Fig. 4). At the other extreme, when there is little or no HH attraction ($\epsilon \rightarrow 0$), the m value tends to be large. This follows from Equation A.12 and the fact that $\langle h \rangle_D$ is small for small $|\epsilon|$. The exact value of $m(\epsilon)$ at $\epsilon = 0$ is sequence dependent.

Effects of mutations

When a mutation changes sequence S to sequence S′, the change in stability and m value are, respectively,

$$\Delta\Delta G(\epsilon) \equiv \Delta G(\epsilon)' - \Delta G(\epsilon), \qquad (A.18)$$

and

$$\begin{aligned} \Delta m(\epsilon) &\equiv m(\epsilon)' - m(\epsilon) \\ &= \Delta h_N + [\langle h \rangle_D(\epsilon)] - [\langle h \rangle_D(\epsilon)]', \\ \Delta h_N &\equiv h_N' - h_N, \end{aligned} \qquad (A.19)$$

where the unprimed and primed terms correspond to the wild-type sequence S and the mutated sequence S′, respectively. Note that when the HH attraction is very strong,

$$\lim_{\epsilon \rightarrow -\infty} \Delta m(\epsilon) = 0 \qquad (A.20)$$

by Equation A.17, implying that no mutation can lead to a change in m value at infinite $|\epsilon|$. It follows that for large $|\epsilon|$ the distribution of $\Delta m(\epsilon)$ over all possible mutations should peak narrowly around zero.

The ϵ-dependence of $\Delta m(\epsilon)$ is illustrated by Figure 6, which shows the distribution of $\Delta m(\epsilon)$ at different ϵ for mutations of $n = 16$ sequences. Because of Equation A.20, $\Delta m(\epsilon)$ approaches zero for all mutations at large ϵ (the three bottom panels of Fig. 6 for $\epsilon = -10kT$).

Because the variation in h is largest at intermediate ϵ for individual sequences (see above), the variation in $\langle h \rangle_D$ among sequences is also expected to be largest at intermediate ϵ. Because the variation in $\Delta m(\epsilon)$ among sets of mutations with a specific Δh_N is dependent upon the change in $\langle h \rangle_D$ before and after the mutation (Equation A.19), this variation is also largest at intermediate ϵ (compare the three $\epsilon = -3kT$ panels of Fig. 6 with results at other ϵ).

According to Equation A.19, $\Delta m(\epsilon)$ is controlled by the change in h_N and the change in $\langle h \rangle_D$. The change in h_N is zero for H0 → P0 (Fig. 6A) and HP → PP (Fig. 6B) mutations, therefore $\Delta m(\epsilon)$'s for these two classes of mutations are determined only by the change in the average number of HH contacts $\langle h \rangle_D$ in the D state. For small $|\epsilon|$, the distribution $g(h)$ is concentrated at small h for all sequences, hence it is impossible to have a large mutational change in $\langle h \rangle_D$. This is why the distributions of $\Delta m(\epsilon)$ in Figure 6A and B (top panels) at $\epsilon = -0.1kT$ have a narrow peak near zero. For the PH → HH mutations (Fig. 6C), the number of N state HH contacts increases by one; thus the distribution of $\Delta m(\epsilon)$ is shifted to more positive values at small $|\epsilon|$ because of the positive contribution from $\Delta h_N = +1$. However, in this case $\langle h \rangle_D - \langle h \rangle_D'$ is negative because an additional H monomer always tends to increase the average number of HH contacts in the D states. At intermediate $|\epsilon|$, when the variance in h is largest, this negative contribution from the change in $\langle h \rangle_D$ overcompensates the positive $\Delta h_N = +1$ contribution and thus leads to a net decrease in $m(\Delta m < 0)$ for many sequences, as shown in Figure 6C ($\epsilon = -3kT$ panel).