PROTEINS: Structure, Function, and Genetics 24:335–344 (1996)

Comparing Folding Codes for Proteins and Polymers

Hue Sun Chan and Ken A. Dill
Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-1204

ABSTRACT **Proteins fold to unique compact native structures. Perhaps other polymers could be designed to fold in similar ways. The chemical nature of the monomer "alphabet" determines the "energy matrix" of monomer interactions—which defines the folding code, the relationship between sequence and structure. We study two properties of energy matrices using two-dimensional lattice models: uniqueness, the number of sequences that fold to only one structure, and encodability, the number of folds that are unique lowest-energy structures of certain monomer sequences. For the simplest model folding code, involving binary sequences of H (hydrophobic) and P (polar) monomers, only a small fraction of sequences fold uniquely, and not all structures can be encoded. Adding strong repulsive interactions results in a folding code with more sequences folding uniquely and more designable folds. Some theories suggest that the quality of a folding code depends only on the number of letters in the monomer alphabet, but we find that the energy matrix itself can be at least as important as the size of the alphabet. Certain multi-letter codes, including some with 20 letters, may be less physical or protein-like than codes with smaller numbers of letters because they neglect correlations among inter-residue interactions, treat only maximally compact conformations, or add arbitrary energies to the energy matrix.** © 1996 Wiley-Liss, Inc.

Key words: **protein design, synthetic heteropolymer design, energy matrix, sequence degeneracy, structural encodability**

INTRODUCTION

Proteins are special among polymers because they can fold to specific unique compact structures. Proteins have a *folding code,* that is, the native three-dimensional structure of a protein is encoded in its monomer sequence by virtue of two properties of the monomer set: (1) the *alphabet size,* the number of letters in the code (for proteins this number is 20, representing the set of naturally occurring amino acids); and (2) the *energy matrix,* the set of energies that describe the interactions among the monomers. For example, the simplest model of the folding code of proteins involves a two-letter alphabet of monomer types, hydrophobic (H) and polar (P), called the HP model[1–8] (see also ref. 9 and references therein). Many simple folding codes have been studied recently using lattice models.[1–36] Here we study the different folding codes that result from different energy matrices, for the following purposes: (1) to learn about sequence/structure relationships of proteins; (2) to assess different simplified models, which are based on different energy matrices; and (3) to explore how protein-like and non-protein-like structures might be designed into *foldable polymers* using other monomer chemistries.

We study two properties of folding codes—*degeneracy* and *encodability.* The degeneracy of a monomer sequence is its number of lowest-energy conformations. For example, one sequence for a given monomer alphabet might have only a single lowest-energy conformation, whereas another might have 100 different conformations with equal lowest possible energy. In these cases, we refer to the sequences as having degeneracies 1 or 100, respectively. Although the sequences of globular proteins have very low degeneracies in general, their degeneracies are not always equal to 1. Often there are flexible loops and other forms of conformational freedom in real proteins. Some of these properties are captured, at least qualitatively, by the HP folding code for chains configured in two dimensions—most sequences fold to very few stable states, sometimes with degeneracies equal to 1, but sometimes more.[1–9]

Whereas degeneracy is a property of a *sequence,* encodability is a property of a *structure.* A chain conformation is defined as encodable if it is the unique native (lowest-energy, or ground-state) conformation of at least one sequence. Not all chain conformations are encodable by sequences of monomers having only spatially short range interactions. *Open* chain conformations are not encodable because there is no monomer sequence, of any set of such monomer types, that can cause that conformation to be more stable than the many other open conformations of similar energies. Hence an open conforma-

Received August 28, 1995; revision accepted October 12, 1995.

Address reprint requests to Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, CA 94143-1204.

tion is not encodable. To be encodable, a conformation must be relatively compact, but even some compact conformations are not encodable, depending on the energy matrix. It sometimes happens, at least in lattice models, that no sequence can be designed that will cause a particular target conformation, a, to be lower in energy than some other conformation, b. Then conformation a is not encodable. Less than 20% (for chain lengths $n \leq 14$) of the maximally compact conformations are encodable in a binary HP code in two-dimensional lattice models[3]; the fraction is probably smaller in three dimensions.[7,8] Moreover, the ability to encode structures requires a monomer alphabet containing more than one letter. No structure is encodable in a homopolymer except for very special cases such as the rod-like conformation of a highly charged chain.

MODELS

Here we study how different energy matrices determine sequence degeneracy and structure encodability in simple lattice models. We consider short chains having only pairwise nearest-neighbor contact interactions on two-dimensional square lattices. The full space of conformations is determined by exhaustive enumeration for all possible sequences with $n \leq 18$ monomers. The ij element ϵ_{ij} of the energy matrix $\mathscr{E}$ gives the energy of a nearest-neighbor contact between any monomer of chemical species i with any monomer of species j. We consider several different model energy matrices below. In the HP model, contacts of type (HH, HP, PP) have energies of $(-|\epsilon|, 0, 0)$.[1] Since the encoding of a native fold is unchanged when all ϵ_{ij}s are uniformly scaled by a positive multiplicative constant, the energy matrix $\mathscr{E}$ is only specified up to a positive constant. For instance, the energy matrix of the HP model may be expressed as

$$\mathscr{E} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{P} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{P} \end{array} \\ \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} \end{array} \qquad (1)$$

where all entries have been scaled by $|\epsilon|^{-1}$.

This HP energy matrix involves only attractions (H with H) and neutral interactions (P with P, and P with H). Other folding codes include repulsions.[23–25,29,30,32,36] Some folding studies explore a restricted ensemble of only maximally compact states,[10,11,18,33] and some folding codes neglect correlations (see below) among different interresidue interactions[10–12,19,20,22] (see also ref. 27). We call these uncorrelated folding codes. We study the different properties of these various types of folding codes.

Repulsions

In addition to the HP model, we also study the "AB" model, which involves two monomer types A and B having the energy matrix

$$\mathscr{E} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array} \\ \begin{pmatrix} -1 & +1 \\ +1 & -1 \end{pmatrix} \end{array}. \qquad (2)$$

This energy matrix favors contacts between monomers of the same type and disfavors contacts between different monomer types. This folding code does not correspond to a hydrophobic/polar distinction. It leads to "left–right" separations of monomer types A and B, rather than inside/outside separations of hydrophobic and polar residues as in proteins.[18] This applies also to modified AB models with an attractive background interaction.[16,21]

In addition, we consider a variant of the HP model that includes repulsions. It is based on "shifting" the average interaction to zero. For a general energy matrix, the average interaction is defined as

$$\langle\epsilon\rangle \equiv \frac{2}{\mu(\mu+1)} \sum_{i\leq j}^{\mu} \epsilon_{ij} \qquad (3)$$

where μ is the number of monomer types; $\mu = 2$ for the two-letter sequences in Eqs. (1) and (2). The entries of a shifted energy matrix $\mathscr{E}'$ are given by

$$\epsilon'_{ij} \equiv \epsilon_{ij} - \langle\epsilon\rangle. \qquad (4)$$

Since the average interaction $\langle\epsilon'\rangle$ of the shifted energy matrix is zero, it must contain both favorable interactions (negative energies) and unfavorable (positive) ones. Shifted energy matrices have been used in protein folding models.[23–25,29,30,32] For both the HP and AB model described above, $\langle\epsilon\rangle = -1/3$. Therefore, after multiplying by a scaling factor of 3, the shifted HP energy matrix is

$$\mathscr{E}' = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{P} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{P} \end{array} \\ \begin{pmatrix} -2 & +1 \\ +1 & +1 \end{pmatrix} \end{array} \qquad (5)$$

whereas a uniform scaling by a factor of 3/2 results in the shifted AB matrix

$$\mathscr{E}' = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array} \\ \begin{pmatrix} -1 & +2 \\ +2 & -1 \end{pmatrix} \end{array}. \qquad (6)$$

Strong Attractions

One popular model folding code assumes that all monomers are strongly attracted to all others and that the chemical distinctions between monomers is only a small perturbation of this strong background attraction.[10,11,33] We call these "perturbed homopolymer" models. By definition, the native states of all sequences in these models are maximally compact.[37] To explore perturbed homopolymer folding codes, we add a strong background monomer/monomer attraction to both the HP and AB potentials, so that the contact energies are

$$\epsilon_{ij} - C, \quad C \rightarrow \infty. \qquad (7)$$

This energy matrix may also be regarded as shifted, but we reserve the term "shifted matrix" only for cases of $\langle\epsilon'\rangle = 0$. In the limit of $C \rightarrow \infty$, each contact is infinitely favorable so only maximally compact conformations can be native structures.

Correlations

We also explore the *correlation* in a folding code. Suppose an alphabet consists of monomer types X, Y and Z. A given sequence of X, Y and Z may have multiple possibilities of forming XY contacts, for example. If residue pair (1,4) and residue pair (35,90) both consist of monomer types X and Y, then it is clear that both pairs of contacts must have the same energy. *Correlation* is our term describing this physical requirement that all instances of an interaction between residue types i and j must have the same energy.[27] However, some simplified models neglect correlation.[10–12,19,20,22] In such models, the interaction between any pair of residue can have an energy independent of the interaction energy of all other residue pairs.

Correlation is absent in "Gō folding codes" with "strong limit specificity" because they allow independent variation of the contact energy between any residue pair.[38,39] Some recent spin-glass-inspired models also neglect correlation in folding codes.[10–12,19,20,22] Such models specify pairwise contact energies b_{ij}, which are intended as approximations to the quantities $\epsilon_{t_it_j}$ of the energy matrix, where t_i and t_j are the monomer types of monomer i and monomer j [t_i, $t_j \leq \mu$; see Eq. (3)]. The main assumption in these studies is that each b_{ij} can be modeled as an *independent* random variable, which is often assumed to obey a Gaussian distribution. Neglecting correlation is equivalent to having an infinitely large monomer alphabet.[20] In that case, for a random sequence of finite length, multiple instances of a particular XY pair would become vanishingly small; hence correlation is negligible. It follows that in general the importance of correlation should diminish with increasing alphabet size.

How serious an error is the neglect of correlations? For polymers constructed from a two-letter alphabet, X and Y, in long chains there will naturally be large numbers of residue pairs of identical type, since only the pairs XX, YY, and XY are possible. Hence correlation is bound to be important. Proteins use 20-letter alphabets; thus correlation is less important than in two-letter models but is nonetheless present for any polypeptide sequence with at least one amino acid type occurring more than once. This includes all polypeptides with chain lengths larger than 20. The degree of correlation does not depend solely on the alphabet size. Different amino acid pairs that share similar chemical properties (for example, hydrophobic-hydrophobic) have similar or related interactions, which would tend to increase correlation. It is likely that correlation is not negligible in proteins.

On two-dimensional square and three-dimensional simple cubic lattices (and all hypercubic lattices in higher spatial dimensions), the number of possible contacts between pairs of monomers along a sequence of length n is given by

$$m = \begin{cases} (n-2)^2/4 & \text{for } n \text{ even.} \\ (n-1)(n-3)/4 & \text{for } n \text{ odd.} \end{cases} \tag{8}$$

Therefore, if there are v possible contact energies for every contact, the total number of b_{ij} sequences is v^m. To make exhaustive enumeration of all sequences computationally feasible for at least some short chains, we consider b_{ij} sequences with minimal variability, i.e., with $v = 2$ possible contact energies. Two cases are considered: (1) $b_{ij} = -1, 0$ (either $b_{ij} = -1$ or $b_{ij} = 0$); and (2) $b_{ij} = \pm 1$. These may be regarded as the uncorrelated version of the HP [Eq. (1)] and AB [Eq. (2)] potentials, respectively. Native conformations of b_{ij} sequences are determined by exhaustive conformational searches. We study also the combined effects of no correlation with a strong background favorable intrachain interaction by limiting our search of native structures of b_{ij} sequences to the maximally compact ensemble. In these two cases, the set of 2^n sequences with correlated interactions, when represented as b_{ij} sequences, is always a subset of the 2^m sequences with uncorrelated interactions. Since $m > n$ for $n \geq 8$, when the chain length n is sufficiently long, $2^n << 2^m$, so there are many more uncorrelated sequences than correlated sequences. Thus, correlation is a severe constraint on the heterogeneity and diversity of intrachain interactions in heteropolymer sequences.

RESULTS

All These Folding Codes Have Low Median Degeneracy

All the folding codes we studied have low median sequence degeneracies (Fig. 1A,C; see also refs. 1,3,6,9,15,18). For example, half of all HP sequences of chain length $n = 18$ have no more than 47 ground-state conformations, and 1/8 have no more than 5 ground-state conformations, out of a conformational space of 5,808,335 conformations.[2] (By comparison, an all-H $n = 18$ homopolymer has 1,673 maximally compact ground-state conformations.[2,37]) Moreover, while the size of the conformational space grows exponentially with chain length ($\sim (n-1)^{1/3}(2.64)^n$),[2] the median degeneracy grows only approximately linearly ($\sim 3.48n + 14.1$ for $n \geq 10$). As noted before, the distributions of degeneracy g are sharply peaked at very low values, $g \sim 1$.[1,3,15,18] This applies to all the folding codes we studied, and to a modified HP model studied elsewhere containing an explicit repulsion between H monomers and solvent.[40]

Figure 1 shows that introducing unfavorable interactions into a folding code reduces the average

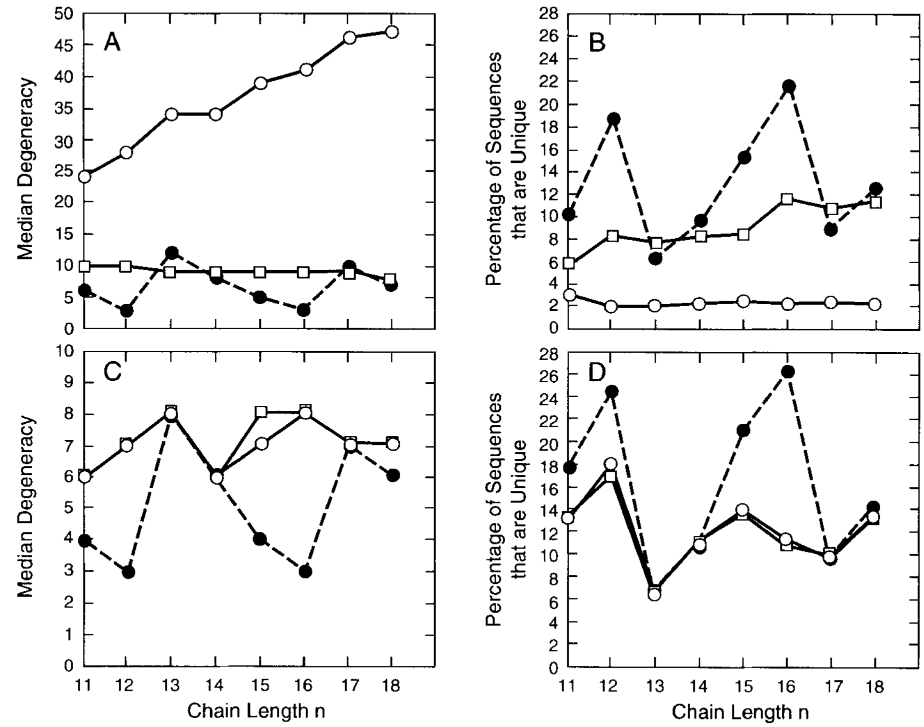

Fig. 1. **A and C**: Median degeneracy g vs. chain length n. For given n, one half of the 2^n sequences have ground-state degeneracies lower than the median g. Six different energy matrices are considered. A: Circles represent HP sequences; squares represent "shifted-HP" ($\langle\epsilon'\rangle = 0$) sequences; and dots connected by dashed lines represent "perturbed-homopolymer-HP" sequences with native structures restricted to maximally compact ($\rho = 1$) conformations. C: Circles represent AB sequences; squares represent "shifted-AB" ($\langle\epsilon'\rangle = 0$) sequences; and dots connected by dashed lines represent "perturbed-homopolymer-AB" sequences with native structures restricted to maximally compact ($\rho = 1$) conformations. **B and D**: Percentages of sequences that are unique for sequence types in A and C, respectively. The HP data in B are from refs. 3 and 6. $\rho = t/t_{\text{max}}$ where t_{max} is the maximum possible number of intrachain contacts for a given chain length.[37]

TABLE I. Number of Unique ($g = 1$) Uncorrelated (b_{ij}) Sequences of Lengths $n = 11, 12$*

	$b_{ij} = -1, 0$		$b_{ij} = -1, +1$	
n	$\rho = 1$	Exact	$\rho = 1$	Exact
11	248,988 (23.8)	198,686 (19.0)	248,988 (23.8)	207,934 (19.8)
12	11,553,705 (34.4)	7,164,974 (21.4)	11,553,705 (34.4)	7,758,122 (23.1)

*Percentages of sequences that are unique are in parentheses. For chains configured on the two-dimensional square lattice with only two different possible b_{ij}s, there are 2^{20} and 2^{25} uncorrelated sequences for $n = 11$ and 12, respectively. Exact results are obtained by exhaustive conformational searches. Results obtained by *restricted* conformational searches of only the maximally compact ensemble ($\rho = 1$) are also given. Note that the $\rho = 1$ statistic is independent of the magnitudes of b_{ij}s. In general it depends only on the number of possible values for b_{ij}.

degeneracy. Restricting native structures to the ensemble of maximally compact conformations ($\rho = 1$ where ρ is a measure of compactness[37] by including a strong background favorable intrachain interaction[10,11,16,19–22] also decreases the median degeneracy. We find that the average degeneracy does not increase with chain length when repulsions are included (Fig. 1A,C). In addition, Figure 1B and D shows that there are more "good" sequences (i.e., sequences that fold to unique native states) when the energy matrix involves repulsions. Hence introducing repulsive interactions into a monomer alphabet should help reduce the conformational diversity of the stable states.

Lower degeneracies can be achieved by introducing repulsions into the monomer alphabet. Lower degeneracies can also arise from neglecting or reducing correlations. This can be achieved by increasing the size of the alphabet, but the complete neglect of correlation is nonphysical for real polymers. Although 3 and 2% of HP sequences, and 13 and 18% of AB sequences of chain lengths $n = 11$ and 12 fold uniquely, neglecting correlations leads to 19 and 21% and 19 and 23% of unique folders. Further re-

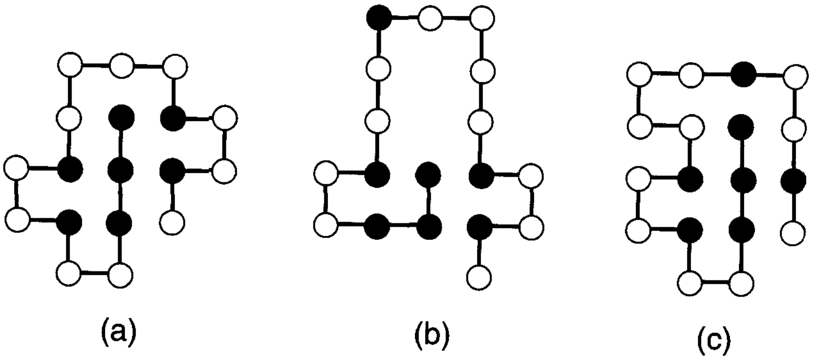

Fig. 2. Shifting the energy matrix can change the native (ground-state) structure. **a**: An HP sequence in its unique native structure. **b**: The sequence in (a) shifted by the $\langle\epsilon'\rangle = 0$ condition has a different unique native structure. **c**: The corresponding perturbed-homopolymer-HP sequence has yet another unique (maximally compact) native structure.

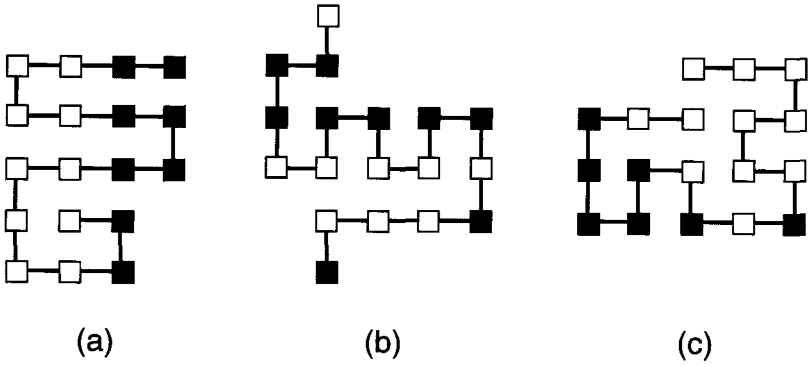

Fig. 3. Native structures of AB-type sequences. The two types of monomers tend to separate to opposite sides of the native conformations. Three different AB sequences are shown in their unique native structures. **a**: The corresponding shifted-AB ($\langle\epsilon'\rangle = 0$) and perturbed-homopolymer-AB sequences have the same unique native structure. **b**: The corresponding shifted-AB ($\langle\epsilon'\rangle = 0$) sequence has the same unique native structure, but the corresponding perturbed-homopolymer-AB sequence is not unique ($g = 2$). **c**: The corresponding perturbed-homopolymer-AB has the same unique native structure, but the corresponding shifted-AB ($\langle\epsilon'\rangle = 0$) sequence is not unique ($g = 83$).

striction to the maximally compact ensemble gives even larger percentage of unique folders—24% for $n = 11$ and 34% for $n = 12$ (Table I).

We reach two conclusions. First, introducing repulsions or increasing the size of the monomer alphabet should help reduce sequence degeneracy. These principles may be useful for designing foldable heteropolymers. Second, it is not true that any change in a model that favors more unique folding is necessarily more physical or protein-like. Some modeling efforts have chosen energy matrices based on their ability to cause a high fraction of arbitrarily chosen sequences to fold uniquely, but no experimental evidence indicates that a high percentage of random sequences of amino acids will fold uniquely. Therefore model folding codes with the lowest median sequence degeneracy are not necessarily the most protein-like.

In addition, we find that shifting energy matrices to arbitrary values of mean energy or neglecting correlation, which is commonly done to help reduce conformational diversity, changes the physics of the model.[10–12,19,20,22–25,30,32] For example, Figure 2 shows a given HP sequence that has different unique native structures for three different shifted energy matrices. Figure 3 shows three different AB sequences for which the ground-state degeneracies are changed by shifting their energy matrices. Thus, adding arbitrary constants to knowledge-based potentials, such as the Miyazawa and Jernigan[41] energy matrix, may change both the native states and their degeneracies that would have been predicted without them.

Encodabilities of Structures

We now consider how the nature of the energy matrix determines the encodabilities of compact structures in monomer sequences. To do this, we first define *potential encodability*. A chain fold is potentially encodable if it is the unique ground-state conformation of *some* energy matrix. The present models only consider contact interactions. Thus a

TABLE II. Number of Potentially Encodable Conformations Among Two-Dimensional Square Lattice Chains of n = 11–18 Monomers*

n	No. of potentially encodable conformations
11	154
12	519
13	898
14	2,836
15	4,954
16	15,048
17	26,494
18	77,635

*A potentially encodable conformation has a set of contacts (i.e., a contact map) that can only be realized by itself and not by any other conformation.

conformation is potentially encodable if its set of contacts (contact map) is satisfied only by itself but no other conformations (Table II). Otherwise the conformation can never be the unique native structure of any sequence in any energy matrix. Figure 4 shows that the fraction of structures that are potentially encodable increases with compactness. The fraction of ρ = 1 maximally compact conformations that are potentially encodable is very high.

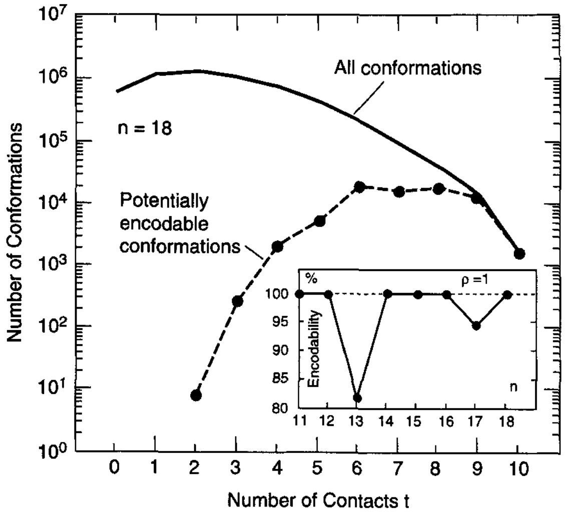

Fig. 4. Number of conformations (solid curve) as a function of number of intrachain contacts t for chain length n = 18. Dots connected by dashed lines give the number of *potentially encodable* conformations. There are no potentially encodable conformations for t < 2. The inset shows percentages of maximally compact (ρ = 1) conformations that are potentially encodable, for n = 11–18.

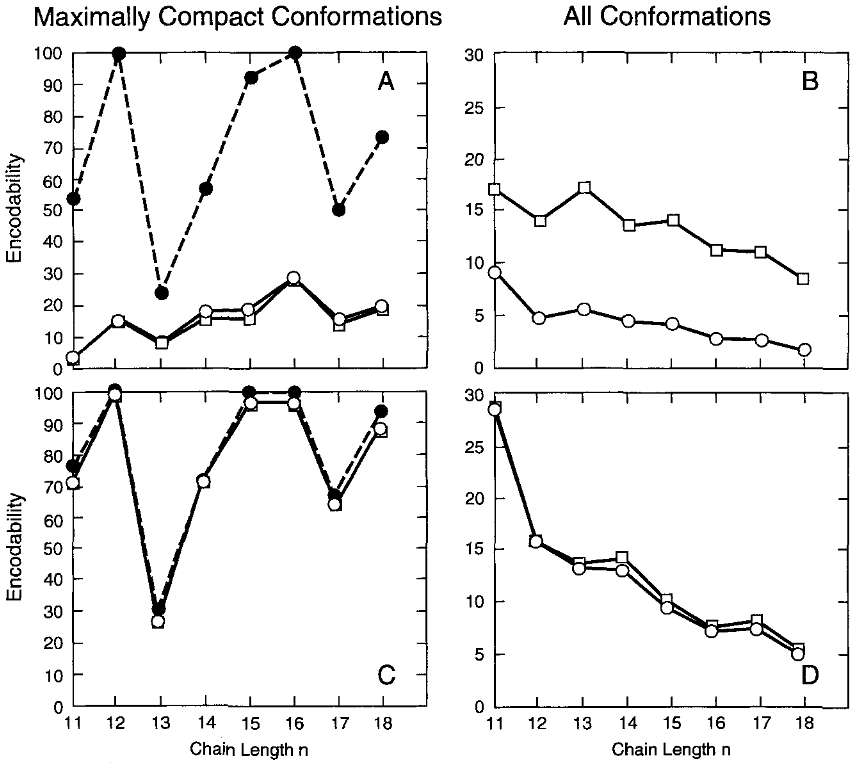

Fig. 5. **A and C:** Encodability of maximally compact (ρ = 1) conformations vs. chain length n. Circles, squares, and dots connected by dashed lines represent different "HP" sequence types in (A), and different "AB" sequence types in (B), as in Figure 1A and C. **B and D:** Total encodability vs. n. Circles and squares represent, respectively, the unshifted and $\langle \epsilon' \rangle$ = 0 HP (B) and AB (D) sequences. For any given energy matrix, encodability is the fraction of potentially encodable conformations in a given ensemble that are actually encodable by at least one sequence.

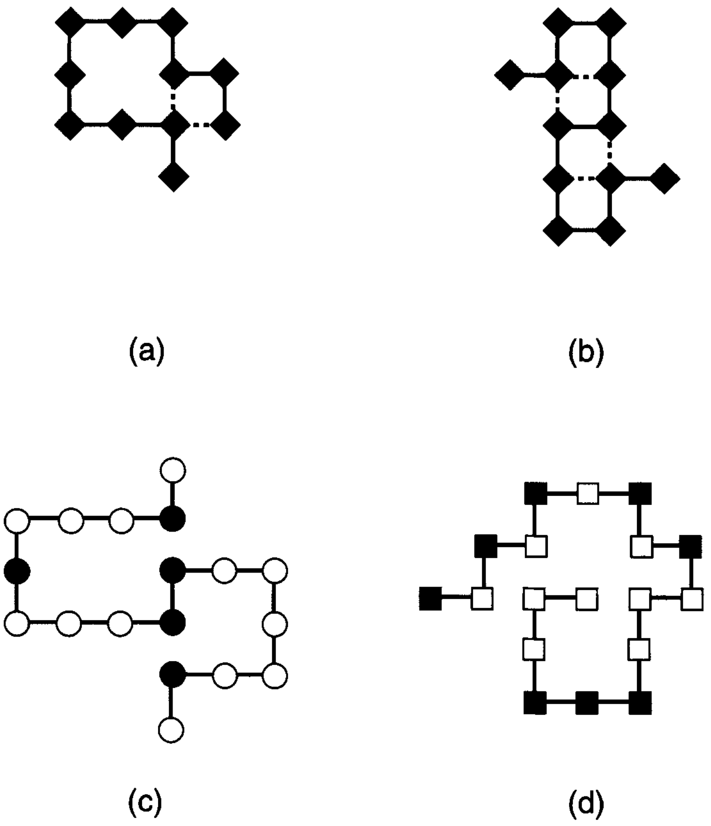

Fig. 6. Unfavorable or uncorrelated interactions allow encoding of certain open conformations. **a:** The open structure is encodable by 172 uncorrelated $b_{ij} = \pm 1$ sequences and 15 shifted HP ($\langle \epsilon' \rangle = 0$) sequences, but not by any other sequence type considered in this paper. **b:** The relatively compact structure is encodable by all sequence types considered except the HP sequences. Dots in a and b represent favorable interactions. **c:** The open structure is the unique ground state of the shifted HP ($\langle \epsilon' \rangle = 0$) sequence shown. The corresponding HP sequence has a ground-state degeneracy of 19. **d:** The open structure is the unique ground state of the shifted AB ($\langle \epsilon' \rangle = 0$) sequence shown. The corresponding AB sequence has a ground-state degeneracy of 3.

The number of chain folds that can actually be encoded in a sequence depends on the energy matrix. The upper bound of this number is the number of potentially encodable conformations. Figure 5 shows that only 4–29% of compact structures are encodable in the HP or the shifted HP potentials. There are many compact conformations that cannot be encoded using these simple alphabets.[3,4] Recent studies also indicate that not all three-dimensional compact structures are encodable in the HP potential.[5,7,8] It is not known what fraction of compact conformations are encodable in real proteins, so we cannot judge the HP model on this basis. If a conformation is maximally compact ($\rho = 1$), the odds are very good that it can be encoded in the AB and shifted AB potentials (Fig. 5C). This implies that *strong* repulsions can enhance structure encodability, although *weak* repulsions do not help much.

We found it interesting that adding repulsions to an energy matrix causes some semi-open conformations to be encodable (compare Fig. 5A and 5B). Figure 6 shows some examples of open conformations

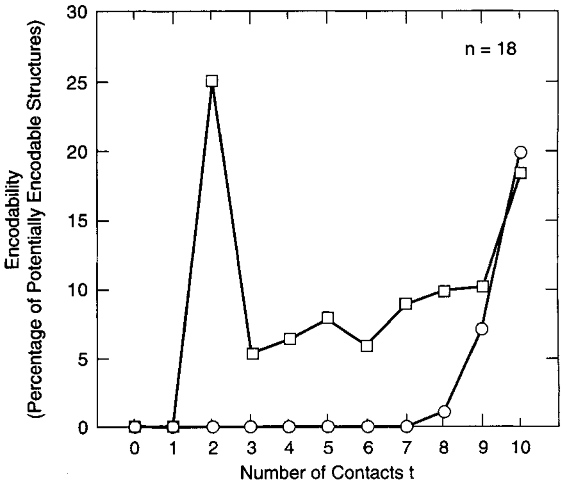

Fig. 7. Unfavorable interactions extend encodability to more open conformations. Circles: Encodability of $n = 18$ conformations by HP sequences. Only relatively compact conformations are encodable. No structures with fewer than $t = 8$ intrachain contacts can be encoded. Squares: Encodability by shifted ($\langle \epsilon' \rangle = 0$) HP sequences with unfavorable interactions. Some open structures with as few contacts as $t = 2$ are encodable.

TABLE III. Percentages of Structures That Are Encodable by Uncorrelated (b_{ij}) Sequences, Obtained by Exhaustive Searches Over All Conformations and *All Possible Sequences* of Lengths n = 11,12*

	b_{ij} = −1, 0		b_{ij} = −1, +1	
n	ρ = 1	Total	ρ = 1	Total
11	100	55.8	100	100
12	100	68.2	100	100

*"Total" encodability refers to the percentage of all potentially encodable conformations that are encodable by some sequence(s) with the given intrachain interactions. All (100%) maximally compact (ρ = 1) conformations with n = 11, 12 are potentially encodable (see Fig. 4).

that are unique ground states for energy matrices that are uncorrelated and/or having repulsions. Figure 7 shows that a considerable number of open conformations become encodable when energy matrices include repulsions. Indeed, proteins may make use of this, since they have charged residues, which can be repulsive. On the other hand, proteins may not encode open or loop regions with repulsions since electrostatic interactions are weak in water. One caveat is that our study does not consider long-range electrostatic repulsions; the repulsions we model here are only short-range disaffinities. Even if such open conformations are not encoded in proteins in this way, this principle may be useful in the design of other classes of polymers.

Figure 5B, D also shows that the fraction of encodable conformations decreases with increasing chain length. Moreover, Figure 5A and C shows that, whereas shifting the energy matrix affects degeneracies, it does not much affect encodabilities of compact structures. In fact, the set of encodable ρ = 1 conformations is almost identical for the HP and shifted HP potentials and is exactly identical for the AB and shifted AB potentials.

As with degeneracies, Table III shows that neglecting correlations also enhances encodabilities. Govindarajan and Goldstein[33] have also observed this before. As we noted before with degeneracies, whereas models that assume that all native states are maximally compact (or that neglect correlations) can give high encodabilities, this does not imply they are more physical.

These aspects of encodabilities are simple to understand. Good sequences involve "negative design," the designing *out* of incorrect chain folds, as well as "positive design," the designing *in* of the desired native structure.[4] Repulsions are more effective for negative design than neutral interactions are, because they raise the energy more of the competing conformations. Also, neglecting correlations helps negative design. With a small monomer alphabet, changing monomer types X and Y will have energy implications at multiple sites in the native and the non-native conformations of the molecule. On the other hand, if each pair of monomer contacts is independent of each other, then there is more freedom to tune individual contacts more specifically without affecting other parts of the structure or other conformations. Hence more structures are encodable by uncorrelated b_{ij} sequences than by their correlated counterparts. Table III shows that structure encodability is highest when unfavorable interactions are present *and* interactions are uncorrelated.

CONCLUSIONS

We have used simple lattice modeling to study different possible energy matrices used in folding codes that translate monomer sequences into folded structures. The simplest folding code involves the HP binary alphabet. The HP folding code leads to a sequence space having many sequences that do not fold to unique native structures and it leads to an inability to encode some structures. Adding repulsive interactions leads to folding codes where more sequences fold to unique conformations and more compact structures can be encoded within sequences.

We also study nonphysical aspects of folding codes that have been used in some models, namely: (1) the neglect of correlations among the interactions; (2) the restrictive assumption that all native states are so highly compact to be perfectly cubical in lattice models; and (3) the addition of energy constants to "shift" energy matrices. There is no evidence supporting the suggestion that shifting corresponds to experimentally feasible changes in solution conditions.[29] These assumptions have mainly been for convenience or tractability. Earlier studies by Gō and co-workers[38,39] showed that sequences with less correlation are more stable and fold faster.[23–25,29,30,32,36] Repulsive interactions prevent trapping in local energy minima,[29,36] which are the bottlenecks in many lattice folding models.[6,12,26,42] Folding funnels can also be more "focused" by introducing unfavorable interactions (N. D. Socci, personal communication; see also refs. 29,36).

Nevertheless, while these modeling tricks make model proteins fold more uniquely, or faster, and allow a greater breadth of chain folds, they also change the native states of given sequences and can be poorer physical models. For example, in some cases, adding energy constants[23] to Miyazawa and Jernigan[41] matrices leads to burial of charged residues.* Although it is possible to constrain the design

*The energy matrix used in ref. 23 is modified from Table VI of Miyazawa and Jernigan,[41] which gives "the preference for the specific contact pair *i-j* over the average contacts of the *i*th and *j*th types of residues." All entries in the matrix used in ref. 23 are identical to this particular matrix of Miyazawa and Jernigan except for the following pairs: tyrosine-tyrosine, glycine-phenylalanine, and glutamine-phenylalanine. Reference 41 gives 0.06, −0.38, and −0.29 for these entries, but ref. 23 gives −0.06, 0.38, and 0.49 (E.I. Shakhnovich, personal communication).

to put some hydrophobic residues into the protein core,[32] many native structures[24,25,29,30,32] of design sequences with energy matrices similar to that in ref. 23 or Table VI of ref. 41 have monomer distributions quite different from those of real proteins, with many polar residues in the core.

What energy matrices most accurately represent real protein folding codes? It is not clear. The true folding code for proteins has 20 letters, and clearly considerations beyond pairwise contact energies are ultimately necessary. Whereas it is a challenge to design sequences in the HP model that fold uniquely, stably, and quickly, in 20-letter models studied in refs. 23–25,29,30, and 32 having many more parameters, sequence design is relatively simple,[23,30] and the rate of folding is rapid.[23–25,29,30,32] This alone, however, does not imply that any model 20-letter or other multi-letter code is a better model of proteins than a code with fewer letters. The alphabet size is only one factor affecting how well a model code mimics the true code. Some codes assume independent variations of energies b_{ij}, which neglects physical correlations.[10–12,19,20,22] Some 20-letter codes shift their energy matrices, which can change the physics, change the native states, and bury charge, which is nonphysical.[23–25,29,30,32] Our study shows that when we fix the number of letters in the code alphabet to two, other important physical factors in the energy matrix than just the alphabet size determine native conformations, degeneracies, and encodabilities. Because the true degeneracies or encodabilities of real proteins are not known, it is not yet clear which energy matrix models are most protein-like.

The present work also aims to give some guidance in choosing monomer sets for design of other foldable polymers.[43,44] In particular, there may be some conformations, particularly involving open chain structures, that might be encodable in other polymers but are not currently found in proteins.

NOTE ADDED IN PROOF

Recently, a rigorous formulation of the sequence design problem has been provided.[45,46] A discussion of the validity of knowledge-based "energy potentials" for proteins has been given in ref. 47.

ACKNOWLEDGMENTS

We thank Joe Bryngelson, Josh Deutsch, Alexander Gutin, Henri Orland, José Onuchic, Eugene Shakhnovich, Nick Socci, Dave Thirumalai, Paul Thomas, Peter Wolynes, and Kaizhi Yue for helpful discussions, and the National Institutes of Health for financial support.

REFERENCES

1. Lau, K.F., Dill, K.A. A lattice statistical mechanical model of the conformational and sequence spaces of proteins. Macromolecules 22:3986–3997, 1989.
2. Chan, H.S., Dill, K.A. Polymer principles in protein structure and stability. Annu. Rev. Biophys. Biophys. Chem. 20:447–490, 1991.
3. Chan, H.S., Dill, K.A. "Sequence space soup" of proteins and copolymers. J. Chem. Phys. 95:3775–3787, 1991.
4. Yue, K., Dill, K.A. Inverse protein folding problem: Designing polymer sequences. Proc. Natl. Acad. Sci. USA 89: 4163–4167, 1992.
5. Yue, K., Dill, K.A. Sequence structure relationship of proteins and copolymers. Phys. Rev. E 48:2267–2278, 1993.
6. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100:9238–9257, 1994.
7. Yue, K., Dill, K.A. Forces of tertiary structural organization of globular proteins. Proc. Natl. Acad. Sci. USA 92: 146–150, 1995.
8. Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S., Shakhnovich, E.I., Dill, K.A. A test of lattice protein folding algorithms. Proc. Natl. Acad. Sci. USA 92:325–329, 1995.
9. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S. Principles of protein folding—a perspective from simple exact models. Protein Sci. 4:561–602, 1995.
10. Shakhnovich, E.I., Gutin, A.M. Enumeration of all compact conformations of copolymers with random sequence of links. J. Chem. Phys. 93:5967–5971, 1990.
11. Shakhnovich, E.I., Gutin, A.M. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346:773–775, 1990.
12. Shakhnovich, E., Farztdinov, G., Gutin, A.M., Karplus, M. Protein folding bottlenecks: A lattice Monte Carlo simulation. Phys. Rev. Lett. 67:1665–1668, 1991.
13. Leopold, P.E., Montal, M., Onuchic, J.N. Protein folding funnels: A kinetic approach to the sequence-structure relationship. Proc. Natl. Acad. Sci. USA 89:8721–8725, 1992.
14. O'Toole, E.M., Panagiotopoulos, A.Z. Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. J. Chem. Phys. 97:8644–8652, 1992.
15. Camacho, C.J., Thirumalai, D. Minimum energy compact structures of random sequences of heteropolymers. Phys. Rev. Lett. 71:2505–2508, 1993.
16. Shakhnovich, E.I., Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. USA 90:7195–7199, 1993.
17. Shakhnovich, E.I., Gutin, A.M. A new approach to the design of stable proteins. Protein Eng. 6:793–800, 1993.
18. Gutin, A.M., Shakhnovich, E.I. Ground state of random copolymers and the discrete random energy model. J. Chem. Phys. 98:8174–8177, 1993.
19. Šali, A., Shakhnovich, E., Karplus, M. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614–1636, 1994.
20. Šali, A., Shakhnovich, E., Karplus, M. How does a protein fold? Nature 369:248–251, 1994.
21. Socci, N.D., Onuchic, J.N. Folding kinetics of protein-like heteropolymers. J. Chem. Phys. 101:1519–1528.
22. Karplus, M., Šali, A. Theoretical studies of protein folding and unfolding. Curr. Opin. Struct. Biol. 5:58–73, 1995.
23. Shakhnovich, E.I. Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. 72: 3907–3910, 1994.
24. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Free energy landscape of protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. J. Chem. Phys. 101:6052–6062, 1994.
25. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. Biochemistry 33:10026–10036, 1994.
26. Thirumalai, D. Theoretical perspectives on in vitro and in vivo protein folding. In: "Statistical Mechanics, Protein Structure, and Protein-Substrate Interactions." Doniach, S. (ed.). New York: Plenum, 1994:115–134.
27. Chan, H.S. Kinetics of protein folding. Nature 373:664–665, 1995.
28. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.

29. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I. Is burst hydrophobic collapse necessary for protein folding? Biochemistry 34:3066–3076, 1995.
30. *Gutin*, A.M., Abkevich, V.I., Shakhnovich, E.I. Evolution-like selection of fast-folding model proteins. Proc. Natl. Acad. Sci. USA 92:1282–1286, 1995.
31. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z., Socci, N.D. Toward an outline of the topography of a realistic protein-folding funnel. Proc. Natl. Acad. Sci. USA 92: 3626–3630, 1995.
32. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Domains in folding of model proteins. Protein Sci. 4:1167–1177, 1995.
33. Govindarajan, S., Goldstein, R.A. Searching for foldable protein structures using optimized energy functions. Biopolymers 36:43–51, 1995.
34. Govindarajan, S., Goldstein, R.A. Optimal local propensities for model proteins. Proteins 22:413–418, 1995.
35. Betancourt, M.R., Onuchic, J.N. Kinetics of proteinlike models: The energy landscape factors that determines folding. J. Chem. Phys. 103:773–787, 1995.
36. Socci, N.D., Onuchic, J.N. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. J. Chem. Phys. 103:4732–4744, 1995.
37. Chan, H.S., Dill, K.A. Compact polymers. Macromolecules 22:4559–4573, 1989.
38. Taketomi, H., Ueda, Y., Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. Int. J. Pept. Protein Res. 7:445–459, 1975.
39. Gō, N., Taketomi, H. Respective role of short- and long-range interactions in protein folding. Proc. Natl. Acad. Sci. USA 75:559–563, 1978.
40. Sun, S., Brem, R., Chan, H.S., Dill, K.A. Designing amino acid sequences to fold with good hydrophobic cores. Protein Eng. 8:1205–1212, 1995.
41. Miyazawa, S., Jernigan, R.L. Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
42. Camacho, C.J., Thirumalai, D. Kinetics and thermodynamics of folding in model proteins. Proc. Natl. Acad. Sci. USA 90:6369–6372, 1993.
43. Simon, R.J., Kania, R.S., Zuckermann, R.N., Huebner, V.D., Jewell, D.A., Banville, S., Ng, S., Wang, L., Rosenberg, S., Marlowe, C.K., Spellmeyer, D.C., Tan, R., Frankel, A.D., Santi, D.V., Cohen, F.E., Bartlett, P.A. Peptoids: A modular approach to drug discovery. Proc. Natl. Acad. Sci. USA 89:9367–9371, 1992.
44. Cho, C.Y., Moran, E.J., Cherry, S.R., Stephans, J.C., Fodor, S.P., Adams, C.L., Sundaram, A., Jacobs, J.W., Schultz, P.G. An unnatural biopolymer. Science 261:1303–1305, 1993.
45. Kurosky, T., Deutsch, J.M. Design of copolymeric materials. J. Phys. A 28:L387–393, 1995.
46. Deutsch, J.M., Kurosky, T. New algorithm for protein design. Phys. Rev. Lett. 76:323–326, 1996.
47. Thomas, P.D., Dill, K.A. Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. In press.