Use of Quantitative Structure-Property Relationships to Predict the Folding Ability of Model Proteins

Aaron R. Dinner,[1,2] **Sung-Sau So,**[1] **and Martin Karplus**[1,2,3]*
[1]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts*
[2]*Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts*
[3]*Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, Strasbourg, France*

ABSTRACT **We investigate the folding of a 125-bead heteropolymer model for proteins subject to Monte Carlo dynamics on a simple cubic lattice. Detailed study of a few sequences revealed a folding mechanism consisting of a rapid collapse followed by a slow search for a stable core that served as the transition state for folding to a near-native intermediate. Rearrangement from the intermediate to the native state slowed folding further because it required breaking native-like local structure between surface monomers so that those residues could condense onto the core. We demonstrate here the generality of this mechanism by a statistical analysis of a 200 sequence database using a method that employs a genetic algorithm to pick the sequence attributes that are most important for folding and an artificial neural network to derive the corresponding functional dependence of folding ability on the chosen sequence attributes [quantitative structure-property relationships (QSPRs)]. QSPRs that use three sequence attributes yielded substantially more accurate predictions than those that use only one. The results suggest that efficient search for the core is dependent on both the native state's overall stability and its amount of kinetically accessible, cooperative structure, whereas rearrangement from the intermediate is facilitated by destabilization of contacts between surface monomers. Implications for folding and design are discussed. Proteins 33:177–203, 1998.** © **1998 Wiley-Liss, Inc.**

Key words: lattice model; Monte Carlo; protein folding; QSPR; genetic algorithm; neural network

INTRODUCTION

An important problem in structural biology concerns the properties of a polypeptide that allow it to have a unique, stable three-dimensional native state and to fold to that state from a denatured (coil) state in a reasonable time.[1] This problem can be approached by generating a large number of random sequences, testing each for its ability to fold, and then comparing the sequences to determine the essential differences. Such combinatorial strategies have been used in experiments to show that a wide range of sequences can lead to ordered molten globules (for reviews, see refs. 2 and 3). These experiments are analogous to computational studies with protein models in which the folding abilities of a large number of sequences with known ground states are compared.[4,5] Although the computer "experiments" can screen only hundreds of sequences, rather than millions, they have the advantage that they can detect nonfolding sequences (unlike the experimental selection techniques that so far have relied on resistance to proteolysis) so that it is possible to compare sequences that fold to the native structure with those that do not.

The first computational study that compared folding and nonfolding sequences employed a 27-mer random heteropolymer model of a protein subject to Monte Carlo (MC) dynamics on a simple cubic lattice.[4,6] Two hundred such sequences with unique ground states were generated, and each was tested for its ability to fold in 10 independent MC trials. It was found that a necessary and sufficient condition for folding is that the native state be a pronounced global energy minimum with a large energy gap between the ground state and the remainder of the states. The energy gap criterion has been confirmed in similar studies by other authors.[7,8] Moreover, it was shown that the gap can be simplified to a consideration of the difference in energy between the ground state and the first fully compact ($3 \times 3 \times 3$) excited state.[4] There is no significant correlation between the rate of folding and any other sequence attributes, such as the amount of secondary structure in the native state. The correlation between folding and the energy gap is explained by a nonspecific folding mechanism. Folding proceeds by a fast ($\sim 10^4$ MC steps) collapse to a semicompact random globule, followed by a slow ($\sim 10^7$ MC steps), nondirected search through the ($\sim 10^{10}$) semicompact

Grant sponsor: National Science Foundation.

*Correspondence to: Martin Karplus, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge, MA 02138. E-mail: marci@tammy.harvard.edu

Received 20 February 1998; Accepted 1 June 1998

structures for one of the ($\sim 10^3$) transition states that lead rapidly ($\sim 10^5$ MC steps) to the native conformation.[6] The restriction of the search to the compact portion of the conformation space and the large number of transition states lead to the resolution of the folding problem for the 27-mer. The energy gap criterion results in a native state that is stable at a temperature high enough for the folding polypeptide chain to overcome barriers between random semicompact states.

The 27-mer folding mechanism appears to be limited to small proteins.[6] Since the folding time is dominated by the random search, it scales with the ratio of compact configurations to transition states. This ratio is expected to increase exponentially with chain length and yield unrealistically long folding times for chains of more than about 80 residues.[4] Although it has been suggested that a 27-mer corresponds to a helical protein of about 60 residues,[9] which would increase somewhat the upper length limit of folding by the 27-mer mechanism, the arguments for this are neither very compelling nor sufficient to explain how larger domains (particularly those with a nonhelical structure) could fold by the 27-mer mechanism. Since experimentally the folding time appears not to be correlated with the size of a protein, other factors must be important in restricting the search time for the native state of longer chains. To examine this question, a corresponding comparative approach was employed for 125-mers.[5] This length was chosen because its maximally compact form is a $5 \times 5 \times 5$ cube with a surface-area-to-volume ratio that corresponds roughly to that found in the native states of globular proteins. To obtain folding in a reasonable time, sequences were engineered with pronounced global energy minima and high secondary structure content in the native state. The introduction of nonrandom properties to facilitate folding does not bias the results, since, as in the 27-mer study,[4] the conclusions are based only on comparisons of similarly engineered sequences with varying degrees of folding ability. The comparative approach differentiates the present study and the related ones[4–8,10–12] from most experiments and simulations that consider only sequences that fold (but see refs. 2, 3, and 13).

The detailed study of a few 125-mer sequences showed that they fold by an elaboration of the three-state mechanism elucidated for 27-mers.[4–6] The chain quickly collapses to a disordered globule. Then it makes a relatively slow search through the compact states for a specific set of about 30 contacts (nonbonded spatial nearest neighbors), which make up a core that serves as the transition state for folding to a near-native intermediate. Completion of folding requires rearrangement and condensation of surface residues, which is often slow due to the need to disrupt native structure in the intermediate. For a database of 200 sequences, folding is correlated with the overall stability of the native state and with the amount of its kinetically accessible cooperative secondary structure, in particular, antiparallel sheets connected by tight turns.

In the present paper, we make an in-depth analysis of the 125-mer results to determine the sequence attributes that are essential for folding. Because of the complexity of the phenomenon (there are many sequence properties that could be involved in determining folding ability, and each one may have only a weak correlation), we apply a method that has been used to derive quantitative structure-property relationships (QSPRs) for ligand binding.[14,15] This method employs a genetic algorithm to pick the sequence attributes that are most important for folding and an artificial neural network to derive the functional dependence of folding ability on the chosen sequence attributes. The predictive power of the QSPR is assessed both by cross-validation and application of the QSPR to the prediction of the folding ability of a set of 20 additional sequences that were not used in developing the QSPR.

The utility of the present analysis is demonstrated by the substantial increase in predictivity obtained when multiple sequence attributes are used at the same time. The Pearson linear correlation coefficients of the cross-validated predictions exceed 0.8. The genetic algorithm repeatedly selects measures related to three features of the sequences: the overall stability of the native state, the amount of kinetically accessible cooperative structure in the native state, and the stability of the surface contacts in the native state. The functional dependence of the neural network on the inputs demonstrates that folding is accelerated by increases in the native state stability and the amount of kinetically accessible cooperative structure in the native state and by a decrease in the stability of the contacts between surface monomers. This finding is in agreement with the earlier qualitative analysis of the folding mechanism.[5] Implications of the results for protein folding and design, as well as for the use of QSPRs to predict both model and real protein behavior, are discussed.

METHODS

The protein model has been described in some detail in earlier publications[4–6] but is briefly reviewed here. Subsequently, we describe the QSPR methodology and the sequence properties chosen for analysis.

Model

The energy function for the 125-mer self-avoiding heteropolymer chain on a simple cubic lattice consists of a sum over all contacts (nonbonded spatial nearest neighbors). It has the form:

$$E = \sum_{i<j} \Delta(\vec{r}_i - \vec{r}_j)(B_{ij} + B_0). \tag{1}$$

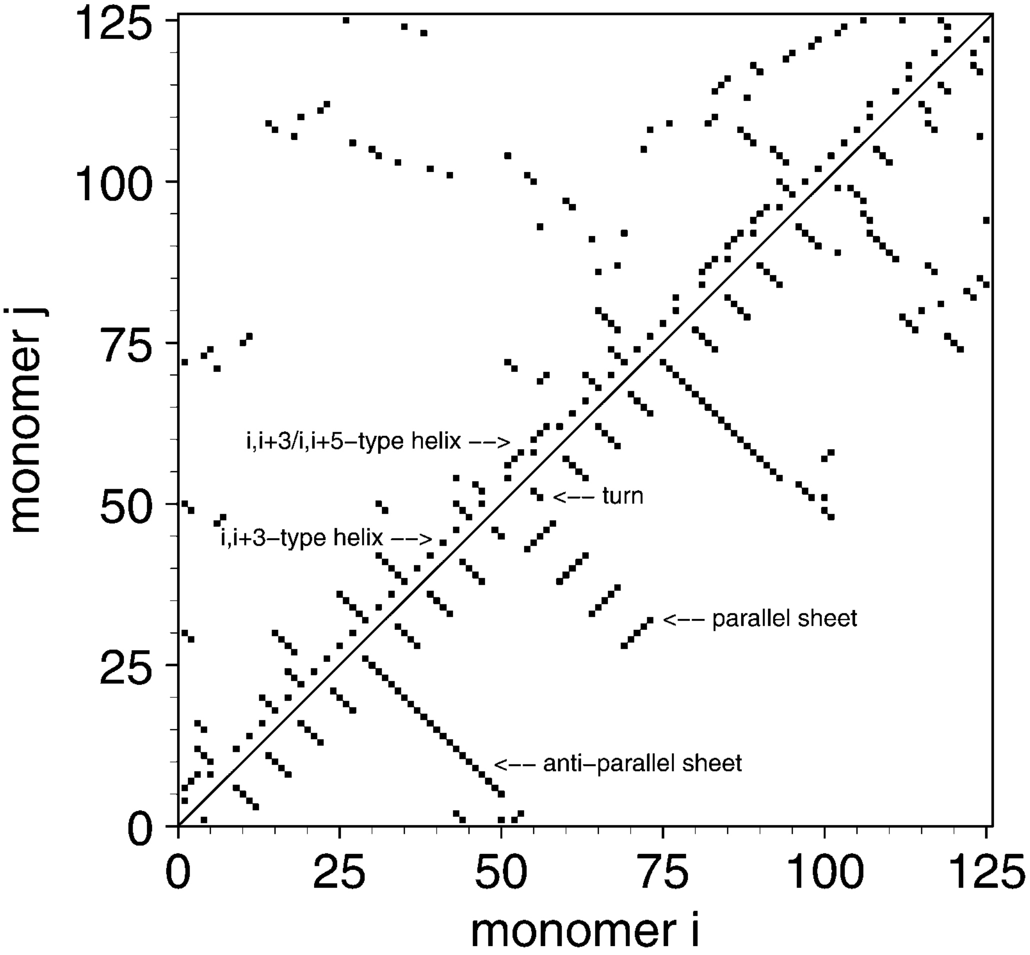

Fig. 1. Examples of secondary structure defined by patterns on a contact map. The contact maps are for the helix (upper half) and sheet (lower half) sequences shown in Fig. 2a and b, respectively.

The function $\Delta(\vec{r}_i - \vec{r}_j)$ selects the interacting monomer pairs; it is unity if i and j, located at $\vec{r}_i$ and $\vec{r}_j$, respectively, are in contact and zero otherwise. The B_{ij} give the specific interaction energies between monomers i and j, and a complete set of B_{ij} defines a sequence. The quantity B_0 ($B_0 < 0$) is a mean attraction between monomers that corresponds to an overall hydrophobic term. The choice of B_{ij} , B_0, and other parameters is discussed below.

Sequence Optimization

A comparative approach requires that the sequences that make up the database have a range of folding abilities. The probability of finding random sequences that satisfy both the thermodynamic and kinetic requirements for folding is very small for a 125-mer. Unlike the 27-mer case, in which 56 of 200 randomly chosen sequences were found to fold,[4] none that folded was found in 10 125-mer sequences. If the fraction of sequences satisfying the thermodynamic requirement is independent of chain length, as proposed by Shakhnovich and Gutin,[16] this suggests that the stability criterion is not sufficient for satisfying the kinetic requirement for folding in the 125-mer. Consequently, to obtain a 125-mer database with a significant fraction of folding sequences, nonrandom properties had to be introduced.

In the earlier study of the 125-mer,[5] we were interested in explaining the effect of the presence of secondary structure in the native state on folding ability. For lattices, the most direct measure of secondary structure is based on a contact map[17] (Fig. 1). Helices are composed of i, $i + 3$ and i, $i + 5$ contacts; thus, they line the diagonal of the contact map. Antiparallel sheets are such that if i and j are in contact, either $i - 1$ and $j + 1$ are in contact or $i + 1$ and $j - 1$ are in contact; these contacts form lines on the contact map perpendicular to the diagonal. Parallel sheets have either $i + 1$ and $j + 1$ in contact or $i - 1$ and $j - 1$ in contact, given a contact between i and j; these contacts form lines on the contact map parallel to the diagonal. Turns are an i, $i + 3$ contact with an $i - 1$, $i + 4$ contact. Any element of secondary structure must have at least two contacts; turns, by definition, have exactly two. A given contact may belong to more than one type of secondary structure since the definitions are not exclusive; for example, the shortest contacts of an antiparallel hairpin loop count as both antiparallel sheet and turn.

From an ensemble of fully compact 5 × 5 × 5 (176-contact) conformations, we selected 100 otherwise random structures with a large fraction of contacts in sheets and 100 otherwise random structures with a large fraction of contacts in helices (Figs. 2 and 7).[18] Each of the 200 chosen structures served as the starting point for a design process that created a sequence for which that structure is the native (ground) state. The resulting sequences can be regarded as representing, respectively, the β-sheet and α-helical classes of proteins.[19]

To generate a sequence given a native structure, we first create a matrix ($\mathbf{B}_\epsilon$) consisting of two types of entries: ϵ for monomer pairs in contact in the native state and 0 for all other pairwise interactions,[20] and then we add to that matrix N_B matrices ($\mathbf{B}_k$), which are analogous to the random matrices used in the 27-mer studies[4,6]:

$$\mathbf{B} = \mathbf{B}_\varepsilon + \sum_{k=1}^{N_B} \mathbf{B}_k. \tag{2}$$

The entries of each of the added matrices (B_{ij}^k where k is an index between 1 and N_B) are chosen at random from a Gaussian probability distribution:

$$P(B_{ij}^k) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\left(\frac{B_{ij}^k}{\sqrt{2}\sigma}\right)^2} \tag{3}$$

where σ is the standard deviation of the elements of each of the added matrices ($\mathbf{B}_k$). For all the sequences discussed, ϵ is –1, N_B is 5, and σ is 0.25. These values were chosen by trial and error to minimize the likelihood of changing the native state while maximizing the standard deviation of the resulting B_{ij} distribution (σ_B). The final standard deviation (σ_B) is not equal to the standard deviations of the individual added matrices (σ) and is typically about 0.6. We wanted to maximize σ_B to minimize the correlation between the native structure and the complete **B** matrix, from which the B_{ij} are obtained; this increases the similarity to sequences generated in a purely random way and allows better comparison with the 27-mer results. The sequence generation procedure is equivalent to drawing the B_{ij} from two Gaussian distributions with standard deviations equal to σ_B: one for native B_{ij} with a mean of -1 and one for non-native B_{ij} with a mean of 0. However, since there are far more non-native contacts than native ones (3,606 compared with 176), the overall distribution of the B_{ij} used for the computations are essentially Gaussian in character and centered on 0 (Fig. 3). Addition of B_0 to all contacts, as in Equation 1, translates the distribution as a whole.

The optimization procedure is similar to others that have been used in the literature[21,22] in that it lowers the energy of the native state relative to the majority of states in the spectrum and tends to increase the correlation between the energy of a state and its similarity to the chosen native state. It is expected to increase the stability of the native state, smooth the overall energy landscape, and accelerate the rate of folding. However, the shift in the distribution of folding abilities does not affect the analysis, since we base it on comparisons of sequences with varying folding abilities generated in the same manner. This comparative approach, which is similar to that used by Šali et al.,[4,6] contrasts with most experimental studies and other folding simulations in which only native "proteins" (sequences that repeatedly fold) are considered.

Monte Carlo

The 200 sequences were each subjected to 10 independent Metropolis Monte Carlo trials,[23] each of which started with a different random configuration. Trials continued for a maximum of 50 × 10^6 MC steps and were terminated earlier if the chain found the native state (first passage time). A single step consists of picking a random monomer, making a random change to the chain configuration (move) at that site, and then applying the Metropolis criterion. If the move could not be applied due to violation of either the excluded volume requirement or the connectivity of the chain, a new site and move were selected without counting the failed attempt as a step. The allowed chain moves are tail flips, internal bead flips, crankshaft rotations,[24] and a composite move in which two to four single monomer or crankshaft moves are performed before applying the Metropolis criterion. We do not include the "pivot" move,[25] in which one changes a single dihedral angle and thereby creates a displacement of a large segment of the chain; this move decreased the speed of finding the native state in preliminary trials with shorter chains. During MC simulations, single, crankshaft, and composite moves are chosen at random with specific weights (see next section).

On the average, a Monte Carlo folding simulation of 1 × 10^6 steps takes about 2 CPU minutes on a Hewlett Packard Apollo 735/125. However, this time is not distributed evenly, since the chain spends much more time trying to find a geometrically feasible move when it is in a dense conformation; 1 × 10^6 steps starting from a random, relatively open, configuration take about 1.5 CPU minutes, whereas the same number of steps starting from a 5 × 5 × 5 structure can take over 6 CPU minutes.

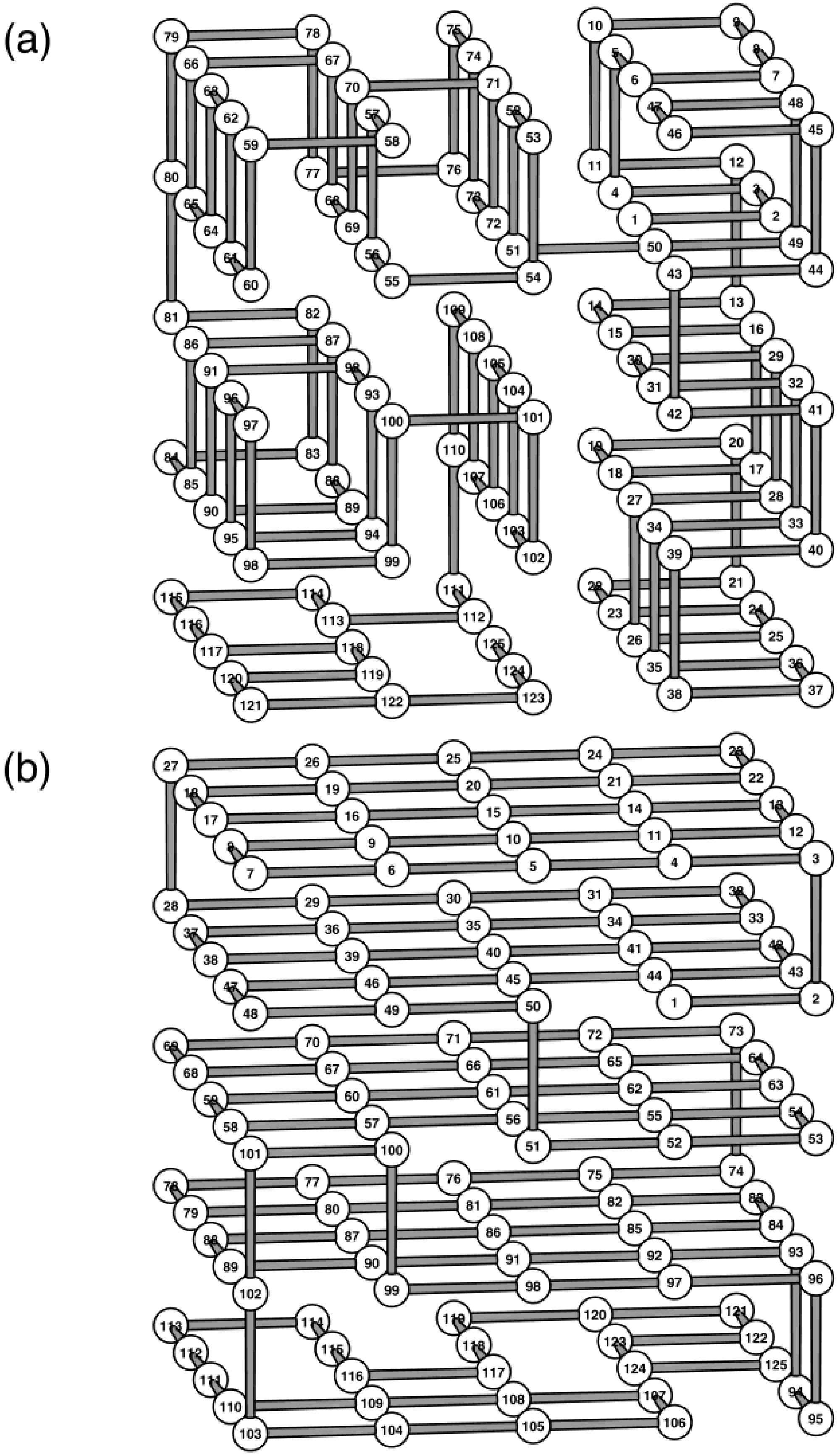

Fig. 2. Typical (a) helix and (b) sheet 5 × 5 × 5 native structures.

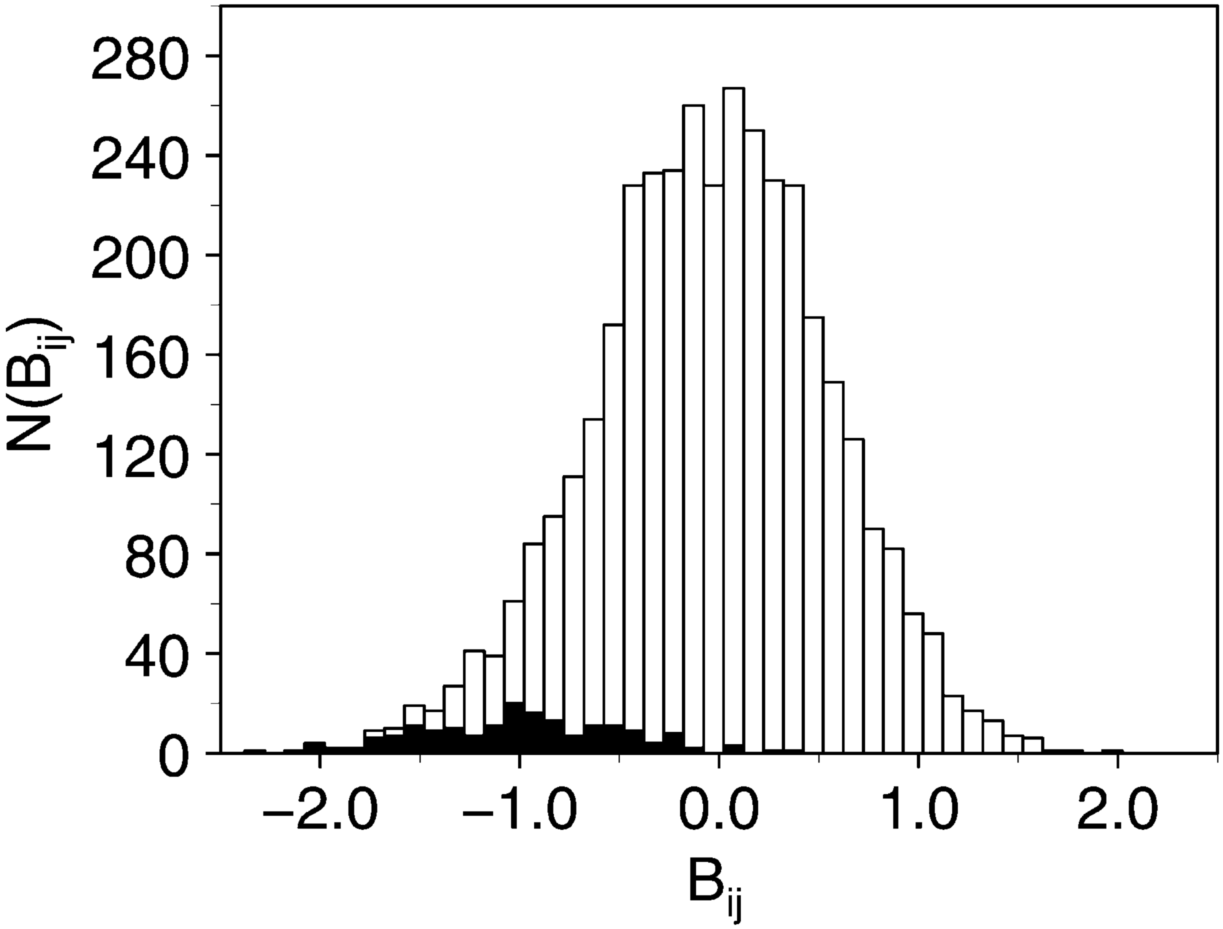

Fig. 3. Distribution of native (black) and non-native (white) B_{ij} for a typical 125-mer sequence.

Folding Parameters

To test a large number of sequences for their ability to fold to the native state in the most effective manner, it is necessary to optimize certain folding parameters; these include B_0, T, and the frequency of picking each type of MC move (Fig. 4). The optimal fraction of attempted crankshaft moves was taken from the 27-mer studies[4]; all other parameters were tested by exploring a limited range of conditions for a small number of sequences. The resulting optimized move parameters are attempt fractions of 0.8, 0.2, and 0.4 for crankshaft (f_{cr}), single monomer (f_s), and composite (f_{cm}) moves, respectively. These percentages do not add up to 1 because the first two fractions are the total breakdown of single monomer and crankshaft moves within both composite (up to four single monomer and crankshaft moves) and noncomposite moves; in other words, the fractions overlap because one first chooses whether or not to make a composite move and only subsequently chooses between crankshaft and single monomer moves. The actual distribution of moves is not necessarily the same as that of the assigned fractions because the type of move picked for a monomer may not be possible given the configuration of the chain. With the chosen attempt fractions, there are roughly 4.4 single monomer moves to 1 crankshaft move, although this ratio depends on the compactness of the chain. The percentage of moves tried that then pass the Metropolis criterion are roughly 30% for simple moves (individual single monomer and crankshaft moves), 19% for composites of two simple moves, 11% for composites of three simple moves, and 7% for composites of four simple moves. This yields an overall acceptance rate of roughly 23% for the entire simulation.

In principle, one could fix either B_0 or T and vary σ and N_B instead. However, the case of constant σ and N_B (σ= 0.25, N_B = 5) is closer to the case of real proteins in which one constructs sequences from a fixed set of amino acids. The value chosen for T was 0.8. This value differs from that reported in Dinner et al.[5] because all energy and temperature values in that paper were scaled so that the average standard deviation of the B_{ij} distributions was equal to unity for better comparison with the 27-mer work.[4,6] Since the unscaled average σ_B is 0.6, the scaled tempera-

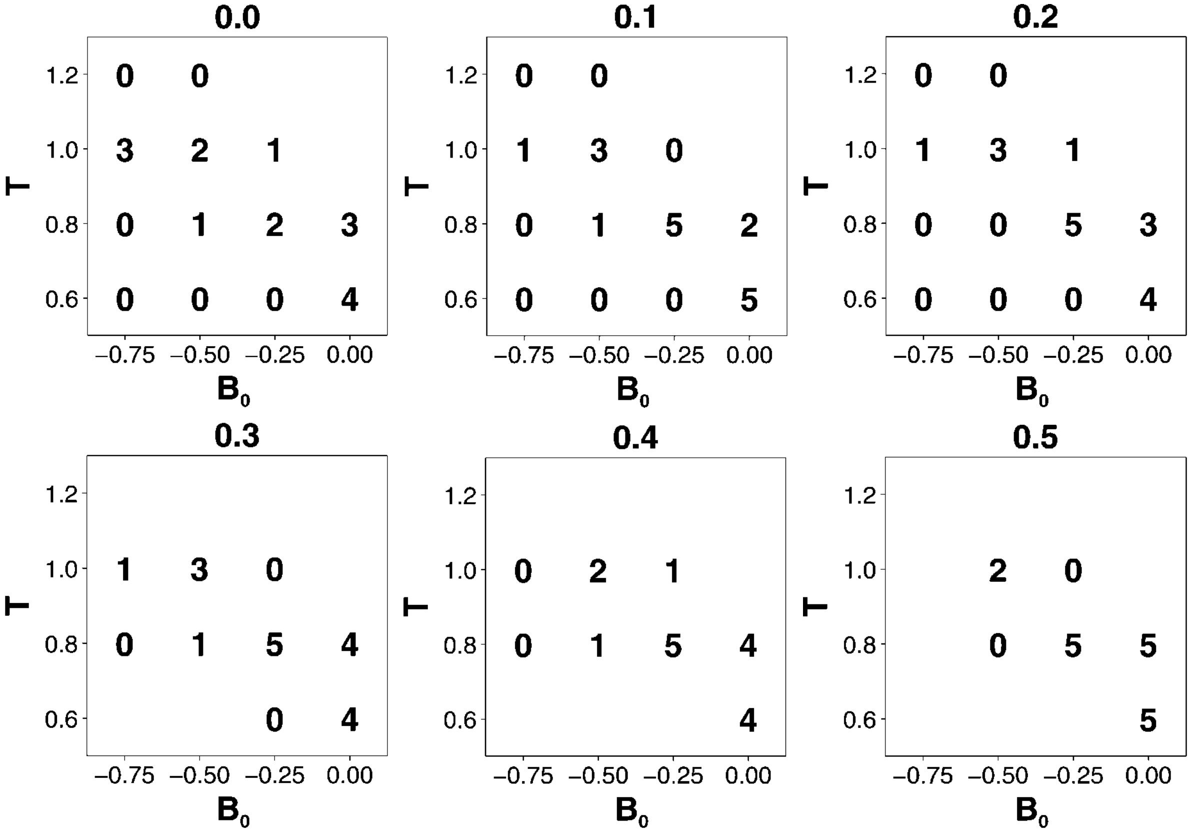

Fig. 4. Optimization of folding parameters B_0, T, and f_{cm} for sequence 4. Each number in the plot indicates the number of trials out of 5 that the sequence folded to the assigned native state within 50×10^6 MC steps; coordinates without numbers were not tested. Numbers above each panel indicate the fraction of composite moves (f_{cm}) (see text).

ture is $T = 0.8/0.6 \approx 1.33$, which corresponds to roughly the same optimum as for a fast-folding sequence in the 27-mer.[4,26] From equilibrium sampling over a range of temperatures for a few sequences, we estimate the melting temperatures (T_m) of the sequences studied to be in the range $0.9 < T_m < 1.1$, so that $T = 0.8$ is expected to be lower than T_m for most sequences.

On the other hand, B_0 has a smaller magnitude relative to the 27-mer. Here, $B_0 = -0.25$ (-0.42 when scaled by σ_B[27]), as opposed to $B_0 = -2.0$ in the 27-mer case. The choice of a weaker collapse parameter derives from the fact that our sequences are strongly optimized relative to random 27-mers. At stronger B_0 (more negative), less optimized sequences fold faster since a greater part of conformational space is eliminated, but more optimized ones fold slower since rearrangement is faster when the chain is noncompact[28] (Fig. 4).

Quantitative Structure-Property Relationships

In the present paper, we use a genetic neural network (GNN) to obtain QSPRs[14,15] for a more complete analysis of the relation between different system properties and folding ability. As described in the papers that introduced the GNN algorithm for ligand design,[14,15] the sequence attributes (descriptors) are selected by a genetic algorithm (GA),[29] and the functional dependence of the response data on the chosen descriptors is derived by an artificial neural network[30] (for reviews of the application of these methods to chemical and biological problems, see refs. 31–34). The GNN method combines the evolutionary and parallel features of genetic algorithms with the nonlinear features of neural networks to obtain optimal descriptions of complex relationships. The specific GNN protocol has been published.[15] In the current implementation, we used an evolutionary programming genetic algorithm[14,35] and trained the neural network with a pseudo-second derivative algorithm, the scaled conjugate gradient (SCG) method.[36] The SCG has been shown to yield good convergence with no critical dependence on arbitrarily defined scalar parameters.[15,36] All the QSPRs presented in the present study are three-descriptor models of the folding ability. Since both the genetic algorithm and the training of the neural network are nondeterministic (stochastic), we performed three trials for each QSPR, each with a different seed for the random number generator. We used 200 individuals and 50 evolutionary programming genetic cycles in each trial; these numbers are of the order of those employed previously.[15] Models with more than three descriptors tended to yield only marginally higher correlations with less consistency from trial to trial of the descriptors chosen by the GA, thereby complicating interpretation of the results.

Separate QSPRs were derived for each secondary structure class (helix and sheet) and for the entire database. In each of the three cases, results were obtained for each sequence using all the sequences available (training set) and using all the sequences except the one in question (cross-validated predictions). The success of the model derived by the GNN was evaluated by the Pearson linear correlation coefficient:

$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_x} = \frac{\sum_i^n (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\left[\sum_i^n (x_i - \langle x \rangle)^2 \sum_i^n (y_i - \langle y \rangle)^2\right]}} \quad (4)$$

between the observed (x) and predicted (y) values. A complete GNN trial using all 200 sequences requires approximately 17 CPU hours on a R5000 Silicon Graphics Indy workstation.

Measures of Pairwise Structural Similarity

To characterize the similarity of two structures (denoted n and m), we use three different measures. The first of these is the fractional contact overlap between the two structures (Q_{nm}). As a shorthand, we denote the overlap between non-native and the native states of the same sequence as Q_0.

Second, we use the root-mean square deviation between all distance pairs ($DRMS$):

$$\frac{\sqrt{2 \sum_{i>j+1} (d_{ij}^m - d_{ij}^n)^2}}{n - 2} \quad (5)$$

where d_{ij}^n is the distance between monomers i and j in structure n. We employ this measure of pairwise similarity in addition to Q_0 because similar structures that are shifted slightly relative to the primary sequence indices will have a low contact overlap (low Q_0), but the corresponding monomers will be close spatially (low $DRMS$).

Third, we use the fractional overlap of "pseudo-dihedral" angles in the structures (ϕ_{nm}). Here we take a pseudo-dihedral to be any one of the six possible conformations that are unrelated by reflection or rotation for a set of three bonds (four monomers). This is a purely local measure and so is more indicative than the other two measures of structures that have the same secondary structure but different tertiary structures.

Measures of Folding Ability

The folding behavior of a sequence is measured both by the number of times it reached the native state in 10 trials (N_f) and by the average contact overlap between the native structure and the lowest energy structure (E_{min}) sampled during each trial ($Q_m = \overline{Q}_{0E_{min}}$). Although Q_{0Emin} is not the highest Q_0 (computation of the latter requires more time), it was typically observed to be within 0.05 of the highest Q_0.[5] Q_m identifies sequences that repeatedly get close to the native state but fail to find it. We did not use the mean first passage time to quantify folding ability (as was done by Klimov and Thirumalai[11,12]) even though it may be a more sensitive measure because its determination requires that each trial reach the native state, which was not computationally feasible for such large chains.

Sequence Descriptors

We considered 47 different sequence attributes as inputs to the QSPR GNN. They fall into five categories: 1) properties of the B_{ij} distribution, 2) the overall stability of the native state, 3) the structure content of the native state, 4) correlations between sequence and native structure, and 5) breakdowns of the energy and secondary structure content of the native state by monomer exposure. Most of those in the first four categories were chosen because they have been suggested elsewhere[4,17,20,37–39] to be of importance for folding. The descriptors we added that have not been the subject of previous studies were motivated by our analysis of the folding mechanism of a few sequences[5] and are designed to characterize differences between the contacts that make up the core of the native structure and those that make up its surface.

B_{ij} distribution

Although the distributions of B_{ij} for all the sequences are similar, slight differences exist. We consider the following measures of the overall distribution: the mean ($\langle B_{ij} \rangle$), the standard deviation (σ_B), and the third moment normalized by the standard deviation (α_{B^3}), the minimum B_{ij} value ($B_{\min}$), and the maximum B_{ij} value ($B_{\max}$). The third moment measures skewness of the distribution. In addition, we consider two measures of the native B_{ij} (B_{ij}^{nat}): the ratio of the average native B_{ij} to the average of all B_{ij} ($\langle B_{ij}^{\text{nat}} \rangle / \langle B_{ij} \rangle$) and the standard deviation of the native interactions (σ_B^{nat}).

Stability of the native state

Since it is not possible to obtain an exact measure of the stability of the native structure (for example,

T_m) without excessive amounts of computer time for a system as large as the 125-mer, we use several estimates of stability. The simplest is the energy of the assigned native state (E_0). The efficacy of E_0 as a measure of stability depends on the rest of the spectrum being essentially the same for all sequences; if this assumption is true, then the lower E_0, the greater the stability. To include explicitly details of the rest of the spectrum, we have three measures of the energetic difference between the native state and the quasi-continuous part of the energy spectrum. First, we have the difference between the native state and an average maximally compact state[5,21,40,41]:

$$\Delta/\sigma_B = \frac{\langle E\rangle - E_0}{\sigma_B} \tag{6}$$

where $\langle E\rangle = C_0(\langle B_{ij}\rangle + B_0)$ with C_0 the number of contacts in a $5 \times 5 \times 5$ structure (176). This measure is similar to the Z-score of the native state but differs in sign and normalization (the normalization used for the Z-score is the standard deviation of the distribution of energies rather than B_{ij}).[42] Second, we have the separation between the ground state and the upper limit of the discrete spectrum[16]:

$$\Delta_b = \langle E\rangle - N\sigma_B\sqrt{2\rho \ln \gamma} - E_0 \tag{7}$$

where N is the number of monomers (125), ρ is the average number of contacts between monomers (176/125 = 1.41), and γ is the scaling factor for the number of conformations as a function of length [estimated from the 27-mer compact self-avoiding ensemble with 103,346 structures[43] to be $(103{,}346)^{1/27} = 1.53$]. Third, we have the separation between the ground state and the lower limit of the continuous spectrum[43]:

$$\Delta_c = \langle E\rangle - \sigma_B\sqrt{2\rho N} - E_0. \tag{8}$$

The differences between Δ/σ_B, Δ_b, and Δ_c are shown in Figure 5.

We did not include the energy gap between the ground and first excited compact states (ΔE_{10}), which was found to be a good indicator of the native stability in Šali et al.,[4] because it could not be calculated easily for the 125-mer.

Native structure

To characterize the overall native structure, we count both the number of contacts with a particular contact order [$C_0(|i - j|)$] and the number of contacts involved in each secondary structure (C_h, C_{ps}, C_{as}, and C_t for helices, parallel sheets, antiparallel sheets, and turns, respectively). For the $C_0(|i - j|)$, we consider $i - j = 3, 5, 7$, and 9 and the ranges $3 \leq |i - j| \leq 9$, $11 \leq |i - j| \leq 19$, $21 \leq |i - j| \leq 123$. The $C_0(|i - j|)$ give information about the distribution of short- and long-range contacts, but lack information about the spatial arrangement of contacts that is important for assessing cooperativity and kinetic accessibility. We consider contacts "cooperative" if formation of any one contact increases the probability of the formation of the others. On the lattice, sheets are strongly cooperative, turns and $i, i + 3/i, i + 5$-type helices are weakly cooperative, and $i, i + 3$-type helices are non-cooperative (Fig. 6).[17] As for kinetic accessibility, by which we mean contacts that are easily found by a random search, turns and helices involve only short-range contacts and so are always kinetically accessible. The kinetic accessibility of sheets depends on their shortest range contacts. Antiparallel sheets are typically more kinetically accessible than parallel

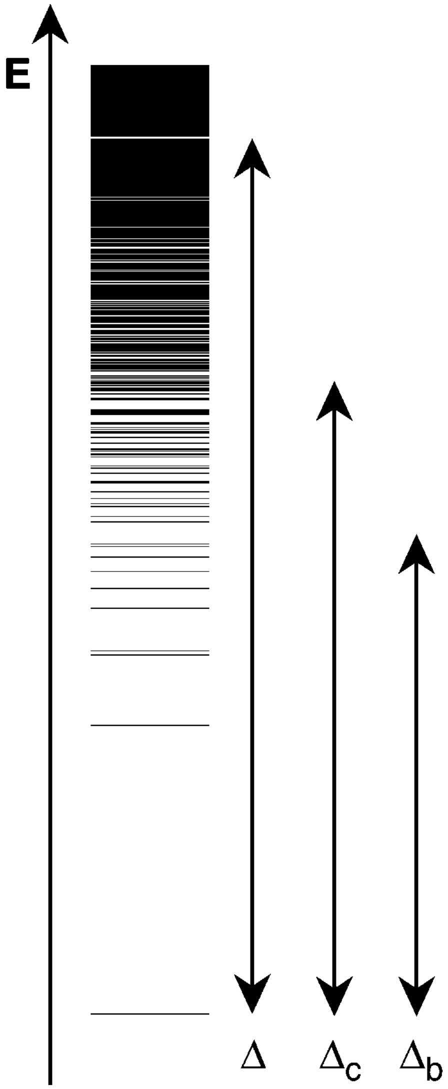

Fig. 5. Schematic illustrating the differences between three measures of the native state stability. Shown are the gaps between the native state and the estimated top of the discrete spectrum (Δ_b), between the native state and the bottom of the quasi-continuous part of the spectrum (Δ_c) and between the native state and the unweighted average energy (Δ).

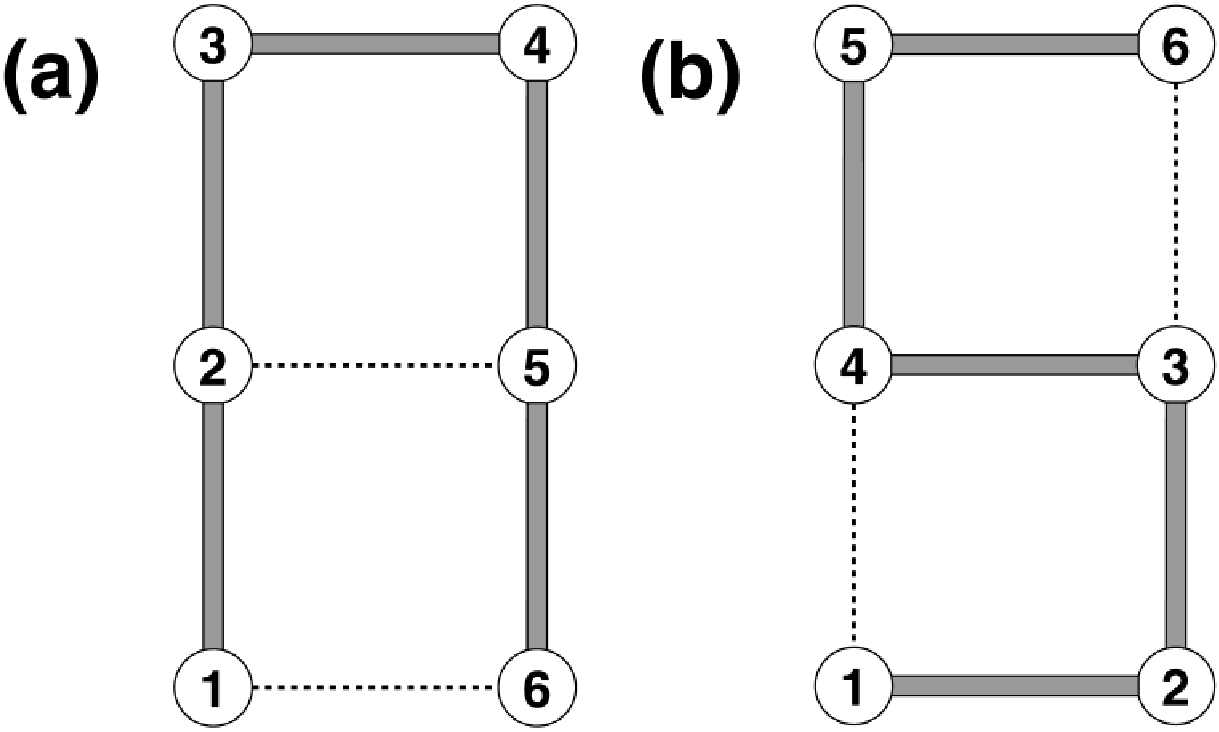

Fig. 6. Comparison of the cooperativity of lattice secondary structures. (a) A turn or sheet structure; formation of contact (2,5) increases the likelihood that (1,6) will be found in a random search by decreasing the available conformational space. (b) An *i,i*+3 helix structure; formation of contact (1,4) does not substantially affect the likelihood that (3,6) will be found in a random search.

ones because parallel sheets must loop around before they can initiate, thus requiring long-range contacts.

Correlations between sequence and structure

The correlation between sequence and native structure is assessed by a variety of measures. First, we study r_{BC}, the Pearson linear correlation coefficient (Eq. 4) between the B_{ij} and the C_{ij} where $C_{ij} = -1$ if i and j are in contact in the native structure and $C_{ij} = 0$ otherwise. The sign of the native C_{ij} was chosen to yield a positive correlation with favorable interactions ($B_{ij} < 0$). Next, we introduce several measures that are based on contact weighted averages. To this end, we define the "position" of a contact in the native structure between monomers i and j to be

$$\vec{r}_{ij} = \frac{(\vec{r}_i + \vec{r}_j)}{2}\,. \tag{9}$$

Then, the unweighted and B_{ij}-weighted averages are

$$\langle\vec{r}_{ij}\rangle = \frac{1}{C_0}\sum_{i<j}\Delta(\vec{r}_i - \vec{r}_j)\vec{r}_{ij} \text{ and}$$

$$\langle\vec{r}_{ij}\rangle_B = \frac{1}{E_0}\sum_{i<j}\Delta(\vec{r}_i - \vec{r}_j)\vec{r}_{ij}(B_{ij} + B_0) \tag{10}$$

where $\Delta\vec{\rho}_i - \vec{r}_j)$ is the kronecker delta that picks out contacts in the structure. The quantities that we study are the distance between these two averages ($|\langle\vec{r}_{ij}\rangle - \langle\vec{r}_{ij}\rangle_B|$), the total B_{ij}-weighted contact moment:

$$I_B = \sum_{i<j}\Delta(\vec{r}_i - \vec{r}_j)(B_{ij} + B_0)|\vec{r}_{ij} - \langle\vec{r}_{ij}\rangle_B|^2, \tag{11}$$

and the B_{ij}-weighted average contact radius squared ($R_B{}^2 = I_B/E_0$). Unlike I_B, $R_B{}^2$ separates the arrangement of the contacts from their magnitude. In addition, we test for correlations with the B_{ij}-weighted average contact-contact spatial distance:

$$\langle|\vec{r}_{ij} - \vec{r}_{kl}|\rangle_B = \sum_{i<j}\sum_{k<l}\Delta(\vec{r}_i - \vec{r}_j)\Delta(\vec{r}_k - \vec{r}_l)|\vec{r}_{ij} - \vec{r}_{kl}| \cdot \frac{B_{ij} + B_{kl} + 2B_0}{2E_0} \tag{12}$$

the B_{ij}-weighted average contact order ($\langle| i - j |\rangle_B$), and the Pearson correlation coefficient between the B_{ij} and the contact order ($r_{|i-j|\,Bij}$).

Breakdowns by monomer exposure

We break the native state energy, the number of native contacts, and the native secondary structure down by the exposure of the monomers participating in the contacts that contribute to each of these totals. Monomers are classified as either buried (4 contacts,

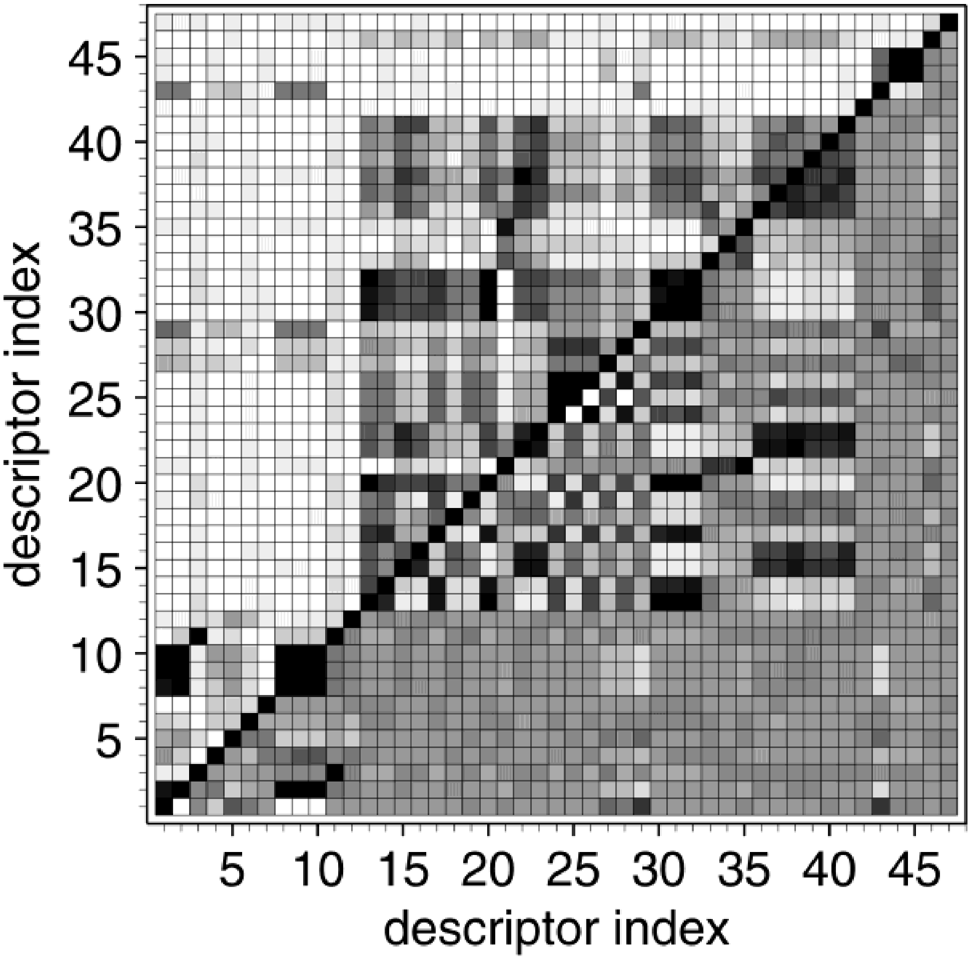

Fig. 7. Correlations between descriptors for all 200 sequences. (above) The absolute value of the Pearson linear correlation coefficient between the two descriptors corresponding to the *x*- and *y*-axes values (Table I). Values range between (white) 0 (no correlation) to (black) 1 (perfect correlation). (below) The signed value of the correlations shown above the diagonal. Values range from (white) –1 (perfect anti-correlation) to (black) 1 (perfect correlation).

denoted by *b*) or on the surface (3 contacts, denoted by *s*) of the 5 × 5 × 5 native structure. The superscripts *bb*, *sb*, and *ss* denote contacts between two buried monomers, a surface monomer and a buried monomer, and two surface monomers, respectively.

Pairwise correlations between descriptors

Many of the descriptors exhibit significant correlations. We show these correlations graphically in Figure 7.

Above the diagonal, we show the magnitude of each descriptor-descriptor correlation; a dark square indicates that two descriptors are correlated (or anticorrelated) in the 200 sequence database. Below the diagonal, we show the signed correlation coefficient; light squares correspond to anticorrelations and dark squares to correlations. Although some pairs of descriptors are almost perfectly correlated [such as Δ/σ_B (8), Δ_b (9), and Δ_c (10)], none actually is. The inclusion of highly correlated descriptors leaves the choice of which is a better predictor of folding ability to the genetic algorithm. Significant correlations are observed between the descriptors 13 through 41 because most (except for 27, 28, and 29) are native structure measures (see Native Structure, above).

RESULTS

The goal of this study is to determine which attributes of a sequence correlate with its ability to fold rapidly. Consequently, the database should not be homogeneous with respect to any attribute of interest since a dependence on that attribute would be obscured if all sequences had the same value. Because the native structure is the starting point for sequence generation, it is possible to maximize the variability in the database by including only structures that are dissimilar. Three measures of similarity are shown in Figure 8.

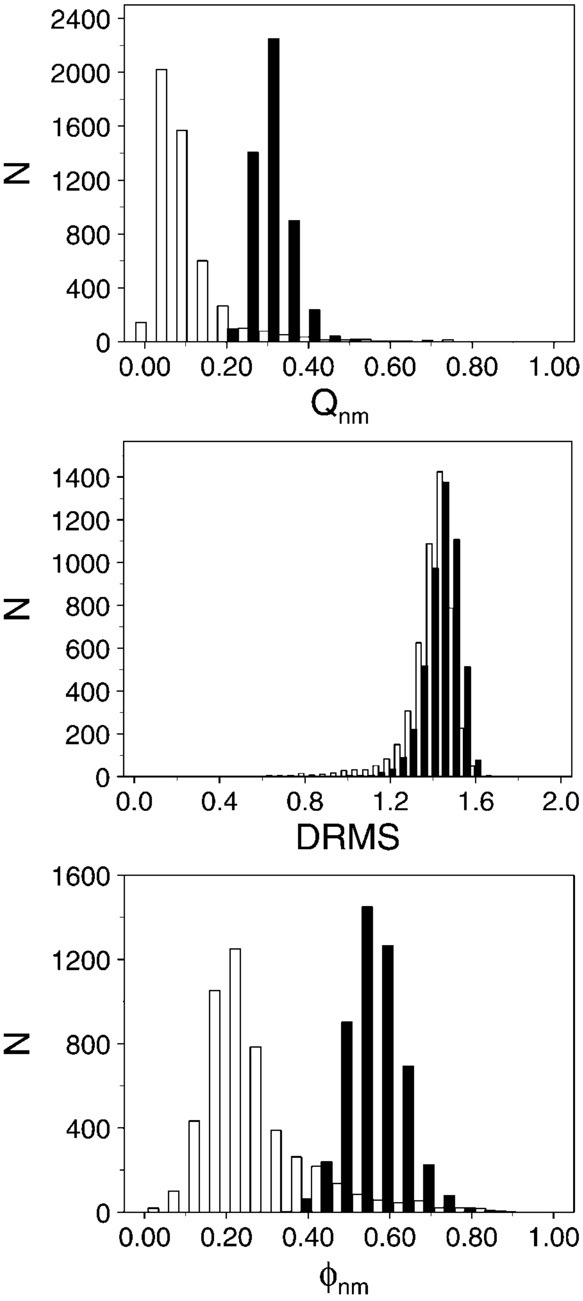

Fig. 8. Pairwise similarity of helix (black) and sheet (white) native structures. (a) Q_{nm}. (b) *DRMS*. (c) ϕ_{nm}.

Each of these is a pairwise similarity and is calculated for all helix-helix structure pairs and all sheet-sheet structure pairs separately; the helix-sheet structure pairs (not shown) exhibit low similarity by all three measures. The first of the measures is the pairwise fractional contact overlap of two structures (Q_{nm}). Of all the measures, it is the most closely related to the energy function, which is a sum over contacts. Both the helix and sheet sequences exhibit pairwise Q_{nm} peaked below 0.4, indicating that there is little contact overlap between native structures. The helix peak is higher than that for the sheets because the larger number of short-range helix contacts are more likely to overlap by chance. The next measure is the pairwise $DRMS$. For both types of sequences, the pairs cluster around $DRMS \approx 1.4$. Given that the length of each cube edge is 4 lattice units (5 monomers and 4 bonds), this again indicates that most structure pairs exhibit low similarity. Indeed, if we scale the units by 3.8, which corresponds to the distance in angstroms of a real C_α-C_α pseudo-bond, the peak in the distribution of $DRMS$ falls at 5.32 angstroms, which corresponds to substantial dissimilarity for a fully compact structure. The third measure is the fraction of overlapping pseudo-dihedrals, ϕ_{nm}. As expected, it is high for helix sequences, since if a significant fraction of the structures is helical, the helices must overlap. On the other hand, it is low for sheet sequences, since there is no reason for the sheets to cluster at particular residues. Thus, there is significant local overlap in helix sequences, but little tertiary overlap for all the sequences.

As stated earlier, the sequences are engineered to have substantial amounts of secondary structure in the native state. The distributions of C_h, C_{ps}, C_{as}, and C_t are shown separately for helix and sheet sequences in Figure 9.

Although many of the sheet sequences are composed almost entirely of sheet contacts (particularly antiparallel sheet, Fig. 9c), none of the helix sequences are composed entirely of helix contacts (Fig. 9a). This asymmetry stems from the fact that antiparallel sheet contacts can be both short- and long-ranged, but that helix contacts can be only short-ranged. The stipulation that the native structure be a 5 × 5 × 5 cube mandates a certain fraction of long-range contacts, preventing the creation of structures with all helix contacts. The helix and sheet sequences exhibit comparable numbers of parallel sheet contacts. These are almost all long-ranged ($i - j \geq 11$) since the chain must loop around to form such contacts.

Once the structures have been chosen, the B_{ij} are assigned in the manner described in Sequence Optimization, above. The degree of optimization is assessed by Δ_c (Fig. 10) and r_{BC} (Fig. 11). The helix and sheet distributions for both these descriptors are almost the same, indicating that the energetic optimization is independent of structure type. This independence is confirmed by the lack of correlations between the stability descriptors (indices 1–2 and 8–10) and the structure descriptors (indices 13–26 and 30–41) (Fig. 7). The mean of the Δ_c distribution is about 156.3, and its standard deviation is 7.5. Since the standard deviation is small relative to the mean, most of the sequences are comparably opti-

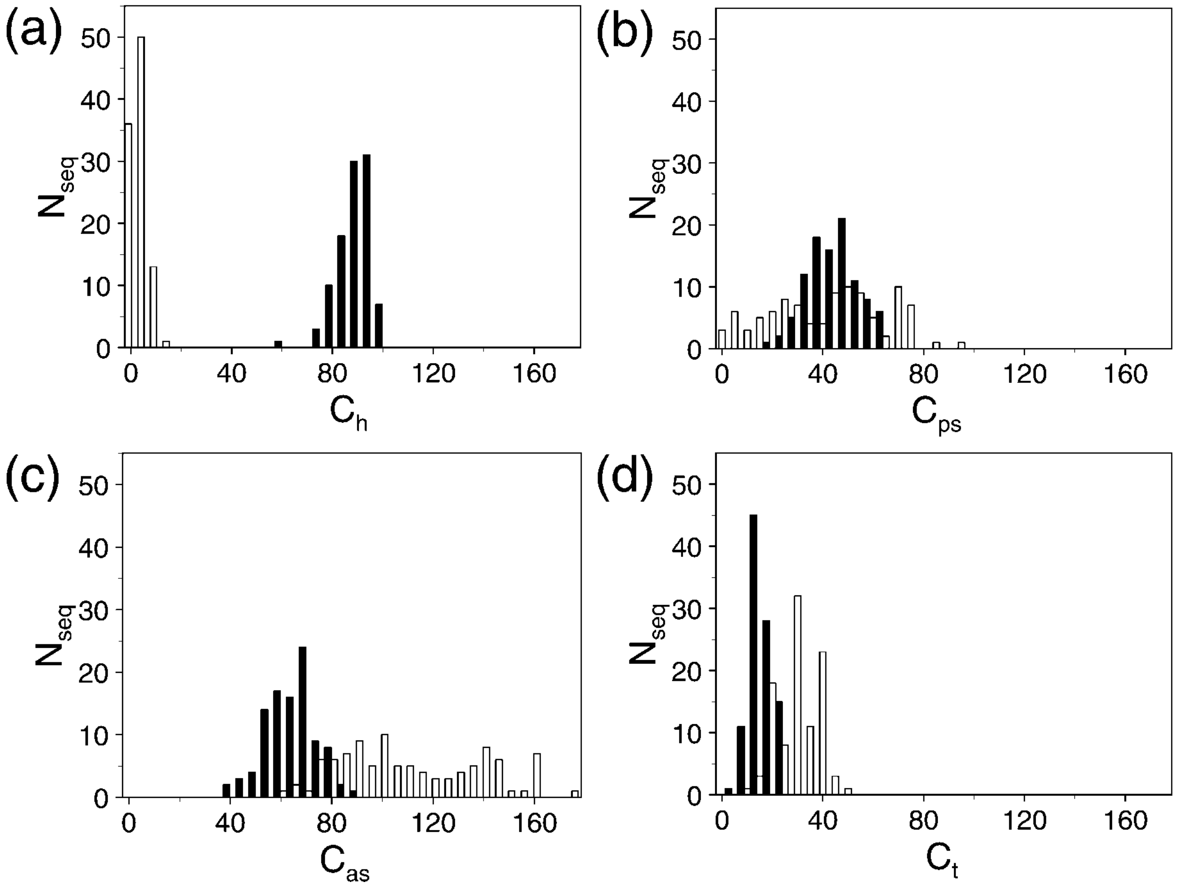

Fig. 9. Distribution of native state secondary structure content for helix (black) and sheet (white) sequences. (a) Helix contacts. (b) Parallel sheet contacts. (c) Anti-parallel sheet contacts. (d) Turn contacts.

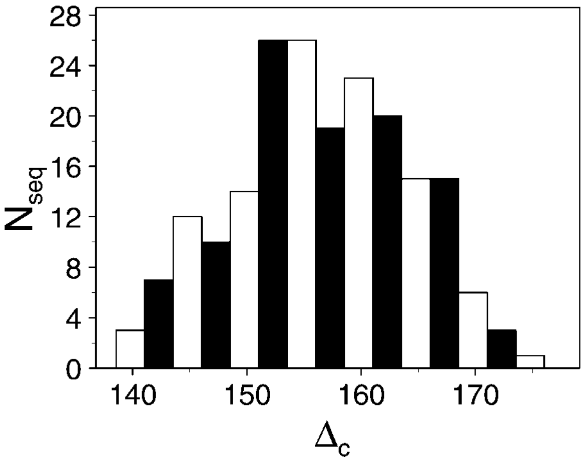

Fig. 10. Distribution of Δ_c for helix (black) and sheet (white) native structures.

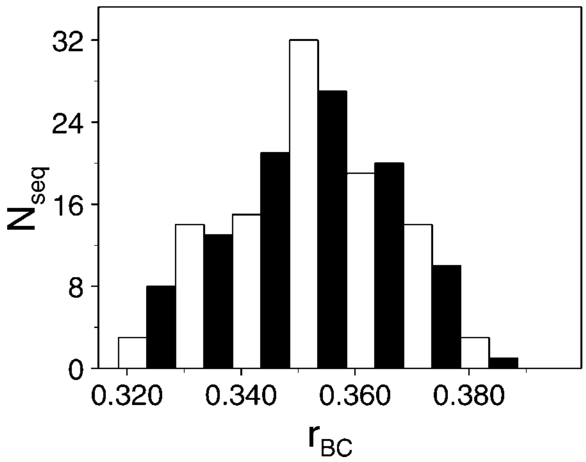

Fig. 11. Distribution of r_{BC} for helix (black) and sheet (white) native structures.

mized; nevertheless, there is a spread in the distribution, so that a comparison of the folding ability with the stability can be made. As for r_{BC}, the distribution is peaked around 0.35. This value is comparable to that for randomly generated 27-mers.[4] However, the sequences are clearly optimized for folding relative to random ones since the much larger fraction of non-native contacts makes the 125-mer correlation more significant. A comparison (Dinner and Karplus, unpublished data) suggests that the present se-

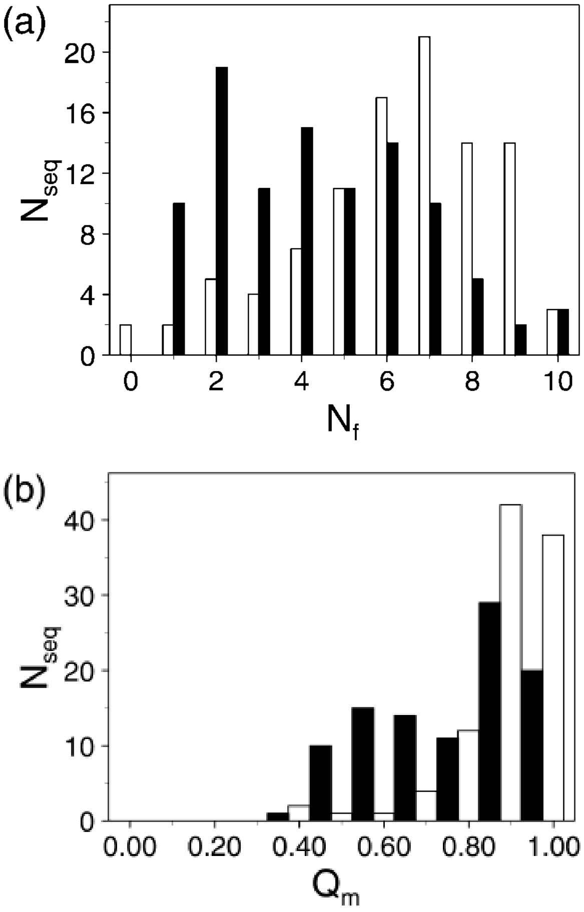

Fig. 12. Distribution of folding abilities for helix (black) and sheet (white) sequences. (a) N_f. (b) Q_m.

quences are comparable in stability to those obtained with other optimization techniques.[21,22]

The optimization increases significantly the fraction of sequences capable of folding in a reasonable time. For the random 27-mer, the distribution of N_f is peaked at $N_f = 0$ and decreases roughly exponentially with N_f (Fig. 4 of Šali et al.[4]). The 125-mer optimized sequences peak at intermediate values of N_f: 2–4 for helix sequences and 6–7 for sheet sequences (Fig. 12). For the distribution of Q_m, both sequence types are peaked in the range $0.8 < Q_m < 1.0$. This implies that the primary role of the optimization is to facilitate reaching a near-native state; its role with respect to rearrangement from that state to the native one is less clear since overstabilization of native contacts can slow this final step of the folding process.[5] From Figure 12, it is evident that the sequences in the database have a much stronger tendency to fold than random sequences. However, such a shift in the distribution does not bias the analysis because conclusions are based only on comparisons of sequences optimized by the same procedure. All that is necessary, as already mentioned, is that there be considerable variability among the sequences that are compared.

In most cases in which optimization procedures have been used, it is not guaranteed that the structure assigned to be the ground state actually is the one of lowest energy. Usually this is verified by long MC searches; in very simple models, this can be shown by exhaustive enumeration.[7,44,45] For the present case, the ground state could not be proved to be correct. However, during the trials used to determine the folding ability of each sequence, only two sequences (one sheet and one helix) were found to have noncompact states of lower energy. The two states were only –0.024 below the assigned $E_0 = -211.464$ for one sequence and –0.373 below the assigned $E_0 = -224.361$ for the other. Due to the high similarity between the assigned and observed ground states ($Q_0 = 0.972$ and $Q_0 = 0.966$, respectively), Q_m for these sequences was not substantially affected, and the N_f differed from those that would be obtained using the lower energy (noncompact) structures by 2 and 1, respectively.

As described above, the observed folding mechanism[5] can involve two slow steps, either of which, in principle, can prevent the chain from folding on the time scale of 50×10^6 MC steps. The distribution of Q_0 for the lowest energy structures from each trial that failed to reach the native state are shown as a function of sequence type and N_f in Figure 13.

More helix sequences get stuck at the first barrier, formation of a stable core, whereas more sheet sequences get stuck at the second, rearrangement from an intermediate to the native state. This can be explained by the differences between helix and sheet contacts. Helix contacts are noncooperative and thus have difficulty forming a stable core. Sheet contacts are cooperative and thus have difficulty breaking to rearrange. This also explains the difference in the observed distribution of N_f for the sequences (Fig 12). A sequence that has difficulty forming the core is more likely to fail repeatedly to find the ground state during the allowed time period (hence the lower peak for helix sequences) than one that has difficulty rearranging due to the relative sizes of the conformational spaces that must be searched at these two steps.

GNN Analysis

In the earlier analysis of the folding simulations[5] we used a linear correlation measure with a limited number of descriptors and found that Q_m correlated most strongly with Δ/σ_B, C_{as}, and C_t. As a first step in the interpretation of the GNN approach, we show in Table I the results obtained from the (nonlinear) neural network in predicting N_f and Q_m using the individual descriptors. To highlight features of the table, the cross-validated Pearson correlation coefficients (r_{cv}) for the prediction of Q_m are shown graphically in Figure 14.

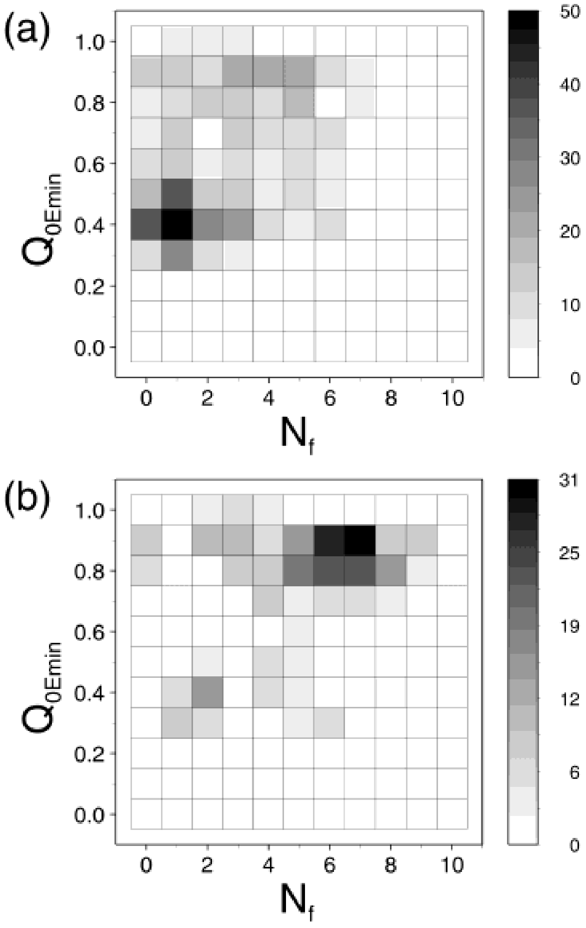

Fig. 13. Distribution of Q_{0Emin} as a function of N_f for trials that did not end in the assigned native state for (a) helix and (b) sheet sequences.

It should be noted that the correlations given are between calculated and predicted values, so that any measure that is of significance should give positive correlations, even if the values of that descriptor are anticorrelated with folding ability. When all sequences are considered, the strongest cross-validated correlations between the predicted and observed Q_m values (between 0.500 and 0.515) are those for kinetically accessible cooperative secondary structure (C_{as}) and for the energy gaps (Δ_b and Δ_c). The correlations for some of the other secondary structure measures (C_h and C_t) and the other measure of the energy gap (Δ/σ_B) are almost as high ($r_{QmCh} = 0.467$, $r_{QmCt} = 0.482$, and $r_{Qm\Delta/\sigma B} = 0.484$). For all sequences, the energy gap is not as good a predictor of Q_m as is the secondary structure of the native state (C_h, C_{as}, and C_t), but, if we restrict the QSPR calculation to either secondary structure class, the energy gap becomes the strongest predictor. This finding indicates that secondary structure of the native state is important but that once a helix or sheet native structure is chosen, the energy gap is the single most important determinant of folding ability. The nonlinear cross-validated predictivities are higher than can be obtained with linear regression, but the results are in qualitative agreement with the earlier analysis of the database.[5]

In the earlier analysis, it was found that consideration of two descriptors at a time substantially enhances predictivity; the best pair was Δ/σ_B and C_t. With the aid of the genetic algorithm described in Quantitative Structure-Property Relationships, above, three descriptor correlations can easily be obtained; the results are shown in Table II. Since both the genetic algorithm and the method for training the neural network are nondeterministic, different results can be obtained with different random number seeds. To illustrate the variation in the results, three lines are shown for each case (all, helix, sheet). The best predictions (as measured by r_{cv}) are shown in Figure 15.The correlations are higher than any of the single descriptor results. Those for Q_m are notably higher than those for N_f. This difference stems from the fact that the N_f data are more noisy because this measure of folding ability is an integer so that a single atypical trial out of the 10 can substantially change its value. It should be mentioned that the population of models in the genetic algorithm simulation includes other descriptor choices than those shown. Only the highest correlating ones are shown. However, most of the other individuals chosen by the GA contain descriptors with high correlations to those shown in Table II, so that those models exhibit comparable predictivities. For example, the second best GA individual for the first seed in the prediction of N_f for all sequences has E_0 instead of Δ_c (the other two descriptors are the same) with $r_{cv} = 0.646$ compared with $r_{cv} = 0.654$. Increasing the number of descriptors did increase the correlations (for example, in going from three to six descriptors, r_{cv} typically increased by about 0.05), but also tended to increase the variation in descriptors chosen with different random number generator seeds. We restrict our analysis to three descriptor models because the results did not show a marked improvement, and it is more likely that spurious correlations will be found with a larger number of descriptors.

To confirm that the model derived with the QSPR procedure is useful in making predictions, a test set of 20 additional sequences (10 helix and 10 sheet) was generated in the same manner as the database and was tested in an identical manner for the ability to fold. The correlations between the predicted and observed values are given in Table II and they are plotted along with the cross-validated predictions in Figure 15.

On the whole, the predictive accuracy is comparable to the cross-validated accuracy (which itself is not substantially lower than the complete training set accuracy). In some cases the test set accuracy is

TABLE I. Predictive Power of Each Descriptor Taken Singly[†]

		All				Helix				Sheet			
		N_f		Q_m		N_f		Q_m		N_f		Q_m	
Index	Descriptor	r_{tr}	r_{cv}	r_{tr}	r_{cv}	r_{tr}	r_{cv}	r_{tr}	r_{cv}	r_{tr}	r_{cv}	r_{tr}	r_{cv}
1	E_0	0.274	0.222	0.490	0.461	0.355	0.248	0.659	0.626	0.328	0.203	0.508	0.400
2	r_{BC}	0.293	0.251	0.507	0.473	0.355	0.252	0.630	0.579	0.313	0.218	0.559	0.460
3	$\langle B_{ij} \rangle$	0.179	0.106	0.199	0.131	0.193	−0.124	0.067	−0.257	0.169	−0.027	0.309	0.083
4	σ_B	0.187	0.099	0.289	0.207	0.175	0.000	0.356	0.269	0.198	−0.011	0.337	0.151
5	α_B^3	0.127	−0.014	0.231	0.175	0.202	0.062	0.345	0.241	0.110	−0.185	0.270	0.110
6	B_{min}	0.117	−0.142	0.177	0.048	0.018	−0.383	0.227	0.016	0.162	−0.033	0.221	0.085
7	B_{max}	0.023	−0.388	0.011	−0.146	0.240	−0.066	0.052	−0.373	0.143	−0.061	0.135	−0.182
8	Δ/σ_B	0.298	0.259	0.518	0.484	0.355	0.256	0.631	0.588	0.310	0.212	0.553	0.458
9	Δ_b	0.308	0.261	0.533	0.500	0.365	0.279	0.656	0.617	0.339	0.237	0.597	0.496
10	Δ_c	0.295	0.253	0.540	0.509	0.367	0.276	0.667	0.628	0.360	0.271	0.618	0.507
11	$\langle B_{ij}^{nat} \rangle / \langle B_{ij} \rangle$	0.233	0.157	0.310	0.261	0.213	−0.012	0.247	0.061	0.197	0.033	0.413	0.230
12	σ_B^{nat}	0.000	−0.017	0.029	−0.084	0.059	−0.327	0.064	−0.246	0.362	0.289	0.366	0.256
13	$C_0(3)$	0.546	0.505	0.500	0.468	0.184	−0.089	0.061	−0.225	0.273	0.069	0.251	0.033
14	$C_0(5)$	0.499	0.440	0.421	0.372	0.112	−0.130	0.036	−0.349	0.391	0.293	0.347	0.270
15	$C_0(7)$	0.520	0.491	0.479	0.452	0.081	−0.165	0.205	−0.010	0.311	0.191	0.281	0.184
16	$C_0(9)$	0.446	0.406	0.414	0.386	0.132	−0.062	0.075	−0.199	0.234	0.077	0.352	0.208
17	$C_0(3–9)$	0.443	0.404	0.393	0.337	0.057	−0.309	0.096	−0.151	0.319	0.199	0.325	0.226
18	$C_0(11–19)$	0.336	0.292	0.310	0.263	0.005	−0.229	0.189	−0.146	0.196	−0.002	0.181	−0.041
19	$C_0(21–123)$	0.255	0.193	0.205	0.139	0.041	−0.407	0.083	−0.160	0.240	0.103	0.211	0.074
20	C_h	0.526	0.495	0.487	0.467	0.016	−0.481	0.214	0.062	0.211	−0.009	0.172	−0.210
21	C_{ps}	0.309	0.261	0.286	0.230	0.167	−0.051	0.227	−0.032	0.355	0.252	0.380	0.181
22	C_{as}	0.590	0.572	0.536	0.515	0.094	−0.194	0.170	0.008	0.469	0.387	0.428	0.317
23	C_t	0.572	0.550	0.508	0.482	0.186	0.018	0.135	−0.121	0.399	0.339	0.350	0.270
24	C_0^{bb}	0.369	0.337	0.298	0.246	0.212	−0.157	0.020	−0.425	0.016	−0.532	0.191	−0.066
25	C_0^{sb}	0.393	0.359	0.298	0.271	0.058	−0.121	0.128	−0.069	0.251	−0.010	0.084	−0.110
26	C_0^{ss}	0.392	0.345	0.297	0.275	0.198	−0.118	0.112	−0.337	0.035	−0.225	0.065	−0.212
27	E_0^{bb}	0.122	0.001	0.128	0.050	0.300	0.084	0.415	0.351	0.287	0.158	0.341	0.230
28	E_0^{sb}	0.367	0.329	0.373	0.336	0.336	0.251	0.204	0.052	0.080	−0.159	0.122	−0.092
29	E_0^{ss}	0.152	0.052	0.043	−0.197	0.008	−0.418	0.292	0.151	0.046	−0.244	0.208	−0.027
30	C_h^{bb}	0.528	0.494	0.484	0.454	0.022	−0.403	0.045	−0.125	0.185	−0.007	0.194	−0.052
31	C_h^{sb}	0.534	0.499	0.488	0.471	0.212	−0.020	0.166	−0.008	0.020	−0.555	0.153	−0.388
32	C_h^{ss}	0.524	0.497	0.487	0.468	0.004	−0.421	0.187	0.038	0.222	−0.097	0.144	−0.214
33	C_{ps}^{bb}	0.383	0.329	0.314	0.233	0.235	0.026	0.100	−0.209	0.285	0.184	0.246	0.116
34	C_{ps}^{sb}	0.220	0.113	0.202	0.085	0.145	−0.101	0.130	−0.119	0.207	0.085	0.255	0.164
35	C_{ps}^{ss}	0.275	0.229	0.229	0.194	0.200	−0.202	0.042	−0.375	0.392	0.291	0.525	0.293
36	C_{as}^{bb}	0.488	0.452	0.440	0.407	0.180	0.046	0.191	0.010	0.369	0.301	0.343	0.207
37	C_{as}^{sb}	0.453	0.421	0.415	0.377	0.075	−0.125	0.102	−0.215	0.225	0.112	0.207	0.058
38	C_{as}^{ss}	0.569	0.542	0.516	0.491	0.054	−0.267	0.176	0.039	0.444	0.352	0.408	0.286
39	C_t^{bb}	0.445	0.411	0.431	0.401	0.109	−0.152	0.229	0.096	0.291	0.185	0.294	0.139
40	C_t^{sb}	0.494	0.468	0.391	0.356	0.208	0.022	0.079	−0.207	0.307	0.209	0.186	0.032
41	C_t^{ss}	0.478	0.452	0.438	0.413	0.120	−0.134	0.008	−0.614	0.252	0.068	0.277	0.155
42	$\lvert \langle \vec{r}_{ij} \rangle - \langle \vec{r}_{ij} \rangle_B \rvert$	0.048	−0.165	0.060	−0.165	0.115	−0.188	0.135	−0.122	0.111	−0.152	0.147	−0.080
43	I_B	0.068	−0.031	0.234	0.171	0.098	−0.227	0.408	0.298	0.027	−0.508	0.133	−0.024
44	R_B^s	0.268	0.181	0.278	0.225	0.197	−0.057	0.243	0.129	0.130	−0.046	0.197	0.045
45	$\langle \lvert \vec{r}_{ij} - \vec{r}_{kl} \rvert \rangle_B$	0.268	0.194	0.268	0.180	0.258	0.022	0.241	0.128	0.114	−0.093	0.195	0.008
46	$\langle \lvert i - j \rvert \rangle_B$	0.275	0.207	0.255	0.191	0.230	0.113	0.046	−0.381	0.235	0.091	0.182	−0.076
47	$r_{\lvert i-j \rvert B_{ij}}$	0.009	−0.686	0.062	−0.116	0.138	−0.102	0.192	−0.006	0.072	−0.163	0.001	−0.672

[†]Predictivity is measured by the Pearson correlation coefficient for training on the entire sequence set (r_{tr}) or on the sequence set without the sequence whose behavior is to be predicted (r_{cv}).

actually somewhat higher. This slight increase is most probably due to a chance increase in the correspondence between N_f and Q_m; the second barrier to folding must be relatively low for these 20 sequences. In any case, the strong predictions on the test set indicate that the data are not overfitted and that the dependence of folding ability on the chosen descriptors is quite robust.

The trends in descriptor choice are quite striking. The energy gap Δ_c is selected in 15 of 18 trial cases. In the three in which it is not, $E_0{}^{bb}$ ($r_{\Delta cE0bb} = –0.336$), the interaction energy of the buried-buried contacts appears twice, and E_0 ($r_{\Delta cE0} = –0.969$) appears once. Thus, the stability of the native state, particularly the core, is clearly an important element in determining folding ability. The neural network functional

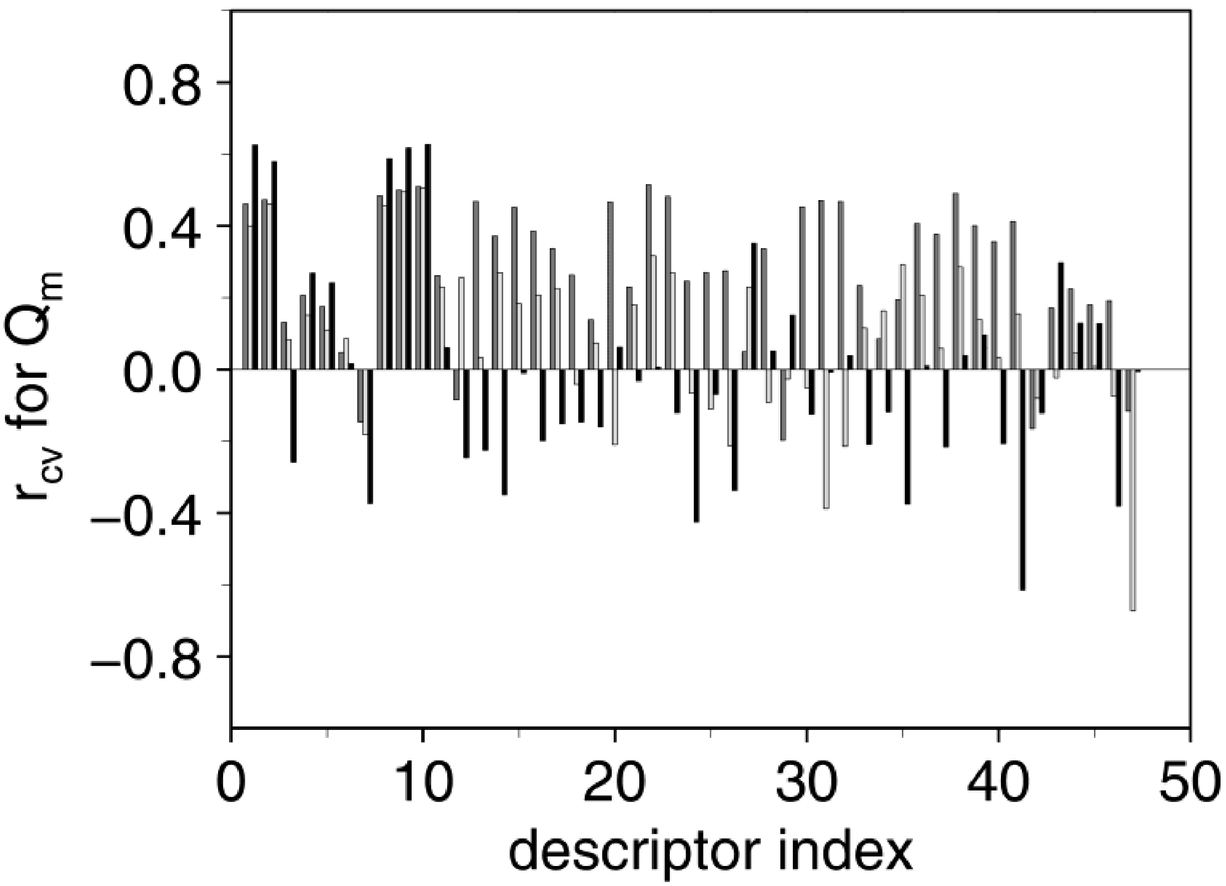

Fig. 14. Predictive power of single descriptors for all (dark gray), helix (black), and sheet (light gray) sequences. Descriptor indices are given in Table I.

TABLE II. Predictive Power of Three Descriptors At Once[†]

Class	Predicted	Seed	Descriptors chosen			r_{tr}	r_{cv}	r_{test}
All		1	Δ_c	C_{as}	E_0^{ss}	0.690	0.654	0.742
	N_f	2	Δ_c	C_{as}	E_0^{ss}	0.691	0.657	0.738
		3	Δ_c	C_{as}	E_0^{ss}	0.693	0.652	0.729
		1	Δ_c	C_{as}	E_0^{ss}	0.808	0.774	0.725
	Q_m	2	Δ_c	C_{as}	I_B	0.790	0.771	0.736
		3	Δ_c	C_{as}	E_0^{ss}	0.807	0.783	0.725
Helix		1	Δ_c	C_t	I_B	0.578	0.450	0.233
	N_f	2	E_0^{bb}	C_t^{sb}	E_0^{sb}	0.565	0.467	0.340
		3	E_0^{bb}	C_t^{sb}	$\langle\lvert\vec{r}_{ij} - \vec{r}_{kl}\rvert\rangle_B$	0.543	0.434	0.193
		1	Δ_c	E_0^{bb}	α_B^3	0.747	0.699	0.544
	Q_m	2	Δ_c	C_{as}^{bb}	C_h	0.732	0.681	0.570
		3	Δ_c	C_{as}^{bb}	C_h	0.734	0.679	0.547
Sheet		1	Δ_c	C_t	B_{min}	0.657	0.557	0.556
	N_f	2	E_0	C_t	B_{min}	0.630	0.542	0.695
		3	Δ_c	C_t	B_{min}	0.654	0.559	0.592
		1	Δ_c	C_{as}^{bb}	E_0^{ss}	0.857	0.834	0.903
	Q_m	2	Δ_c	C_{as}^{bb}	E_0^{ss}	0.853	0.833	0.929
		3	Δ_c	C_{as}^{bb}	E_0^{ss}	0.863	0.832	0.907

[†] Predictivity of the training set (r_{tr}), of cross-validated (r_{cv}), or on the test set (r_{test}).

dependence on a descriptor can be probed by fixing the values of the other two descriptors chosen by the GA (neural network inputs) and varying the descriptor of interest between its minimum and maximum values in the database. The result of this procedure is shown for Δ_c with fixed C_{as} and $E_0{}^{ss}$ in Figure 16a.

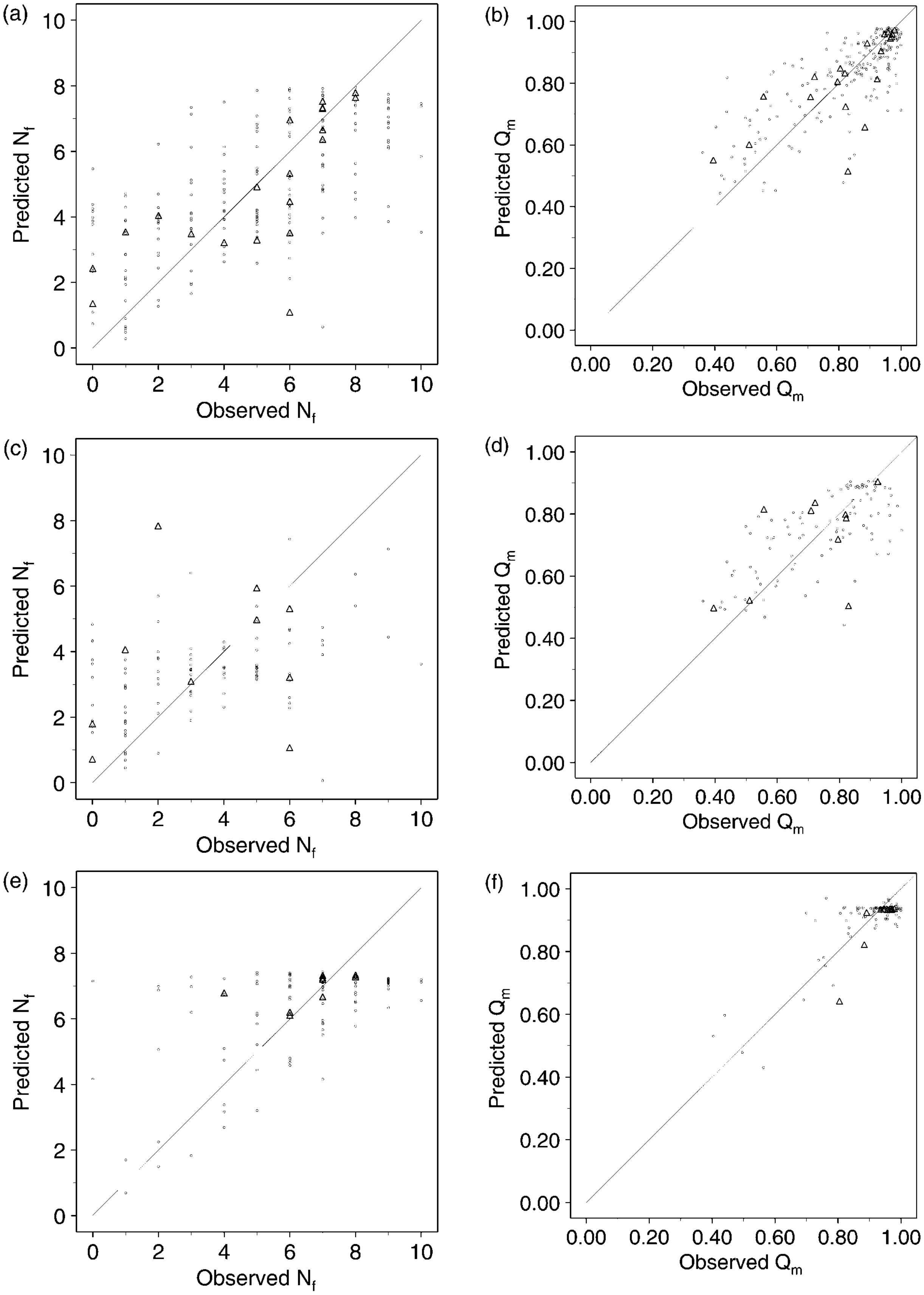
(a)
Predicted N_f
Observed N_f
0
2
4
6
8
10
(b)
Predicted Q_m
Observed Q_m
0.00
0.20
0.40
0.60
0.80
1.00
(c)
Predicted N_f
Observed N_f
(d)
Predicted Q_m
Observed Q_m
(e)
Predicted N_f
Observed N_f
(f)
Predicted Q_m
Observed Q_m

Figure 15.

Three such curves are shown: one in which the fixed descriptors (C_{as} and E_0^{ss}) are given their average values in the database, one in which the fixed descriptors have values corresponding to one of the most strongly folding sequences ($N_f = 10$ and $Q_m = 1$), and one in which the fixed descriptors correspond to the worst (as measured by Q_m) folding sequence ($N_f = 0$ and $Q_m = 0.361$). The functional dependence on Δ_c is similar for all three cases, indicating that it is not very sensitive to the values of the other parameters. Furthermore, there is clearly nonlinear behavior, which corresponds to a threshold-like response followed by saturation. This suggests that there is a critical stability that is necessary for folding but that an additional increase in the energy gap beyond a certain value does not enhance the likelihood of folding within the allotted time.

The next most commonly chosen features are measures of kinetically accessible cooperative secondary structure (see Native Structure, above); they are: C_{as}, C_{as}^{bb}, C_t, and C_t^{sb}. One of these is chosen in all except one of the 18 trials. This demonstrates their importance for folding to the native structure. Dependency curves for C_{as} (for all sequences, both helix and sheet) are shown in Figure 16b. Similar to those for Δ_c, the dependence on C_{as} begins to saturate about halfway through the range of observed values. However, there is a much softer response in the first half of the curve. Interestingly, the response curve for the mean value of the other (fixed) parameters is above that of the strongly folding sequence; the values of Δ_c and E_0^{ss} (152.6 and –101.51) are not optimal in this sequence.

The third parameter chosen is more variable. The one most commonly picked (8 of 18) is E_0^{ss}, the total energy of the contacts between two surface monomers. This parameter was suggested to be important in the earlier analysis of the database[5] because the off-pathway intermediates involve formation of native-like local domains involving primarily surface monomers. A stronger (more negative) surface energy makes it harder for the chain to break those contacts to allow rearrangement and condensation of the surface monomers. As can be seen in Figure 16c, as the surface energy becomes weaker (less negative), the folding ability increases. In contrast to the dependence on Δ_c, the dependence on E_0^{ss} is strongly dependent on the choice of the fixed parameter values; the shape of the curve obtained using values corresponding to the slowest folding sequence differs from the other two, which are obtained with more optimal fixed parameter values. The importance of the choice of fixed values (Δ_c and C_{as}) explains why the predictivity of E_0^{ss} by itself (Table I) is much lower than the predictivities of the descriptors that appear more frequently in the three descriptor models.

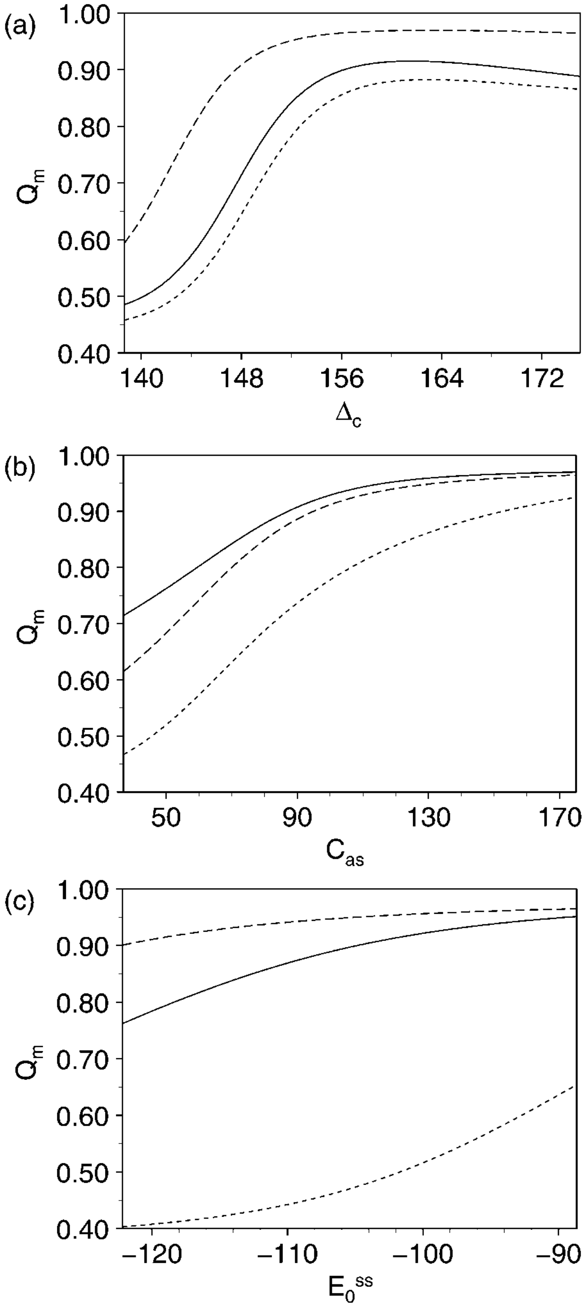

Fig. 16. Dependence of Q_m on chosen descriptors for all sequences (first seed). Each curve is obtained by fixing the other two descriptors and varying the one in question linearly between its minimum and maximum values in the database. Fixed values are set at their average values (solid). Fixed values are set at those of sequence 4, a sheet sequence which is the fastest folding sequence (long dash). Fixed values are set at those of sequence 127, a helix sequence which is the slowest folding sequence (short dash). (a) Δ_c is varied. (b) C_{as} is varied. (c) E_0^{ss} is varied.

Fig. 15. Comparison of observed and predicted measures of folding ability. Cross-validated (dots) and test set (triangles) predictions. (a) N_f all sequences. (b) Q_m all sequences. (c) N_f helix sequences. (d) Q_m helix sequences. (e) N_f sheet sequences. (f) Q_m sheet sequences. The data shown are for the best seed of each set in Table II (as measured by r_{cv}).

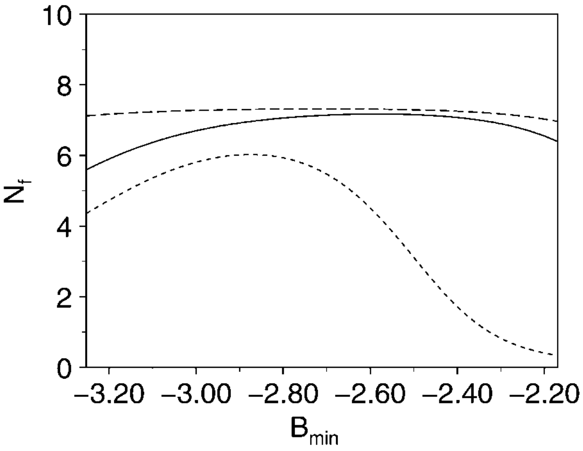

Fig. 17. Dependence of N_f on B_{min} for sheet sequences (first seed). Fixed values (Δ_c and C_t) are set at their average sheet values (solid). Fixed values are set at those of sequence 4, one of the fastest folding sheet sequences (long dash). Fixed values are set at those of sequence 54, the slowest folding sheet sequence (short dash).

A few other descriptors were chosen more than once. For example, the minimum B_{ij} value, B_{min}, was chosen for all three seeds of the prediction of N_f for sheet sequences. The functional dependence curve for this descriptor (Fig. 17) is essentially flat for most of the sequences, but it reaches a maximum at $B_{min} \approx -2.8$ for the sequence in the figure with $N_f = 0$ and $Q_m = 0.404$ (sequence 54). Weaker B_{min} yield lower predicted values of N_f probably because they indicate poor overall optimization (since they represent a relatively high lower bound for the native B_{ij}). However, the predicted N_f is lower for very strong B_{min} as well, perhaps because contacts with very strong B_{ij} are difficult to break so that they slow down conformational sampling.

The number of native contacts in helices, C_h, was chosen twice for the prediction of Q_m for helix sequences. The functional dependence for this descriptor is such that folding ability increases with an increase in C_h (Fig. 18). This result was not expected from the single descriptor analysis since, for all 200 sequences, folding ability is strongly anticorrelated with C_h, and, for just helix sequences, there is an insignificant positive correlation between C_h and Q_m and no measurable predictivity[5] (Table I). Both times that C_h is chosen, it is paired with Δ_c and $C_{as}{}^{bb}$. Thus, if a sequence has sufficient stability and core cooperativity, it folds faster with more helix contacts. Physically, C_h is chosen for essentially the same reason as $E_0{}^{ss}$: a larger number of helical contacts facilitates rearrangement of the surface residues in the kinetic intermediate.

The only other quantity that is picked more than once is the contact moment, I_B. This quantity was designed to detect when the lowest energy contacts of a sequence are clustered since such an arrangement would be more likely to form a stable folding core. I_B (Eq. 11) is always negative because $\langle B_{ij} + B_0 \rangle < 0$ and $|\vec{r}_{ij} - \langle \vec{r}_{ij} \rangle_B|^2 > 0$. The functional dependence curve increases monotonically with I_B, indicating that less negative values correspond to better folding (Fig. 19). For a given native state stability (set by Δ_c, which is always chosen with I_B), I_B is made less negative by decreasing the spatial separation of the lowest B_{ij} values. The observed dependence thus supports the idea that it is best to cluster the lowest B_{ij} values to promote both core formation and rearrangement at high Q_0.

It is worth noting that several quantities are never picked in the best models (and are not highly corre-

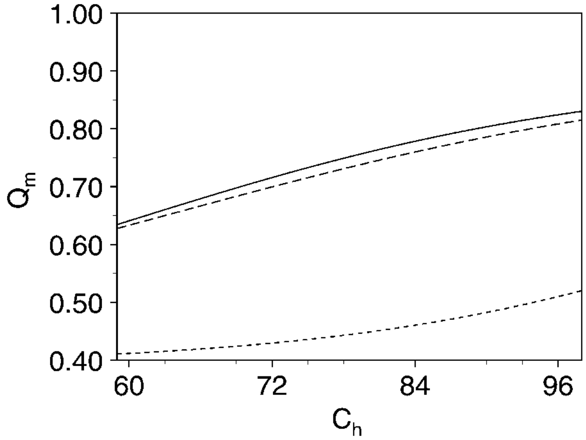

Fig. 18. Dependence of Q_m on C_h for helix sequences (second seed). Fixed values (Δ_c and $C_{as}{}^{bb}$) are set at their average helix values (solid). Fixed values are set at those of sequence 103, one of the fastest folding helix sequences (long dash). Fixed values are set at those of sequence 127, the slowest folding helix sequence (short dash).

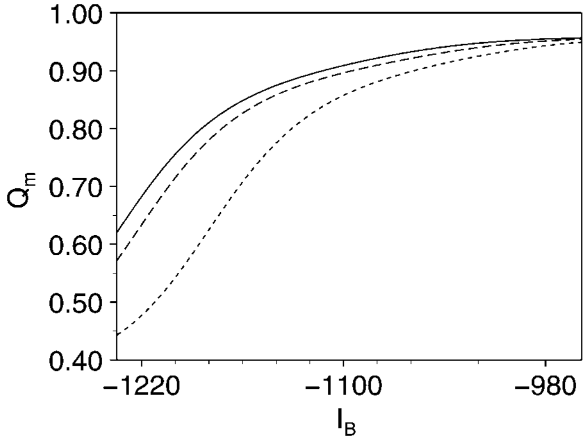

Fig. 19. Dependence of Q_m on I_B for all sequences (second seed). Fixed values (Δ_c and C_{as}) are set at their average values (solid). Fixed values are set at those of sequence 4, a sheet sequence which is the fastest folding sequence (long dash). Fixed values are set at those of sequence 127, a helix sequence which is the slowest folding sequence (short dash).

lated with ones that are picked). These include all the quantities that measure native structure not by secondary structure but by contact order $[C_0(i - j)]$ (including r_{Bij} $|_{i\ -\ j}|$). Since the primary difference between measures that use $|i - j|$ and those that use secondary structure is the amount of information about the spatial arrangement of the contacts, the difference in the GNN results suggests that the spatial arrangement of contacts is crucial. However, we should mention that one secondary structure measure is never picked: C_{ps}. Parallel sheet contacts are cooperative but difficult to access kinetically (they are typically long-ranged because the chain must loop around). The first of these is expected to aid folding and the second to hinder it. These opposing factors probably make the effects of parallel sheet contacts very context dependent, so that there is no predictivity.

DISCUSSION

In the present paper, we have extended the earlier analysis[5] of folding in the 125-mer by using a more sophisticated approach to determine the factors that lead to rapid folding. We generated a large and diverse database of engineered sequences and tested each for its ability to fold. A statistical method that combines a genetic algorithm with a neural network (GNN), employed previously in ligand design,[14,15] was used to select the most important sequence attributes for folding ability, and to determine the predictivity obtained by using the attributes. The method also makes it possible to use the database to derive the functional dependence of folding ability on the selected attributes. The conclusions are based only on comparisons of similarly engineered sequences so that the effect of any bias in the distribution of properties is removed. The present study and related ones[4–8,10–12] differ in an essential way from most laboratory experiments and simulations that consider only sequences that fold. Without explicit consideration of nonfolding or slowly folding sequences, it is impossible to distinguish properties that are crucial to folding from those that are simply common to all the sequences that were studied. The specific model that we studied was a 125-monomer chain subject to Monte Carlo dynamics on a simple cubic lattice. A sequence is defined by a set of assigned pairwise nearest-neighbor interactions. To examine the possible role of secondary structure in folding, sequences with substantial amounts of secondary structure in their $5 \times 5 \times 5$ native states were selected. The interactions were chosen such that the native state was known with high certainty and folding was observed for a significant fraction of the sequences within the allotted simulation time of 50×10^6 MC steps (45 helix and 87 sheet sequences found the native state in 4 or more of the 10 trials).

We have determined the folding mechanism of a few sequences in detail.[5] It is a three-stage process. The chain quickly collapses to a disordered globule due to the presence of an overall hydrophobic attraction. Then, it makes a relatively slow search through the compact states for a set of about 30 contacts, which make up a core that serves as the transition state for folding to a near-native intermediate. Although the folding core in our model forms by a random search restricted by stable, kinetically accessible, cooperative secondary structure, this step as a whole (core formation and the subsequent folding to a near-native state) is similar to the overall folding mechanism described for shorter (36-mer) optimized chains.[41,46] Our core differs in that it appears to be more tolerant of mutations; with only about 80% of its contacts, it can drive the chain to a near-native intermediate. Completion of folding requires condensation and often rearrangement of surface residues, which is slow due to the need to disrupt native structure. It should be noted that the intermediates reported for the shorter chains differ from ours in that they occurred before the formation of the nucleus.[41,47]

The generality of the 125-mer folding mechanism was demonstrated by statistical analysis of a database of 200 sequences. With a simple linear analysis, it was found that folding ability was significantly correlated with the stability of the native state and the amount of kinetically accessible cooperative secondary structure present in that state.[5] However, the overall predictivity (as measured by the single descriptor r_{cv} for prediction of Q_m by linear regression) achieved with either of these descriptors alone is less than 0.5.

The present study extends the analysis of the sequence database by correlating folding ability with multiple sequence attributes (descriptors) selected by a genetic algorithm and by deriving the functional dependence on those attributes with a neural network. The method was applied to the prediction of two different measures of folding ability (N_f, the number of times the native state was reached in 10 trials, and Q_m, the average contact overlap of the native state with the lowest energy state reached in each trial) for the helix and sheet sequences separately as well as all sequences together. The GNN-QSPR method combines the evolutionary and parallel attributes of genetic algorithms with the nonlinear features of neural networks. The resulting predictivity (in the range $0.434 < r_{cv} < 0.834$) is a significant improvement over the single-descriptor linear analysis (in the range $0.199 < r_{cv} < 0.492$). The utility of QSPR as a statistical analysis tool is made apparent by its success with the rather noisy folding ability measures available in this study due to the computational cost that limits the number of 125-mer Monte Carlo simulations for each sequence. The success of the GNN analysis in prediction is demonstrated by both the cross-validated Pearson linear correlation coefficients (r_{cv}) and the application of the QSPRs to a test set.

A physical interpretation of the chosen descriptors is of interest. In almost all cases, the GNN selects a measure of overall stability (Δ_c or E_0), a measure of kinetically accessible cooperative native secondary structure (C_{as}, $C_{as}{}^{bb}$, or C_t), and a measure of the stability of the contacts on the surface of the native structure ($E_0{}^{ss}$ or I_B). Increasing the energy gap stabilizes more native-like structures relative to less native-like ones, which increases the bias of the energy surface toward the native state (more "funnel"-like). As for the kinetic accessibility of the native secondary structure, structures that require long-range contacts in the absence of short-range contacts must surmount an entropic barrier between the disordered and native states. Thus, the first two properties are expected to make the core easier to find by the restricted random search and allow it to maintain its structure sufficiently long to drive folding to high Q_0. The need for less stable contacts on the surface of the native structure is kinetic. It avoids trapping in intermediates and favors the more open structures that can fold rapidly to the native state.

Although no detailed analysis of the type reported here has been made of other protein folding simulations, several lattice model studies have ascribed importance to some of the sequence attributes selected in the present work. As already mentioned, the importance of the overall stability of the global energy minimum (native state) was found in 27-mers with randomly assigned pairwise interactions.[4,6] In that case, the measures of stability used were the gap between the lowest and first excited compact state and the temperature at which an energetic order parameter ($X = 1–\Sigma_i\ p_i{}^2$ where p_i is the Boltzmann probability of conformation i and the sum is over all the 103,346 compact conformations) had a particular value [$T_x = T(X = 0.8)$]. For the fully compact native state of the 27-mer, these two quantities were shown to be closely correlated with the actual stability of the native state (which is a function of the entire energy spectrum). The correlation between folding ability and the energy gap was explained by a nonspecific folding mechanism that consists of a fast collapse to a semicompact random globule, followed by a slow, nondirected search through the semi-compact structures for one of the transition states that lead rapidly to the native conformation.[6] The energy gap criterion results in a native state that is stable at a temperature high enough for the folding polypeptide chain to overcome barriers between random semicompact states. The importance of the native state stability has been confirmed by others[7,8] and is in good agreement with analytical heteropolymer theories.[16,48–50]

The role of the structure of the native state for folding of the 125-mer is apparent even without the GNN-QSPR, in that the two secondary structure classes studied differ in the distributions of folding abilities. Helix proteins tend to fail at core formation whereas sheet proteins tend to fail at rearrangement from the intermediate to the native state. This difference can be understood in terms of the cooperativity of helix and sheet contacts. Cooperativity makes it harder to break structure. Helix contacts are primarily noncooperative and thus helix sequences have difficulty maintaining the core long enough to drive further condensation of structure. Sheet contacts are cooperative and thus sheet sequences have difficulty breaking native structure in the intermediate to allow rearrangement.

No dependence on the secondary structure was observed in the random 27-mers.[4] For optimized sequences, the need for at least some cooperative structure was demonstrated using three 36-mer sequences[38]; two sequences with some long-range, cooperative contacts were found to exhibit the all-or-none behavior associated with folding, whereas one with almost only short-range, noncooperative contacts spent significant amounts of time in partially folded states. Exhaustive enumeration of shorter chains leads to similar conclusions.[51,52] It is important to note that the helix sequences of the present study, in contrast to the one used in Abkevich et al.,[38] do contain substantial numbers of long-range contacts that yield cooperativity and all-or-none kinetics. Although Abkevich et al.[38] are correct in pointing out that cooperativity is crucial to a model's ability to represent real proteins,[38,53] their use of only a few "hand-built" sequences does not provide generality and an unbiased test, whereas their use of short chains obscures the importance of kinetic accessibility.

Kinetic accessibility is one of the central elements of the "hydrophobic zipper" (HZ) hypothesis,[54,55] in which it is suggested that the chain searches only those contacts that are "topologically local" (either short-range or effectively short-range due to the constraints of other contacts). The 125-mer simulations are the first to find a folding mechanism that employs an HZ-like search strategy in a study of random sequences with chains long enough for kinetic accessibility to be an issue. However, we again stress that specific long-range contacts are crucial for an all-or-none folding transition. This essential factor was neglected in one recent study[10] that reported fast folding for 27-mer sequences designed to have strong short-range contacts without mandating a certain fraction of long-range contacts; such sequences model something more like helix-coil transition than protein folding.

The third criterion for fast folding, which concerns the stability of the native surface structure, has not been noted previously. As discussed above, weaker contacts between noncore monomers break more easily and thus allow faster rearrangement from the intermediate states, which typically have native-like local (secondary) structure but incorrect global (tertiary) structure. Since other studies of long chain lattice models have reported trapping in similar intermediates, it is possible that the importance of the surface structure stability was overlooked because each of those studies considered only one sequence. These models include an early simple cubic lattice representation of lysozyme with 116 "backbone" monomers and 15 "side chains" subject to Go-type interactions[56] and an 80-mer subject to HP-type interactions.[57] Given that such intermediates do not appear to be important in shorter chains, it is likely that the stability of the surface residues plays a role only for larger systems with an interior core.

One factor that was suggested to be important for folding[39] but was not found to be so in the present study is the heterogeneity of the native interactions. For 48-mers designed to have large energy gaps (Δ/σ_B) with varying ratios of $\sigma_B{}^{nat}/\sigma_B$,[39] it was found that sequences with less heterogeneous native interactions (smaller $\sigma_B{}^{nat}/\sigma_B$) were less likely to populate intermediates and thus fold faster to the native state. Although that study considered a larger range of $\sigma_B{}^{nat}/\sigma_B$ (0.27–1.77) than did ours (0.82–1.06), the QSPR methodology should be sufficiently sensitive to detect any significant dependence of the folding ability on this parameter. The fact that it was not found indicates that it is less important than $E_0{}^{ss}$ in the present system.

A factor that has been suggested to be of importance by Klimov and Thirumalai[11,12] but was not tested in the present study is the quantity $\sigma = 1 - T_f/T_\theta$, where T_f is the folding transition temperature and T_θ is the collapse transition temperature. It has been argued that σ is simply a measure of the width of the folding transition, which is directly related to the stability of the native state or the energy gap.[53] We did not include σ in the present study because, unlike all the descriptors that were used, it cannot be determined without long MC simulations and thus is available only a posteriori (i.e., it has no predictive quality). One reason that Klimov and Thirumalai used σ was that it is dimensionless, which these authors argued is necessary for comparing a quantity with the folding ability.[11,12] Thus, to compare σ with the energy gap between the ground and first excited compact states, they normalized the energy gap (between the ground and first excited states) by the simulation temperature (T_s), which was a temperature at which a structural order parameter had a particular value. Since the energy gap and T_s are highly correlated, their ratio ($\Delta E_{10}/T_s$) is approximately the same for all sequences, so it is not surprising that the correlation between folding ability and stability was obscured.

In comparing the present study with others, we stress that a number of them, in particular Šali et al.[4,6] and Klimov and Thirumalai,[11,12] used a different simulation temperature for each sequence. The temperature chosen was close to T_m. This provides a physically meaningful approach since it is appropriate to compare the folding of sequences under conditions of corresponding stability. Moreover, a real protein must not only find but also populate its native state, so that sequences with small gaps (low stability) which find their native states by chance in high temperature simulations are not very meaningful. In the present study, we used a single temperature for the entire sequence database. This is valid because the sequences have similar stabilities due to the optimization procedure. A simulation temperature of $T = 0.8$ was used; lengthy equilibrium sampling of a few sequences with varying folding abilities suggested that T_m is in the range 0.9–1.1 for almost all (if not all) the sequences in the database. This T_m estimate, which is based on more extensive simulations, is somewhat higher than the preliminary one ($T_m \approx 0.8$) given in Dinner et al.[5]. Thus there was no need to vary the folding temperature. It should be mentioned that computational costs prohibit determination of T_m for all 200 of the 125-mer sequences. That the stability of the native state is still selected as one of the prime determinants of folding ability demonstrates that its importance is not simply a temperature (Arrhenius) effect. The latter was suggested in a paper[10] that argued incorrectly, as we will show (to be published), that the observed distribution in folding abilities of the 27-mer sequences was due entirely to the distribution in temperature.

A better understanding of what designed properties speed folding should give hints at the transition states, and hence the overall mechanism, of the folding reaction. The design conclusions we draw from the present study are that the fastest folding sequences are those that are optimized both energetically and entropically (structurally). Optimization of sequence attributes that are closely related to Δ_c has been carried out by Shakhnovich and Gutin[21] and Shakhnovich[40]. The present study suggests that folding can be further accelerated by designing the native structure so that highly cooperative contacts are accessible by short-range interactions and by weakening the surface contacts. However, since $E_0{}^{ss}$ is of substantially less predictive value than Δ_c, some weighting of the optimization function would be necessary.

As already discussed by Dinner et al.,[5] care must be used in translating the results of lattice models to real proteins. Each amino acid is represented by a single point, the potential is highly simplified, and the lattice space is extremely restrictive. Also, the choice of a local Monte Carlo move set imposes dynamics that do not include large-scale diffusive motions. Nevertheless, such models are useful in understanding certain generic features of protein folding. It is therefore important to make a comparison of the present results with the available experimental data.

The role of initiation sites and stable cooperative structure is in agreement with the statistics for databases of known protein structures. Brocchieri and Karlin,[58] who studied a database of 172 nonhomologous well-resolved protein structures (52,760 residues in total), classified spatially adjacent residues by sequence separation. In agreement with the conclusion that long-range cooperative structure helps restrict the search for the core, they found that buried amino acids are most commonly separated in the primary sequence by more than 50 residues. Exposed amino acids are most commonly within four residues of each other, in accord with the slowing down of the final stages of folding for structures with more cooperative, longer range surface contacts. The largest number of strand-strand contacts had sequence separations of between 5 and 20 residues, as expected for two strands separated by a tight turn. A breakdown of the statistics by the secondary structure class of the native state was not given, so it is not possible to compare the statistical results directly with our results for the helix and sheet sequence groups considered separately.

Recent submillisecond experiments indicate that helices form about 30 times more quickly than do loops and hairpins.[59,60] These estimates suggest that the fastest α-helical proteins can fold more rapidly than the fastest β-sheet proteins, although fast-folding β-barrel proteins do exist.[61] The distribution of folding abilities observed for the 125-mer and other lattice simulations[38] indicates that in these models helix-dominated structures fold more slowly than β-sheet structures. The failure to reproduce the fast folding of α-helical proteins is likely to be a consequence of the absence of large-scale diffusive motions[62] combined with a lack of cooperativity in lattice helices. Given this limitation, lattice secondary structures should be viewed as modeling cooperativity and kinetic accessibility rather than as a realistic representation of α-helices and β-sheets.

Several different experimental approaches have employed comparative strategies to address issues in folding and design. Like their ligand design counterparts,[15] such studies typically focus on comparisons between molecules with similar overall three-dimensional structure so that small differences are more readily interpretable. The most straightforward experimental approaches are those that exploit the libraries provided by nature, families of homologous proteins. Using fluorescence, circular dichroism, and pulsed hydrogen exchange in conjunction with nuclear magnetic resonance and mass spectrometry, the folding kinetics of members of the lysozyme family from hen, human (52 residues different from hen), Lady Amhurst pheasant (6 residues different), and Japanese quail (6 residues different) have been studied.[63–65] For all these variants, it was observed that there is highly cooperative folding of the α-helical domain to an intermediate lacking substantial protection in the β-sheet domain. Furthermore, it was found that folding proceeded by fast and slow parallel pathways (like the 125-mer).[5] However, clear differences between the variants do exist; the quail protein folds to the intermediate more slowly than either the hen or pheasant proteins, which in turn fold roughly fourfold more slowly than does the human protein. It was suggested that the main difference, at least in the case of hen and human lysozymes,[64] arose from the packing of the hydrophobic core of the α-domain and not the helical propensities of the sequences. These results are consistent with those of the present study in which we find that the stability of the core is the most important factor in determining folding ability.

A much larger number of sequences for comparison can be generated by the use of mutagenesis. Extensive site-directed (and more recently random [66]) mutagenesis for the purpose of studying the effects on folding kinetics has been carried out by Fersht[67,68] and co-workers (the “protein engineering” method). The two proteins studied by that group were the chymotrypsin inhibitor 2 (CI2) and the ribonuclease barnase. For both proteins, it has been suggested that the folding transition states contain substantial native-like structure in the principal hydrophobic core but little elsewhere. The folding mechanism of the 125-mer model more closely resembles that of barnase, which has an intermediate, than that of CI2, which folds directly to the native state (like the optimized 36-mers discussed above[41,46]).

Several studies that used site-directed mutagenesis to design stable sequences found that the mutations influenced the kinetics of folding (and unfolding). For example, mutations to alanines and leucines to repack the hydrophobic core of Rop, a dimeric four-helix bundle, increased the folding and unfolding rates by factors of more than 10,000 relative to the wild type.[69] In addition to the effects on packing in the core (rate-limiting step), the speed-up in folding could derive from a change in the helix propensities. Site-directed mutagenesis studies of the relation of helix propensities to folding rate have been carried out with Che Y, the activation domain of procarboxypeptidase A, an SH3 domain,[70] and a λ-repressor construct.[71] For the λ-repressor construct, it was found that folding could either be accelerated or decelerated by changes in the helix propensities, depending on the structure of the transition state (based on the fraction of solvent-accessible surface area relative to the denatured state). In viewing the protein engineering studies as a whole, the stability and cooperativity of the hydrophobic core are clearly of importance in all cases, whereas the effects of local interactions depend on the rest of the structure. These findings are in agreement with the present study: for all sequence groups (all, helix, and sheet) folding ability depends strongly on the stability of the core, but the QSPRs between folding ability and the number of helix contacts change drastically as we go from all sequences (for which it hinders folding) to helical sequences (for which it helps folding).

Combinatorial strategies, in which many random sequences are compared can be partially realized with random mutagenesis experiments.[2,3] In such an experiment, a library of mutants is created using cassette mutagenesis or polymerase chain reaction to mutagenize a set of positions randomly in the gene. The library is transformed into cells, and functional proteins are identified by screening for activity, expression, or antibody cross-reactivity, or by a freezing-thaw protocol.[72] This method has been applied to the N-terminal domain of λ-repressor, the phage P22 Arc repressor, the GCN4 dimer, cytochrome b_{562}, and a designed four-helix bundle (see ref. 3 and references therein). The random mutagenesis studies are consistent with the site-directed ones, in particular, in regard to their emphasis on the importance of packing of the hydrophobic core. Experimental studies that carry the combinatorial philosophy to its logical extreme and generate completely random sequences have only just begun to be performed.[73,74] It was found that about 1% of sequences composed of leucine, glutamine, and arginine were resistant to intracellular proteolysis, but none of these exhibited significant amide protection, indicating that their structures were molten globule-like and lacked the rigidity of a true protein. Such experiments have the potential to yield interesting and even unexpected results when more data become available and a thorough analysis of the structures resistant to proteolysis can be carried out.

The present study has important methodological implications. GNN-QSPR techniques (including three-dimensional QSPR methods[75]) could be applied to the prediction of native state stabilities (melting temperatures) and folding rates of real proteins. Such an automatic approach could not only better quantify the existing "rules-of-thumb"[76] but also bring out features previously ignored (e.g., consideration of E_0^{ss}, the energy of the contacts between surface monomers, found to be important in the present study). Such analysis would be particularly useful in experiments with random sequences, ideally ones in which both folding and nonfolding sequences were considered. A first step in this direction has been taken recently by Dahiyat and Mayo, who designed several sequences for the coiled coil GCN4-p1[77] and the streptococcal protein G β1 domain[78] and subsequently improved their sequence scoring function by a QSPR analysis based on a genetic algorithm and linear regression. The use of GNN-QSPR for protein design is in the spirit of other automated design procedures[79] and knowledge-based protein structure prediction (for reviews, see refs. 32, 42, 80, and 81). However, it differs in that the quantities to be predicted would be stability and folding rate, not structure. Prediction of these quantities may be easier than structure prediction and useful for protein design.

ACKNOWLEDGMENTS

We thank Themis Lazaridis for a critical reading of the manuscript. Simulations were done on an Iris 4D, an IBM RS6000–550, DEC Alphas, a DECstation 5000/240, HP Apollo 735s, and a Silicon Graphics R5000 Indy. A.R.D. is a Howard Hughes Medical Institute Pre-Doctoral Fellow.

REFERENCES

1. Karplus, M., Shakhnovich, E. Theoretical studies of thermodynamics and dynamics. In: "Protein Folding." Creighton, T.E. (ed.). New York: W.H. Freeman, 1992:127–193.
2. Cordes, M.H.J., Davidson, A.R., Sauer, R.T. Sequence space, folding and protein design. Curr. Opin. Struct. Biol. 6:3–10, 1996.
3. Sauer, R.T. Protein folding from a combinatorial perspective. Folding Design 1:R27–R30, 1996.
4. 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.
5. Dinner, A.R., Săali, A., Karplus, M. The folding mechanism of larger model proteins: Role of native structure. Proc. Natl. Acad. Sci. USA 93:8356–8361, 1996.
6. Šali, A., Shakhnovich, E., Karplus, M. How does a protein fold? Nature 369:248–251, 1994.
7. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100: 9238–9257, 1994.
8. Socci, N.D., Onuchic, J.N. Folding kinetics of proteinlike heteropolymers. J. Chem. Phys. 101:1519–1528, 1994.
9. 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.
10. Unger, R., Moult, J. Local interactions dominate folding in a simple protein model. J. Mol. Biol. 259:988–994, 1996.
11. Klimov, D.K., Thirumalai, D. Factors governing the foldability of proteins. Proteins 26:411–441, 1996.
12. Klimov, D.K., Thirumalai, D. Criterion that determines the foldability of proteins. Phys. Rev. Lett. 76:4070–4073, 1996.
13. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H. Protein design by binary patterning of nonpolar amino acids. Science 262:1680–1685, 1993.
14. So, S.-S., Karplus, M. Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. J. Med. Chem. 39:1521–1530, 1996.
15. So, S.-S., Karplus, M. Genetic neural networks for quantitative structure-activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/GABA$_A$ receptors. J. Med. Chem. 39:5246–5256, 1996.
16. Shakhnovich, E.I., Gutin, A.M. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346:773–775, 1990.
17. Chan, H.S., Dill, K.A. The effects of internal constraints on the configurations of chain molecules. J. Chem. Phys. 92:3118–3135, 1990.
18. Dinner, A.R., Karplus, M. A metastable state in folding simulations of a protein model. Nature Struct. Biol. 5:236–241, 1998.

19. Chotia, C., Levitt, M., Richardson, D. Structure of proteins: Packing of α-helices and pleated sheets. Proc. Natl. Acad. Sci. USA 74:4130–4134, 1977.
20. Gõ, N., Abe, H. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. Biopolymers 20:1013–1031, 1981.
21. 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.
22. 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.
23. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.
24. Hilhorst, H.J., Deutch, J.M. Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. J. Chem. Phys. 63:5153–5161, 1975.
25. Madras, N., Sokal, A.D. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. J. Stat. Phys. 50:109–186, 1988.
26. Karplus, M., Caflisch, A., Săali, A., Shakhnovich, E. Protein dynamics: From the native to the unfolded state and back again. In: “Modelling of Biomolecular Structures and Mechanisms.” Pullman, A., Jortner, J., Pullman, B. (eds.). Boston: Kluwer Academic Publishers, 1995:69–84.
27. In ref. 5, the value of B_0 was reported incorrectly and should have been $B_0 = –0.25/0.6 = –0.42$ in the scaled units.
28. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I. Is burst hydrophobic collapse necessary for protein folding? Biochemistry 34:3066–3076, 1995.
29. Holland, J.H. “Adaption in Natural and Artificial Systems.” Ann Arbor, MI: The University of Michigan Press, 1975.
30. Hertz, J., Krogh, A., Palmer, R.G. “Introduction to the Theory of Neural Computation.” Redwood City, CA: Addison-Wesley, 1991.
31. Zupan, J., Gasteiger, J. “Neural Networks for Chemists: An Introduction.” New York: VCH Publishers, 1993.
32. Pedersen, J.T., Moult, J. Genetic algorithms for protein structure prediction. Curr. Opin. Struct. Biol. 6:227–231, 1996.
33. Clark, D.E., Westhead, D.R. Evolutionary algorithms in computer-aided molecular design. J. Comput. Aided. Mol. Des. 10:337–358, 1996.
34. Burns, J.A., Whitesides, G.M. Feed-forward neural networks in chemistry: Mathematical systems for classification and pattern recognition. Chem. Rev. 93:2583–2602, 1993.
35. Luke, B.T. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure-property relationships. J. Chem. Inf. Comput. Sci. 34:1279–1287, 1994.
36. Møller, M.F. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6:525–533, 1993.
37. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
38. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J. Mol. Biol. 252:460–471, 1995.
39. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Improved design of stable and fast-folding model proteins. Folding Des. 1:221–230, 1996.
40. Shakhnovich, E.I. Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. 72:3907–3910, 1994.
41. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. J. Chem. Phys. 101:6052–6062, 1994.
42. Sippl, M.J. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5:229–235, 1995.
43. 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.
44. Chan, H.S., Dill, K.A. Compact polymers. Macromolecules 22:4559–4573, 1989.
45. Dinner, A., Šali, A., Karplus, M., Shakhnovich, E. Phase diagram of a model protein derived by exhaustive enumeration of the conformations. J. Chem. Phys. 101:1444–1451, 1994.
46. 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.
47. Mirny, L.A., Abkevich, V.I., Shakhnovich, E.I. Universality and diversity of the protein folding scenarios: A comprehensive analysis with the aid of a lattice model. Folding Des. 1:103–116, 1996.
48. Bryngelson, J.D., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. USA 84:7524–7528, 1987.
49. Bryngelson, J.D., Wolynes, P.G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. 93:6902–6915, 1989.
50. Shakhnovich, E.I., Gutin, A.M. Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach. Biophys. Chem. 34:187–199, 1989.
51. Govindarajan, S., Goldstein, R.A. Searching for foldable protein structures using optimized energy functions. Biopolymers 36:43–51, 1995.
52. Govindarajan, S., Goldstein, R.A. Local propensities for model proteins. Proteins 22:413–418, 1995.
53. Shakhnovich, E.I. Theoretical studies of protein-folding thermodynamics and kinetics. Curr. Opin. Struct. Biol. 7:29–40, 1997.
54. Fiebig, K.M., Dill, K.A. Protein core assembly processes. J. Chem. Phys. 98:3475–3487, 1993.
55. Dill, K.A., Fiebig, K.M., Chan, H.S. Cooperativity in protein-folding kinetics. Proc. Natl. Acad. Sci. USA 90:1942–1946, 1993.
56. Ueda, Y., Taketomi, H., Gõ, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A three-dimensional lattice model of lysozyme. Biopolymers 17:1531–1548, 1978.
57. 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.
58. Brocchieri, L., Karlin, S. How are close residues of protein structures distributed in primary sequence? Proc. Natl. Acad. Sci. USA 92:12136–12140, 1995.
59. Eaton, W.A., Muñoz, W., Thompson, P.A., Chan, C.-K., Hofrichter, J. Submillisecond kinetics of protein folding. Curr. Opin. Struct. Biol. 7:10–14, 1997.
60. Muñoz, V., Thompson, P.A., Hofrichter, J. Eaton, W.A. Folding dynamics and mechanism of β-hairpin formation. Nature 390:196–199, 1997.
61. Schindler, T., Schmid, F.X. Thermodynamic properties of an extremely rapid protein folding reaction. Biochemistry 35:16833–16842, 1996.
62. Karplus, M., Weaver, D.L. Protein folding dynamics: The diffusion-collision model and experimental data. Protein Sci. 3:650–668, 1994.
63. Radford, S.E., Dobson, C.M., Evans, P.A. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. Nature 358:302–307, 1992.
64. Hooke, S.D., Radford, S.E., Dobson, C.M. The refolding of human lysozyme: A comparison with the structurally homologous hen lysozyme. Biochemistry 33:5867–5876, 1994.
65. Hooke, S.D., Eyles, S.J., Miranker, A., Radford, S.E., Robinson, C.V., Dobson, C.M. Cooperative elements in protein folding monitored by electrospray ionization mass spectrometry. J. Am. Chem. Soc. 117:7548–7549, 1995.
66. Axe, D.D., Foster, N.W., Fersht, A. Active barnase variants with completely random hydrophobic cores. Proc. Natl. Acad. Sci. USA 93:5590–5594, 1996.
67. Fersht, A.R. Protein folding and stability: The pathway of folding of barnase. FEBS Lett. 325:5–16, 1993.

68. Fersht, A.R. Characterizing transition states in protein folding: An essential step in the puzzle. Curr. Opin. Struct. Biol. 5:79–84, 1995.
69. Munson, M., Anderson, K.S., Regan, L. Speeding up protein folding: Mutations that increase the rate at which Rop folds and unfolds by over four orders of magnitude. Folding Des. 2:77–87, 1997.
70. Muñoz, V., Serrano, L. Local versus nonlocal interactions in protein folding and stability—an experimentalist's point of view. Folding Des. 1:R71–R77, 1996.
71. Burton, R.E., Huang, B.S., Daugherty, M.A., Calderone, T.L., Oas, T.G. The energy landscape of a fast-folding protein mapped by ala·gly substitutions. Nature Struct. Biol. 4:305–310, 1997.
72. Roy, S., Helmer, K.J., Hecht, M.H. Detecting native-like properties in combinatorial libraries of de novo proteins. Folding Des. 2:89–92, 1996.
73. Davidson, A.R., Sauer, R.T. Folded proteins occur frequently in libraries of random amino acid sequences. Proc. Natl. Acad. Sci. USA 91:2146–2150, 1994.
74. Davidson, A.R., Lumb, K.J., Sauer, R.T. Cooperatively folded proteins in random sequence libraries. Nature Struct. Biol. 2:856–863, 1996.
75. Kubinyi, H. "3D QSAR in Drug Design: Theory, Methods and Applications." Leiden: ESCOM Science Publishers, 1993.
76. Shaw, A., Bott, R. Engineering enzymes for stability. Curr. Opin. Struct. Biol. 6:546–550, 1996.
77. Dahiyat, B.I., Mayo, S.L. Protein design automation. Protein Sci. 5:895–903, 1996.
78. Dahiyat, B.I., Mayo, S.L. Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. USA 94:10172–10177, 1997.
79. Desjarlais, J.R., Handel, T.M. De novo design of the hydrophobic cores of proteins. Protein Sci. 4:2006–2018, 1995.
80. Barton, G.J. Protein secondary structure prediction. Curr. Opin. Struct. Biol. 5:372–376, 1995.
81. Bohm, G. New approaches in molecular structure prediction. Biophys. Chem. 59:1–32, 1996.