Sequence Evolution and the Mechanism of Protein Folding

Angel R. Ortiz and Jeffrey Skolnick
Department of Molecular Biology, TPC-5, The Scripps Research Institute, La Jolla, California 92037 USA

ABSTRACT The impact on protein evolution of the physical laws that govern folding remains obscure. Here, by analyzing in silico-evolved sequences subjected to evolutionary pressure for fast folding, it is shown that: First, a subset of residues in the thermodynamic folding nucleus is mainly responsible for modulating the protein folding rate. Second and most important, the protein topology itself is of paramount importance in determining the location of these residues in the structure. Further stabilization of the interactions in this nucleus leads to fast folding sequences. Third, these nucleation points restrict the sequence space available to the protein during evolution. Correlated mutations between positions around these hot spots arise in a statistically significant manner, and most involve contacting residues. When a similar analysis is carried out on real proteins, qualitatively similar results are obtained.

INTRODUCTION

Prediction of protein structure from sequence is widely recognized as one of the most important unsolved problems in molecular biology, and it has become more pressing with the current blossoming of genome sequencing projects (Koonin, 1997, Koonin, et al., 1998). Many of the current approaches to protein structure prediction rely on relating sequence patterns to structural ones, particularly with the availability of multiple sequence alignments (MSAs) for many sequences of interest. But finding sequence–structure relationships requires discrimination from functional or even adventitious patterns in current databases that may not have unique correspondence with structure. Signatures related to protein topology are the needles to be found in the haystack of alignments. Also, because topology is the outcome of the folding mechanism, relationships among sequence space, the folding process, and final topology need to be established. In summary, a better understanding of the constraints imposed during protein evolution by the physical laws that govern folding is required.

The acquisition of such understanding is a formidable task. To make the problem tractable, simplified lattice models of protein-like chains have been developed by a number of authors (Shakhnovich, 1996). In particular, Shakhnovich and coworkers (Mirny, et al., 1998) recently generated a database of fast and slow folding models of homologous proteins (48-mers on a cubic lattice, see Methods and Fig. 2) by using evolution-like pressure for fast folding. Here, after analyzing these fast and slow folding model proteins and after doing similar calculations on real proteins, we report that structurally significant patterns can appear in MSAs of evolutionary related sequences. We observe that correlated mutations tend to occur between contacting residues, and that they seem to arise as a consequence of the evolutionary constraint to fold sufficiently fast to the global energy minimum. They tend to be located closer than expected by chance to the residues forming the protein folding nucleus. These correlations could also be topology dependent, because we show that the topology itself determines which residues form part of the folding nucleus.

METHODS

Database of model proteins

A database formed by in silico evolutionary related sequences of slow and fast folding model proteins was generously provided to us by Dr. E. Shakhnovich (personal communication). This database consists of ~1000 sequences of model proteins created by folding randomly mutated sequences of 48-mers on a cubic lattice subjected to evolutionary pressure for fast folding (Mirny, et al., 1998). The conformation to which all these sequences fold is shown in Fig. 2.

Multivariate analysis of the sequences of model proteins

Factor Analysis (Reyment and Joreskog, 1996) (FA) has been used to derive statistical models that explain the differences between fast and slow folding sequences. FA tries to describe the covariance relationships in a data matrix in terms of underlying, but unobservable, random quantities known as factors. The factors are new variables, or directions in space, built from linear combinations of the original variables in such a way that they account approximately for the same amount of information of the original variables in the data matrix. This is achieved by grouping variables by their correlations. Mathematically speaking, FA seeks finding an underlying orthogonal factor model of an original **X** matrix of the form

$$\mathbf{X} = \mathbf{L}\mathbf{F} + \mathbf{E}. \tag{1}$$

In Eq. 1, **X** is a matrix derived from the MSA and having n rows, corresponding to n positions in the alignment; and m columns, corresponding to m sequences being aligned. On the right-hand side, **L** is the loadings matrix and **F** the scores matrix, whereas **E** is the residual (noise) matrix. Each element l_{ip} of the matrix **L** can be considered as the square root of the weight (or importance) of the variable i on factor p. That is, it is the contribution of variable (i.e., the position) i on building the new direction p. In contrast each component score f_{pj} from the matrix **F** can be considered the projection of the object (i.e., the sequence) j on factor p, or, in other words, the coordinate of object j in the new direction given by factor p.

Received for publication 18 June 1999 and in final form 24 February 2000.

Address reprint requests to Dr. Angel Ramirez Ortiz, Assistant Professor, Department of Physiology and Biophysics, Mount Sinai School of Medicine, One Gustave Levy Pl., Box 1218, New York, NY 10029.

0006-3495/00/10/1787/13 $2.00

The existence of this model implies a special covariance structure for **X.** It can be shown that, under the assumption that **F** and **E** are orthogonal and independent, this covariance structure **XX′** can be expressed as

$$\mathbf{XX'} = (\mathbf{LF} + \mathbf{E})(\mathbf{LF} + \mathbf{E})', \tag{2}$$

$$\mathbf{XX'} = \mathbf{S} = \mathbf{LL'} + \mathbf{U}. \tag{3}$$

Here **X′** is the transpose of **X,** and **U = EE′** is the covariance of the noise matrix, which, under the factor model conditions, contains the specific (not interrelated and therefore not explained by the model) variances of the objects, assumed to be small: $\mathbf{U} \cong \mathbf{0}$**.** The solution to the factor model can then be obtained by diagonalization of the covariance structure **XX′.** This is known as the *principal component solution* to the factor model (Johnson and Wichern, 1992). By using the Eckart–Young theorem (Reyment and Joreskog, 1996), **S** can be expressed as

$$\mathbf{S} = \mathbf{P\Lambda P'} = \mathbf{P\Lambda}^{1/2}\mathbf{\Lambda}^{1/2\prime}\mathbf{P'} = (\mathbf{P\Lambda}^{1/2})(\mathbf{P\Lambda}^{1/2})'. \tag{4}$$

Thus, by comparing Eqs. 3 and 4, we can obtain the loadings as

$$\mathbf{L} = \mathbf{P\Lambda^{1/2}}. \tag{5}$$

Equation 5 indicates that the loadings are no more than the scaled eigenvectors of the symmetric matrix **S.** If principal components analysis is used as a solution of the factor model using Eq. 5 to obtain the loadings, then it is customary to generate the scores by an ordinary (unweighted) least squares procedure (Johnson and Wichern, 1992). Starting from Eq. 1, and assuming $\mathbf{E} \cong \mathbf{0}$**,** then

$$\mathbf{L'X} = \mathbf{L'LF}, \tag{6}$$

$$\mathbf{F} = (\mathbf{L'L})^{-1}\mathbf{L'X}. \tag{7}$$

Using Eqs. 3, 4, and 5, Eq. 7 can be expressed as

$$\mathbf{F} = (\mathbf{\Lambda^{1/2\prime}P'P\Lambda}^{1/2})^{-1}(\mathbf{P\Lambda}^{1/2})'\mathbf{X}, \tag{8}$$

$$\mathbf{F} = \mathbf{\Lambda}^{-1/2}\mathbf{P'X}. \tag{9}$$

Thus, the factor model can be solved using Eqs. 5 and 9 after diagonalizing the covariance structure **XX′.** After a lower dimensionality space of p factors explaining most of the variance of the original **X**-matrix is found, in FA, one proceeds to rotate the factors until a simpler structure is achieved (Reyment and Joreskog, 1996). This is done to obtain a better interpretation of the factors, and it is possible because Eq. 3 is insensitive to any orthogonal transformation of the loadings,

$$\mathbf{XX'} = \mathbf{S} = \mathbf{LL'} + \mathbf{U} = \mathbf{LTT'L'} + \mathbf{U} \quad \text{where} \quad \mathbf{TT'} = \mathbf{I}. \tag{10}$$

The ideal rotation matrix **T** to apply to **L** would generate a new loadings matrix so that each variable has the simplest interpretation in the new space. This would be achieved by having in the rotated loadings matrix **L*** as many zero elements as possible. In this way, a variable would not depend on all common factors but only on a small part of them. This is equivalent to maximizing the total variance of the loading elements in the p selected factors. A popular way to obtain such a multidimensional rotation is by means of the *varimax* rotation, where an iterative pairwise bidimensional rotation method due to Kaiser (1958) is used to find an orthogonal optimal rotation matrix **G** that maximizes a measure of the variances of the loadings known as the Harman function, ϕ:

$$\phi = \sum_{j=1}^{n}\sum_{i=1}^{p}(d_{ij}^2 - \bar{d}_j)^2 = \sum_{j=1}^{n}\sum_{i=1}^{p} d_{ij}^4 - p\sum_{j=1}^{n}\bar{d}_j^2, \tag{11}$$

where

$$d_{ij} = \frac{l_{ij}}{h_i} \tag{12}$$

and

$$\bar{d}_j = p^{-1}\sum_{i=1}^{p} d_{ij}^2. \tag{13}$$

Here, l_{ij} is the *ij* element of the loadings matrix **L,** h_i is the square root of the communality of the variable *i* (that is, the variance explained by the factor model), *n* equals the total number of variables, and *p* is the dimensionality of the loadings matrix. The rotation matrix **G** that maximizes Eq. 11 is used to obtain the rotated loadings by using the transformation,

$$\mathbf{L^*} = \mathbf{LG}. \tag{14}$$

The FA of the MSA was carried out using the correlation matrix, instead of the covariance matrix, as follows: The correlation matrix **R** was derived by first defining an exchange matrix at each sequence position in the alignment where each pair of elements in the position is scored by a modified version of the McLachlan matrix [optimized to maximize the contact prediction ability, (Ortiz and Skolnick, in preparation)], and then computing a Pearson-type correlation coefficient between exchange matrices at any two positions. Thus, the element r_{ij} of **R** is obtained,

$$r_{ij} = \frac{1}{N^2}\sum_{kl}\frac{(s_{ikl} - \langle s_i\rangle)(s_{jkl} - \langle s_j\rangle)}{\sigma_i\sigma_j}. \tag{15}$$

Where *i* and *j* are two different positions in the MSA, and the indices *k* and *l* run from 1 to the *N* of sequences in the family. The parameter s_{ikl} (s_{jkl}) is the comparison score of the amino acids of sequences *k* and *l* at position *i* (*j*) of the alignment. Average values over all aligned sequences at positions *i* and *j* are given by $\langle s_i\rangle$ *and* $\langle s_j\rangle$, whereas σ_i and σ_j correspond to their standard deviation. Afterwards, the PCA solution to the R-mode FA model was obtained by diagonalizating **R.** The first $p = 3$ dimensions were scaled and subjected to a varimax rotation to obtain the loadings matrix. We have seen by trial and error that three factors contain enough variance to explore the main differences in the alignments without being anchored in the rotation by the less significant factors so that they provide optimal separation of residues and sequences, and, therefore, the clearest interpretation of the FA model. From the rotated loadings, the scores were computed by ordinary least squares.

Prediction of contacts and correlated mutation analysis

The procedure is based on computing the **R** matrix from the MSA, as defined above, followed by the sequential application of FA (Reyment and Joreskog, 1996) to eliminate phylogenetic relationships and partial correlation (Johnson and Wichern, 1992) to eliminate indirect effects of intervening or indirect variables. Typically, and for proteins up to about 100 residues, between 5 and 10 contacts are selected. A more detailed account of this methodology for contact prediction can be found in Ortiz, et al. (1999), and an extensive evaluation for real proteins will be given in a subsequent publication. For the proteins studied in this paper, the MSAs were obtained from the HSSP database (Sander and Schneider, 1991) for those proteins present in the protein databank (Bernstein, et al., 1977). As for the CASP3 proteins, a detailed account of the creation of the corresponding MSAs is presented in a separate publication (Ortiz, et al., 1999).

Identification of kinetically hot residues by the GNM

The basic idea behind the Gaussian Network Model (GNM) (Bahar, et al., 1997) is to consider the folded protein as a three-dimensional (3D) elastic network where residues undergo Gaussian fluctuations around their mean positions resulting from harmonic potentials between residues in contact. The Kirchhoff matrix of contacts Γ describes the dynamic characteristics of this network. This matrix is the counterpart of the stiffness matrix used in the analysis of elastic bodies, and its *ij* element is defined as

$$\Gamma_{\mathrm{ij}} = \begin{cases} H(r_{\mathrm{c}} - r_{\mathrm{ij}}) & \text{if } i \neq j \\ -\sum\limits_{\mathrm{i}(\neq \mathrm{j})}^{\mathrm{N}} \Gamma_{\mathrm{ij}} & \text{if } i = j \end{cases} . \tag{16}$$

Where N is the total number of residues; r_{c} is the contacting distance of two residues measured in the C_α positions (a value of $r_{\mathrm{c}} = 7$ was used in this work); r_{ij} is the distance between the C_α positions of the residues i and j; and $H(x)$ is the Heaviside step function, given as $H(x) = -1$ if $x > 0$ and $H(x) = 0$ if $x \leq 0$. The inverse matrix Γ^{-1} yields correlations between the thermal fluctuations per residue. Thus, the cross-correlation in the motion of pairs of residues is associated with the *k*th vibrational mode by the equation,

$$\langle \Delta \mathbf{R}_{\mathrm{i}} \Delta \mathbf{R}_{\mathrm{j}} \rangle_{\mathrm{k}} = (3k_{\mathrm{B}}T/\gamma)\lambda_{\mathrm{k}}^{-1} \lfloor \mathbf{u}_{\mathrm{k}} \mathbf{u}'_{\mathrm{k}} \rfloor_{\mathrm{ij}} . \tag{17}$$

From the above equation, the mean square fluctuation of residue *i* vibrating in a given subset of modes can be obtained from

$$\langle (\Delta \mathbf{R}_{\mathrm{i}})^2 \rangle_{\mathrm{k1-k2}} = (k_{\mathrm{B}}T/\gamma) \sum_{\mathrm{k=k1}}^{\mathrm{k=k2}} \lambda_{\mathrm{k}}^{-1} [\mathbf{u}_{\mathrm{k}}]_{\mathrm{i}}^2 \Big/ \sum_{\mathrm{k=k1}}^{\mathrm{k=k2}} \lambda_{\mathrm{k}}^{-1} . \tag{18}$$

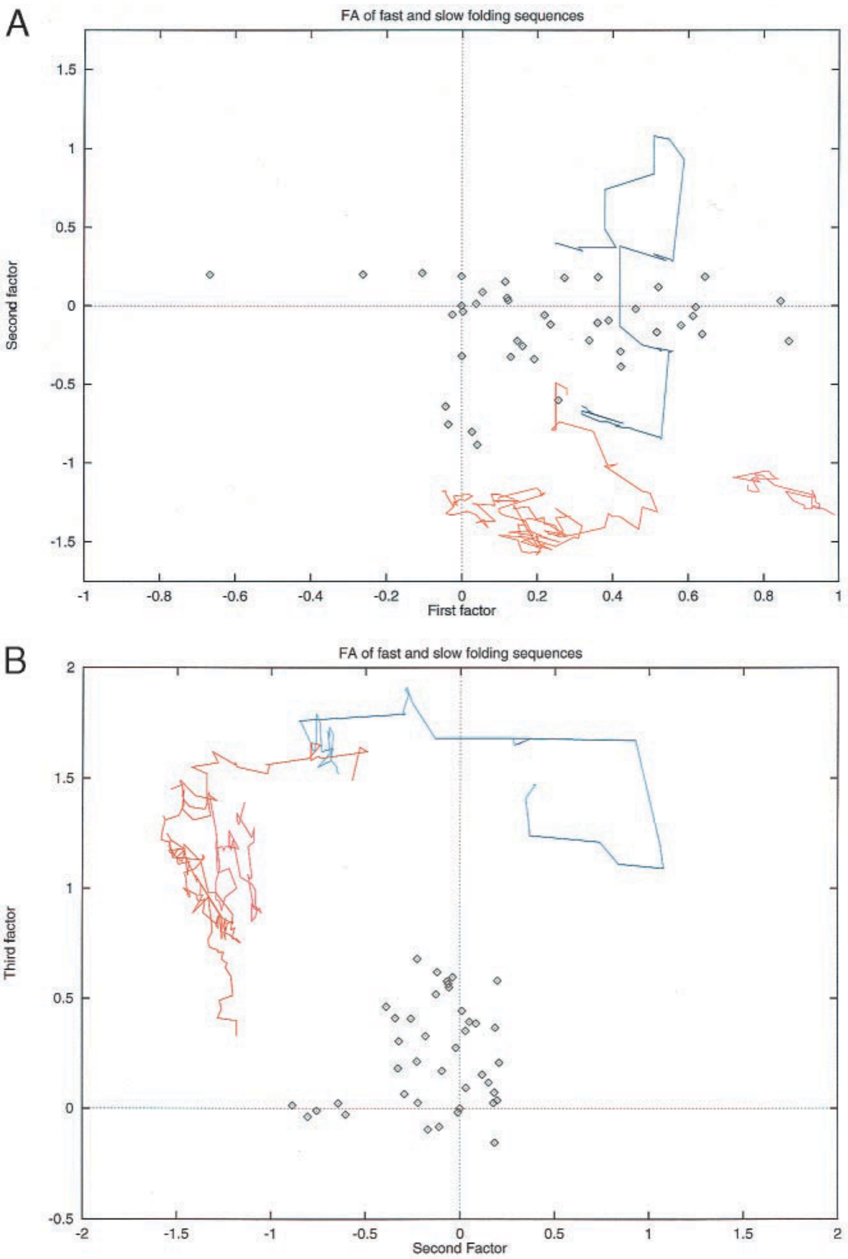

FIGURE 1 Factor analysis of fast- and slow-folding sequences. The analysis included the first 200 slow-folding sequences and the last 217 fast-folding sequences. (*A*) Two-dimensional (2D) plot of the two first factors. (*B*) 2D plot of the second and third factor. In both figures, loadings are represented as diamonds, whereas scores are represented as lines connecting sequences following the evolutionary time. Slow-folding sequences are represented by a blue line, fast-folding sequences by a red line, and very-fast-folding sequences by a magenta line.

Following Demirel et al. (1998), we study the mean square fluctuations induced by the modes $N - 4 \leq k < N$. As in their study, we consider "kinetically hot residues" to be those residues in these modes having an average mean square fluctuation above $6N^{-1}$.

Interaction energy analysis

Let us denote each one of the N positions of the lattice model by i. In the global minimum conformation, each position has a given contact coordination number nc_{i}, with each contact k contributing an energy $E^{\mathrm{S}}(nc_{\mathrm{ik}})$ for a given sequence. The homology averaged energy of each position in the global minimum is given by

$$E_{\mathrm{i}} = \frac{1}{N\,\mathrm{seq} \cdot nc_{\mathrm{i}}} \sum_{\mathrm{S}} \sum_{\mathrm{k}} E^{\mathrm{S}}_{\mathrm{i}}(nc_{\mathrm{ik}}). \tag{19}$$

This value can be averaged over a window $2w + 1$ to consider the mean homology-averaged energy of the different fragments in the structure,

$$\bar{E}_{\mathrm{i}} = \frac{1}{2w + 1} \sum_{\mathrm{j=i-w}}^{\mathrm{j=i+w}} E_{\mathrm{j}}\,. \tag{20}$$

The excess energy (normalized in its structural context) of each residue is then a measure related to the *local frustration* of that residue in its structural context,

$$F_{\mathrm{i}} = (E_{\mathrm{i}} - \bar{E}_{\mathrm{i}}) - \langle E_{\mathrm{i}} - \bar{E}_{\mathrm{i}} \rangle \tag{21}$$

whereas the average stability of the fragment in which the residue is immersed is given by

$$S_{\mathrm{i}} = \bar{E}_{\mathrm{i}} - \langle \bar{E}_{\mathrm{i}} \rangle. \tag{22}$$

Nonparametric regression

The nonparametric regression function $m(\mathbf{X})$ of a response variable Y in measuring a set of variables $\mathbf{X}$ is defined as the conditional mean of Y on $\mathbf{X}$, i.e., $m(\mathbf{X}) = E(Y|\mathbf{X})$. Thus, for a given sample of design variables $\{\mathbf{X}_{\mathrm{i}}\}^{\mathrm{n}}_{\mathrm{i=1}}$, the associated response variables $\{\mathbf{Y}_{\mathrm{i}}\}^{\mathrm{n}}_{\mathrm{i=1}}$ are of the form:

$$Y_{\mathrm{i}} = m(\mathbf{X}_{\mathrm{i}}) + \varepsilon_{\mathrm{i}} \quad \text{with } i = 1, \ldots, n, \tag{23}$$

where $\{\varepsilon_{\mathrm{i}}\}^{\mathrm{n}}_{\mathrm{i=1}}$ are independent errors. Within a given sample, the *nonparametric* estimate $\hat{m}$ of the function m has the general form

$$\hat{m}(\mathbf{X}) = \sum_{\mathrm{i}} w_{\mathrm{i}}(\mathbf{X})y_{\mathrm{i}} \quad \text{with} \quad \sum_{\mathrm{i}} w_{\mathrm{i}}(\mathbf{X})1. \tag{24}$$

Several nonparametric estimates have been proposed in the statistical literature. In this work, we considered the shifted Nadaraya–Watson (SNW) estimate (Mammen and Marron, 1997), which appears stable when data are sparse (Hardle and Marron, 1995). The SNW estimate is defined as

$$\hat{m}_{\mathrm{SNW}}(x) = \frac{\sum^{\mathrm{n}}_{\mathrm{i=1}} K_{\mathrm{h}}(x_{\mathrm{i}} - \xi^{-1}(x))y_{\mathrm{i}}}{\sum^{\mathrm{n}}_{\mathrm{i=1}} K_{\mathrm{h}}(x_{\mathrm{i}} - \xi^{-1}(x))}, \tag{25}$$

with

$$\xi(x) = \frac{\sum^{\mathrm{n}}_{\mathrm{i=1}} K_{\mathrm{h}}(x_{\mathrm{i}} - x))x_{\mathrm{i}}}{\sum^{\mathrm{n}}_{\mathrm{i=1}} K_{\mathrm{h}}(x_{\mathrm{i}} - x)}. \tag{26}$$

TABLE 1 Factor Analysis of fast- and slow-folding sequences

Residue	Factor 1	Factor 2	Factor 3
Residues important in the first factor (loading cutoff: $\|0.60\|$)			
14	0.614	−0.066	0.577
44	0.621	−0.009	−0.018
11	0.638	−0.181	0.328
23	0.645	0.184	−0.155
18	0.846	0.030	0.352
43	0.868	−0.226	0.213
7	−0.666	0.198	0.037
Residues important in the second factor (loading cutoff: $\|0.35\|$)			
16	0.423	−0.388	0.462
30	0.257	−0.602	−0.028
41	−0.042	−0.642	0.023
15	−0.035	−0.756	−0.010
19	0.028	−0.803	−0.037
35	0.042	−0.885	0.014

The analysis included the first 200 slow-folding sequences and the last 217 fast-folding sequences. Residues showing a loading in the corresponding factor above the given cutoff value are shown, together with their loadings in the first three first-rotated factors.

In the previous formulae, $\mathrm{K_h}(\cdot)$ is a renormalized kernel function, taken as a Gaussian density throughout this work, being h^{-1} K $(\cdot/h)$. The parameter h is the bandwidth or *smoothing parameter.* The role of h is essential in kernel regression. On the one hand a too low value leads to highly rugged surfaces and variant or sample dependence. On the other hand, high values introduce bias in the curve estimation and eliminate the fine structure of data. A simple quantification of this trade-off is given by the integrated mean square error (Hardle and Marron, 1995). It can be shown (Hardle and Marron, 1995) that the asymptotically optimal global bandwidth (i.e., the one that minimizes the integrated mean square error) for the SNW estimate is given by

$$h_{\mathrm{opt}} = \left[\frac{\int V(\mathbf{X})}{4n \int B^2(\mathbf{X})} \right]^{1/5}, \tag{27}$$

where $V(\mathbf{X})$ is the variance and $B^2(\mathbf{X})$ is the square bias. These terms can be estimated from the sample itself and plugged into Eqs. 24 and 25. Fast

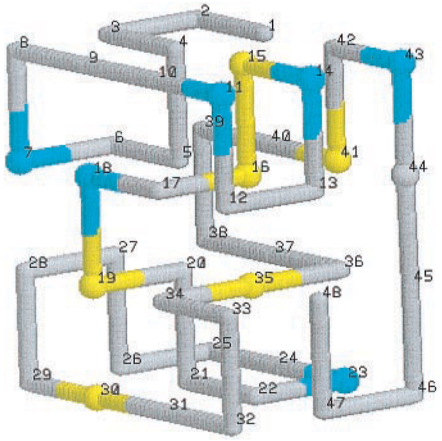

FIGURE 2 Mapping of the residues heavily loading in the first two factors onto the lattice structure, corresponding to the global minimum energy conformation of the 1000 48-mer sequences analyzed in this paper. The first factor is shown in blue and the second factor in yellow.

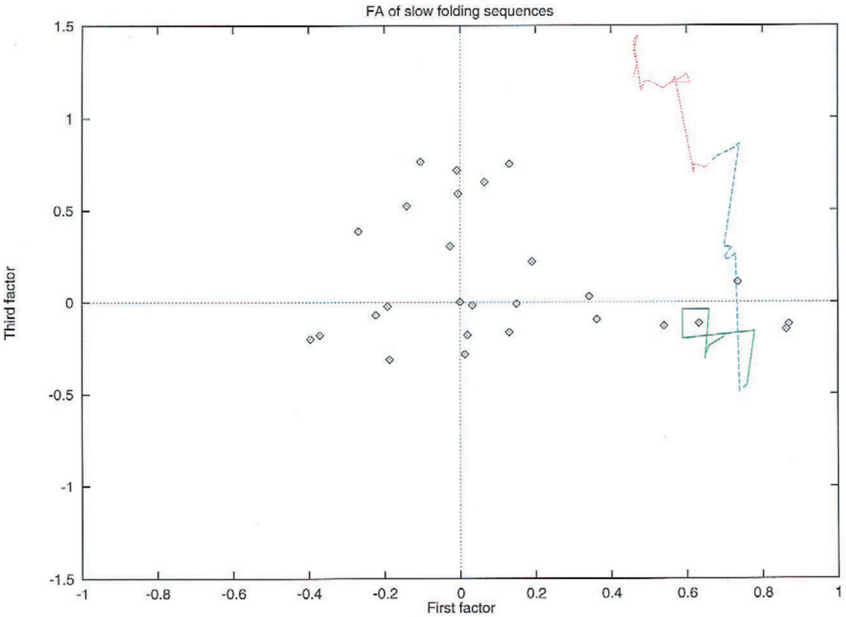

FIGURE 3 Factor Analysis of the slow-folding sequences. The analysis included the first 100 slow-folding sequences. Symbols as in Fig. 1. The 2D plot of the first and third factor is shown.

and simple estimates of these functions can be obtained based on polynomials and histograms, respectively. In the present work, the *block method* of Hardle and Marron (1995) for the estimates of the bias and variance appearing in Eq. 27 proved to be robust. This method is based on dividing the sample in histograms or blocks and fitting a low degree polynomial in each block.

RESULTS

Analysis of the lattice protein models

Sequence factor analysis

We start by trying to identify from the sequences, by using multivariate analysis, clusters of positions that can account for the kinetic differences between slow and fast folding lattice proteins. Fig. 1 and Table 1 summarize the results of an FA (Reyment and Joreskog, 1996) (see Methods for description of this multivariate technique) applied to the fast and slow folding protein sequences. The first 200 slow-folding and the last 217 fast-folding sequences were subjected to the analysis. In Fig. 1, *A* and *B,* the original sequence space spanned by the 200 + 217 = 417 aligned sequences is projected onto the first two dimensions or factors (Fig. 1 *B*), or onto the first and third dimension (Fig. 2). In Fig. 1, loadings are represented as diamonds, whereas scores are represented as lines (see Methods for definitions of loadings and scores). These lines connect the 417 sequences following the in silico evolutionary time of the evolutionary experiment. Sequences in the alignment are classified into three groups: slow-folding sequences (represented by a blue line), fast-folding sequences (red line), and very-fast-folding sequences (magenta line). Note that the coordinates of the line points in the second factor are very different for slow- and fast-folding sequences, i.e., the points of these two groups of sequences are well separated along this axis. Also, there is some separation along the first factor of the very-fast-folding sequences from the rest. Finally, note that there is no significant discrimination along the third factor (Fig. 1 *B*). Thus, this analysis suggests that, at the sequence level, the main differences in the kinetic behavior of the lattice proteins can be explained by two underlying factors.

TABLE 2 Factor Analysis of the slow-folding sequences only

Residue	Factor 1	Factor 2	Factor 3
Residues important in the third factor (loading cutoff: |0.70|)			
41	−0.008	0.180	0.716
35	0.131	−0.013	0.751
15	−0.104	0.012	0.762
Residues important in the second factor (loading cutoff: |0.70|)			
27	−0.026	−0.709	0.304
30	0.363	−0.735	−0.094
47	0.033	−0.754	−0.017
19	0.343	−0.794	0.031
Residues important in the first factor (loading cutoff: |0.70|)			
23	0.735	−0.217	0.110
17	0.863	−0.141	−0.148
18	0.870	−0.192	−0.119

Residues showing a loading in the corresponding factor above the given cutoff value are shown.

The first factor, explaining a higher proportion of the variance, is what we call the *loop factor,* because residues heavily loading in the first component are loop residues (Fig. 2). Within the loop residues, two types of positions can be distinguished, evolving in an anticorrelated fashion: position #7 becomes more hydrophilic, whereas the rest of residues become more hydrophobic. The later are in contact and apparently are required to fix or stabilize the loop

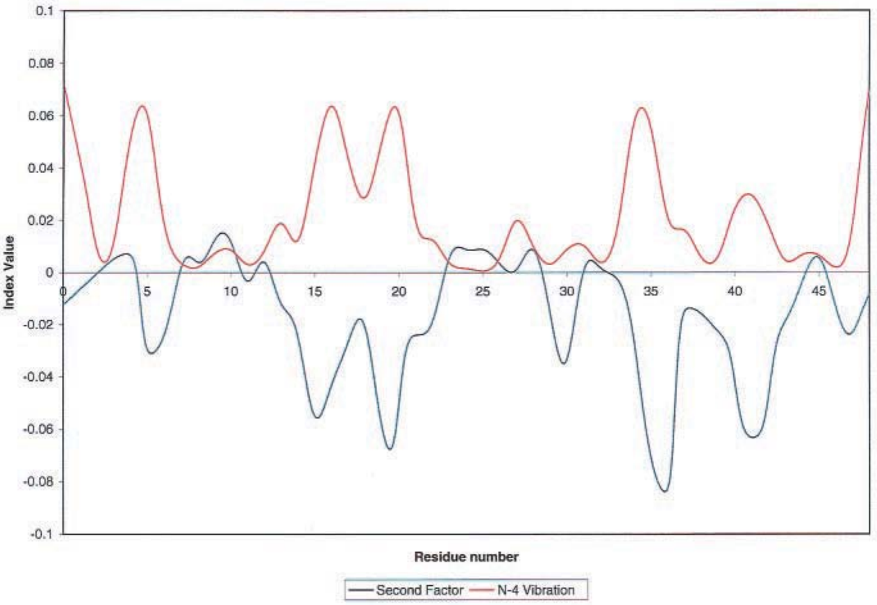

FIGURE 4 The second factor obtained from the FA of fast- and slow-folding sequences is plotted as a function of residue number (blue line), together with the average mean square fluctuation per residue in the last $N - 4$ to $N - 1$ modes, calculated from a GNM analysis of the global minimum of the sequences. Both lines have been previously smoothed by a nonparametric regression, as described in Methods.

conformation, whereas the former possibly avoids competing interactions with the core, making the energy landscape less rugged. Karplus and coworkers (Dinner et al., 1998) have recently obtained similar conclusions for a different system using a different potential energy function. However, they observed that only the weakening of interactions between noncore residues increases the folding rate.

The second factor accounting for a high proportion of the variance is the *nucleus factor.* A subset of the residues belonging to the previously detected thermodynamic folding nucleus (Mirny et al., 1998) is found here (Fig. 2). However, not all the residues previously described as being part of the folding nucleus are discriminating (compare the yellow residues in Fig. 2 with the red residues in Fig. 5). This is, in part, the result of complete conservation of some of these residues between fast- and slow-folding sequences (e.g., for residues #5 and 20). There are also some exceptions, for example, with residue #16. In contrast, residues #19 and 30 contribute to the differences in rate, but are not part of the thermodynamic folding nucleus.

It is important to note that the ranking of factors obtained from the FA model only reflects the fact that there are more representatives of the medium- and fast-folding sequences than of the slow-folding ones, resulting from the fast optimization of the folding rate during the first mutations. Thus, from the viewpoint of explaining the covariance structure, the loop factor is more important (i.e., accounts for a higher proportion of the variance) than the nucleus factor. However, the nucleus factor is far more important in its ability to discriminate fast- from slow-folding sequences, whereas the loop factor only makes a modest contribution. This is evidenced by the separation of the fast- and slow-folding proteins along both axes.

To confirm the results of this analysis, it is also of interest to carry out an FA using the slow-folding sequences alone. Figure 3 and Table 2 show the results. The analysis indicates that the third factor can explain the fast increment in the folding rate of the sequences (Fig. 3), whereas the first two factors have much less explanatory power. Residues loading in Factor 3 (Table 2) are mainly a subset of those residues (#15, 41, and 35) also identified as being part of the thermodynamic folding nucleus by Shakhnovich and coworkers (Mirny, et al., 1998) (Fig. 5).

Thus, differences in kinetic behavior of the model proteins originate from sequence differences detectable by multivariate analysis of the alignments. Folding rate is optimized rapidly, with the bulk of the optimization depending only on mutations in a small number of residues, a subset of the thermodynamic folding nucleus. Once this nucleus is stabilized, a slight increment in folding rate can be achieved by loop mutations.

Structural vibrational analysis

Following the ideas of Demirel et al. (1998), we identify those residues from the protein structure which can be described as being *kinetically hot.* By this, we mean that these residues present a high vibrational frequency in the folded state, implying that their motion is of small amplitude. This is a signature of a steep potential energy surface around their mean position, which arises from the contact coordination number of the residue and the topological restrictions imposed by the structure. The vibrations of the whole structure can be decomposed into modes, with all residues vibrating in a particular mode presenting concerted motions. Thus, residues forming part of the same high-frequency modes correspond to those residues having a dense network of interactions, being rather localized in space and moving coherently in the context of the tertiary fold. As demonstrated by Demirel et al., they are candidates to be part of the folding nucleus.

The main features of the protein fluctuation in the folded state can be obtained by application of the GNM (Bahar, et al., 1997). Thus, a vibrational analysis of the lattice conformation in the global energy minimum has been performed using the GNM (see Methods). The average residue fluctuations in the last $N - 4$ to $N - 1$ modes have been compared with the loadings in the second factor obtained from the FA of the fast- and slow-folding sequences (see above). The values obtained from the raw analysis were smoothed by a shifted Nadaraya–Watson nonparametric regression (Mammen and Marron, 1997) using the Hardle–Marron automatic global bandwidth selector (Hardle and Marron, 1995). Missing values from the loadings (i.e., null values) were removed from the design points and predicted by the regression curve (see Methods). The results are shown in Fig. 4.

TABLE 3 Factor Analysis of fast-folding sequences

Residue A	Residue B	Pearson r	Partial r	Factor	δ
Predicted contacts (no factor correction; $r_{\text{cut}} = 0.40$; $p_{\text{cut}} = 0.40$)					
3	6	0.400	0.534	1	0
10	44	0.480	0.579	2	3
11	13	0.692	0.606	1	1
11	45	0.478	0.694	1	2
12	17	0.509	0.640	1	0
13	48	0.439	0.586	1	0
17	33	0.511	0.666	1	1
17	34	0.443	0.454	1	0

Residue	Factor 1	Factor 2
Residues important in the first factor (loading cutoff: |0.30|)		
3	0.558	0.139
6	0.429	0.224
8	0.482	0.036
11	0.730	0.072
12	0.796	0.076
13	0.720	0.015
17	0.669	−0.111
33	0.613	0.056
34	0.476	0.060
45	0.688	−0.021
48	0.374	−0.059
Residues important in the second factor (loading cutoff: |0.30|)		
7	0.122	0.544
10	−0.098	0.657
18	0.150	0.324
19	−0.312	0.540
21	−0.042	0.516
23	−0.152	0.475
24	0.035	0.462
27	−0.060	0.426
44	−0.144	0.790

The analysis included the last 517 fast-folding sequences. Residues showing a loading in the corresponding factor above the given cutoff value are shown.

Residues #5 and 20, forming part of the thermodynamic folding nucleus detected by Shakhnovich and coworkers but kept constant during the evolutionary experiment, are predicted to contribute significantly to the increased folding rate, particularly residue #20. The rest of the residues not mutated during the computational experiment are predicted to be almost irrelevant for folding-rate optimization.

The similarity in the behavior of both curves in Fig. 4 is striking. We note that one of the curves is derived from the analysis of the sequences alone, whereas the other one derives from an analysis devoid of sequence information, of the topology in the global minimum. This result implies that, at least for this simple lattice model, the topology itself determines which residues will participate both in the folding nucleus and in the optimization of the folding rate. An interpretation of this conclusion is that both analyses are reflecting the same physical process. The multivariate analysis of the sequences uncovers the subset of the most important positions in changing the folding rate, and therefore the free energy of the transition state. In contrast, it is thought that the selection of these residues in the folding nucleus arises in the physical system as the result of a topology-dependent optimal enthalpy–entropy balance in making particular intramolecular contacts (Alm and Baker, 1999; Galzitskaya and Finkelstein, 1999; Munoz and Eaton, 1999). The vibrationally coupled units detected by the GNM in the native structure appear to be sensitive to a similar balance, as empirically shown by Demirel et al. (1998), perhaps as a result of the topological resemblance between the transition and native states.

Interaction energy analysis

Interaction energy analysis has been used to obtain a deeper insight into the energetics underlying the results of the GNM and FA analysis. It would be expected that the residues identified by the GNM and FA interact more favorably than average with the rest of the protein. If so, that would imply that they might constitute a critical folding nucleus. For a heteropolymer to fold to a specific native structure, native interactions must be stronger than those present in alternative nonnative states; that is, the energy gap between the native state and first excited state must exceed some threshold value (Sali et al., 1994). At the same time, to fold fast, it is necessary to lower the free energy barrier of the transition state, corresponding to the formation of the critical nucleus (Shakhnovich, 1997). Thus, stabilization of the interactions of this nucleus with respect to the rest of the interactions increases the folding rate. Is this description consistent with our findings of the sequence–topology connection in the model proteins?

This is supported by the findings shown in Fig. 7. Here, we plot a parameter related to the average local frustration of each residue in the lattice protein versus a parameter reflecting the average fragment stability in which the resi-

TABLE 4 Average sequence distance between predicted contacts and kinetically hot residues, together with its statistical significance, evaluated for a representative set of real proteins

PROT[*]	⟨seq(p1)⟩[†]	⟨seq(p2)⟩[‡]	⟨seq(rnd)⟩[§]	Sdev(rand)[¶]	Zscore(p1)[‖]	Zscore(p2)[**]
1hrc	5.846	15.461	14.273	3.328	−2.532	0.357
1pca	1.846	7.769	8.325	1.987	−3.261	−0.280
2ci2	2.250	10.500	6.433	1.224	−3.417	3.321
1shg	5.778	10.333	7.551	1.040	0.601	−1.081
1ubq	6.300	12.100	7.298	1.439	1.182	−0.537
1tlk	1.250	5.750	5.526	1.977	−2.162	0.113
2adb	1.846	8.769	8.345	1.962	−3.312	0.216
2acy	6.000	24.083	12.776	3.354	−2.020	3.371
3chy	2.000	6.666	10.196	6.620	−1.238	−0.533
T059	2.000	12.250	7.902	2.830	−2.086	1.536
T056	3.917	17.583	9.762	1.794	−3.258	4.359
T077	5.500	9.000	7.550	2.386	−0.859	0.607
T074	0.000	20.500	8.833	5.625	−1.570	2.074
T079	2.333	6.333	5.947	3.323	−1.087	0.116

[*] Protein studied. PDB (Bernstein et al., 1977) entry name or CASP3 entry name is given (for description of CASP3 proteins, see *http://PredictionCenter.llnv.gov/casp3*).

[†] Average distance between the closest element of the predicted contact from the correlated mutation analysis and the closest kinetically hot residue.

[‡] Closest average distance for the other element of the pair.

[§] Average values obtained after 1000 bootstrapping runs.

[¶] Standard deviation of the bootstrapping runs.

[‖] Z-score value for the closest element of the predicted contact pair.

[**] Z-score value for the other element of the pair.

due is immersed (see Methods). A strong segregation of residues can be observed for the fast-folding sequences when compared with the slow-folding ones: in the fast-folding sequences, most of residues are close to inert, with slightly repulsive environments and with a small value of frustration. Among the group of residues with low values of local frustration are residues #16, 41, and 20, all of them important in increasing the folding rate (Fig. 4), with residues #41 and 16 forming a contact and loading together in the FA. Thus, during the optimization of the folding rate, strong favorable interactions are placed between these residues of the folding nucleus, and mild repulsive or inert interactions (with respect to the average) are placed elsewhere. These results support the picture of the sequence–topology connection and suggest that a way to engineer fast folding in real proteins is to select the kinetically hot residues determined by the vibrational analysis of the topology and then to optimize only the interactions of these residues.

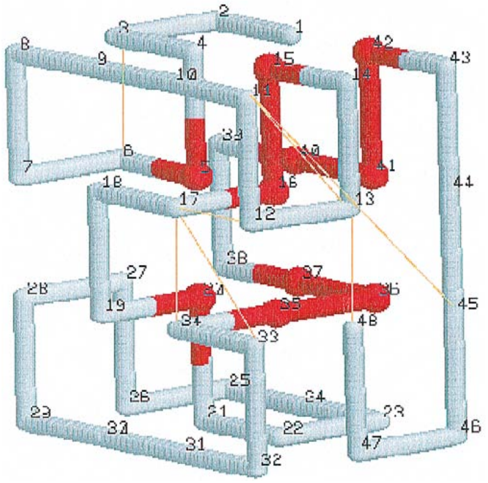

FIGURE 5 Display of the predicted contacts for the model sequences using an alignment based on the last 517 fast-folding sequences. Predicted contacts are shown as dotted lines connecting the corresponding residues. Positions in red correspond to the residues participating in the thermodynamic folding nucleus as described by Shakhnovich and coworkers (Mirny et al., 1998).

Presence of correlated mutations

After the folding rate is optimized, accepted mutations will tend to minimally perturb the stability of the critical nucleus. This foldability requirement should tend to create restrictions in sequence space. It is becoming well established that residues forming part of the folding nucleus tend to be conserved (Ptitsyn, 1998; Shakhnovich, et al., 1996), but it is also possible to imagine some other, more subtle restrictions imposed by the folding nucleus. Thus, the mutational behavior of other residues outside the nucleus could also be restricted in varying degrees, creating patterns of variability, for example, in the form of correlations. Indeed, Table 3 and Fig. 6 show that correlated mutations emerge from the set of evolutionary related sequences under fast folding pressure. Most important, the residues involved in correlation are either forming contacts or shifted in sequence by, at most, two residues in contact map space (Table 4).

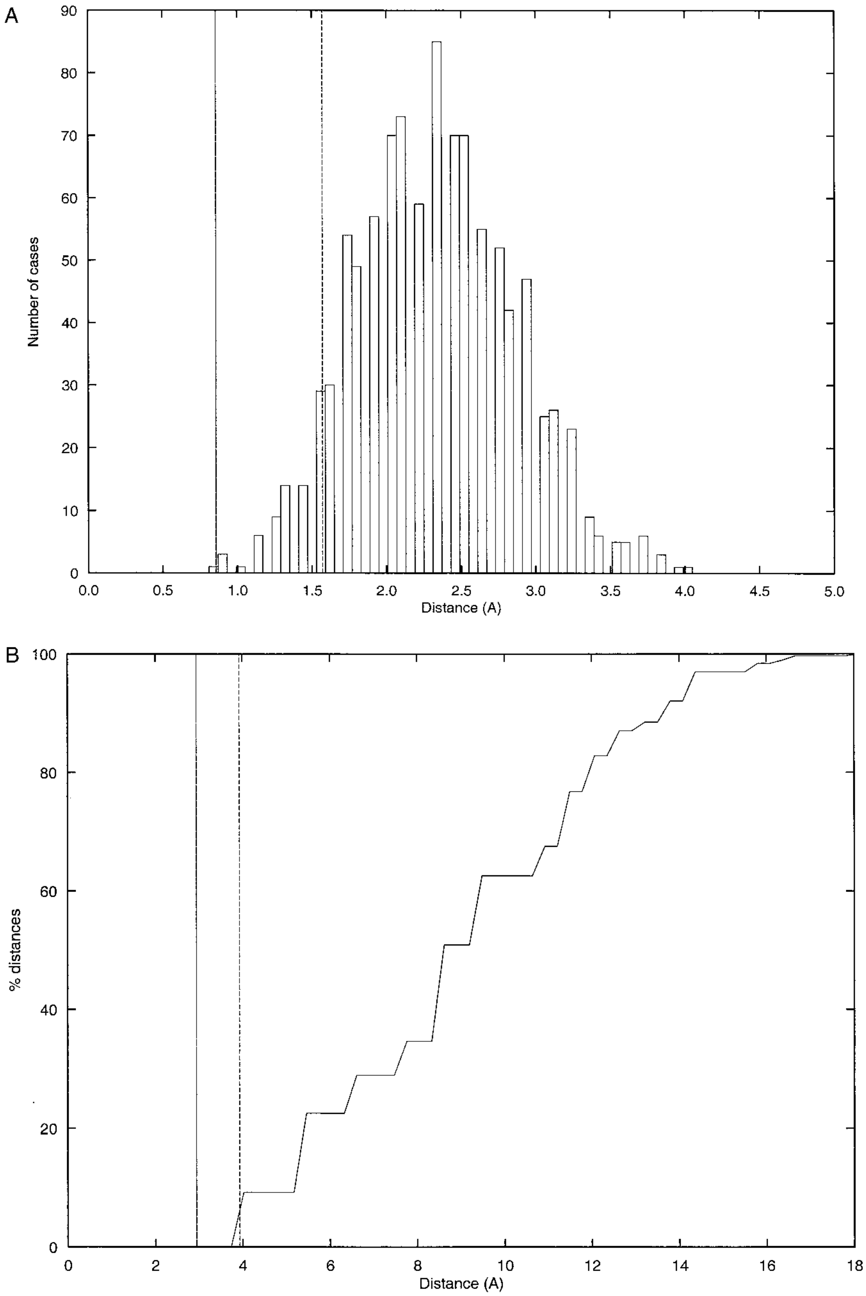

FIGURE 6 (*A*) Probability distribution of the average sequence distance separation between the folding nucleus and the predicted contacts in the fast-folding model sequences. (*B*) Percentage of distances within a given distance for the lattice structure, together with the average values of the observed distances between correlated positions and kinetically hot residues.

An FA of the fast-folding sequences alone shows that the residues having correlated mutations are mostly in a common subspace (first factor of the loadings matrix, accounting for the 9.6% of the total variance, see Table 4). This means that all pairs of correlated mutations change roughly together and in a coherent fashion. The existence of this concerted behavior suggests the presence of a common physical mechanism explaining the changes in all pairs of correlated positions at the same time. The physical basis of this underlying factor appears to be the closeness to the critical residues (Fig. 5). This has been established by comparing the probability of obtaining the observed average sequence distance by chance between the correlated positions and the kinetically hot residues (Fig. 6 *A*). For the closest element of the correlated pair, the observed average closeness in sequence of 0.85 is significant at the 95% confidence interval ($Z_{\text{score}} = -2.06$). However, the results are not significant for the other element of the pair ($Z_{\text{score}} = -1.02$). In 3D space, the results are similar, showing that residues having correlated mutations form a shell around the kinetically hot residues, which cannot be explained by chance (Fig. 6 *B*). Thus, on the basis of this analysis, several structural phases can be distinguished in the sequence space of a protein under the steady-state folding rate regime: the

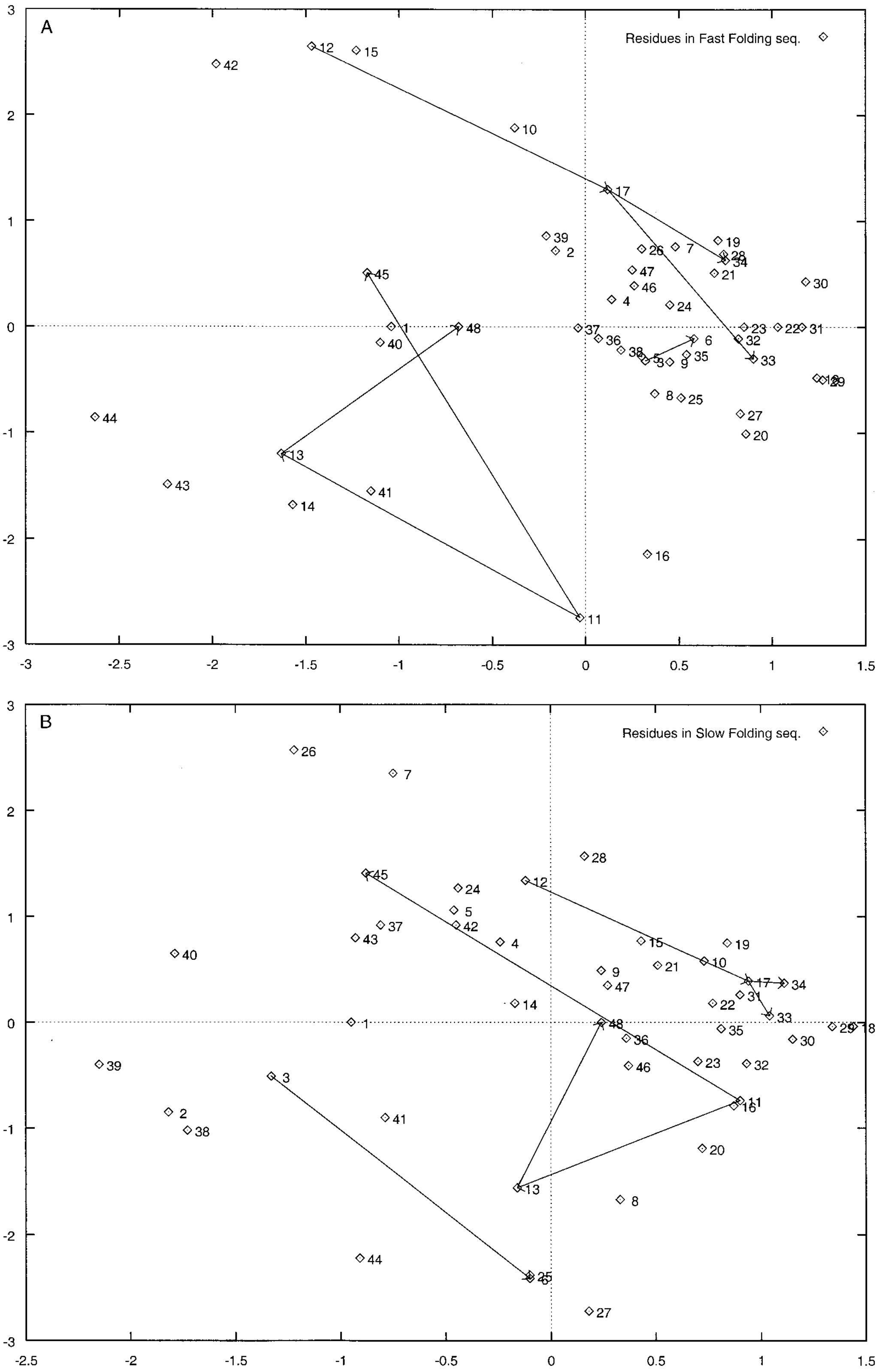

FIGURE 7 Plot of the average fragment stability (Eq. 22) versus the average residue frustration (Eq. 21) for the 1000 48-mer model sequences using the Miyazawa–Jernigan (Miyazawa and Jernigan, 1985) interaction energy matrix. The numbers indicate the corresponding residue number of the 48-mer, and the lines link pairs of residues predicted to be in contact. (*A*) Plot obtained from the fast-folding sequences. (*B*) The same plot from the slow-folding sequences.

TABLE 5 Three-dimensional distance between predicted contacts and kinetically hot residues, together with the results of the computation of its statistical significance

PROT*	⟨dist(p1)⟩[†]	⟨dis(p2)⟩[†]	P-rand[†]	%Rg(p1)[¶]	%dist(p1)[‖]	%Rg(p2)**	%dist(p2)[††]
1hrc	6.215	13.664	7.70E-04	51.00	6.758	108.00	32.412
1pca	2.301	9.371	1.46E-03	21.00	0.000	84.00	17.882
2ci2	5.673	10.208	6.16E-05	54.00	6.923	95.99	24.759
1shg	7.869	10.100	0.42	78.00	16.541	99.00	29.010
1ubq	9.788	14.290	8.53E-03	87.00	18.666	126.00	45.649
1tlk	1.105	11.852	1.25E-02	9.00	0.000	93.00	23.643
2abd	3.949	8.954	7.64E-05	36.00	2.462	75.00	14.637
2acy	10.019	12.866	9.14E-02	81.00	18.030	105.00	33.747
3chy	4.267	6.373	0.36	33.00	1.648	48.00	5.757
T059	2.859	11.720	1.36E-05	27.00	0.000	108.00	34.889
T056	5.445	11.320	1.91E-07	42.00	4.455	86.99	22.527
T-77	8.639	10.724	0.18	69.00	12.901	84.00	21.994
T074	0.000	16.106	1.93E-02	3.00	0.000	132.00	55.475
T079	2.000	8.930	—	15.00	0.000	69.00	10.914

*Protein studied. PDB (Bernstein et al., 1977) entry name or CASP3 entry name is given (for description of CASP3 proteins, see *http://PredictionCenter.11nv.gov/casp3*).

[†] Average distance (in Å) between the closest mutation pair and the kinetically hot residues.

[‡] The same for the other element of the pair.

[§] Probability that the distance distributions of both elements of the pair come from the same underlying distribution according to a two-tail Student's *t* paired sample test (Press et al., 1989).

[¶] Percentage of the radius of gyration corresponding to the distances shown in the first column.

[‖] Percentage of all distances between residues found in the protein below the distances shown in the first column.

**As in ¶, for the second column.

[††] As in ‖, for the second column.

(almost) conserved folding nucleus, the shell of residues around it involved in correlated mutations, and the rest of residues mutating (almost) independently.

Analysis of real proteins

We were interested to see whether there is a qualitative correspondence between the results obtained with the lattice proteins and what can be observed in real proteins. Although the situation with real proteins is more complex, at least a qualitative agreement should be obtained. For such comparison to be meaningful, the same computational procedures should be applied in both cases, with the same coarse graining. Thus, core residues were automatically selected from the 3D structure in the same way it was done for the lattice proteins, based on the GNM calculations, while correlations in the alignments were also computed following the same procedure used with the lattice proteins.

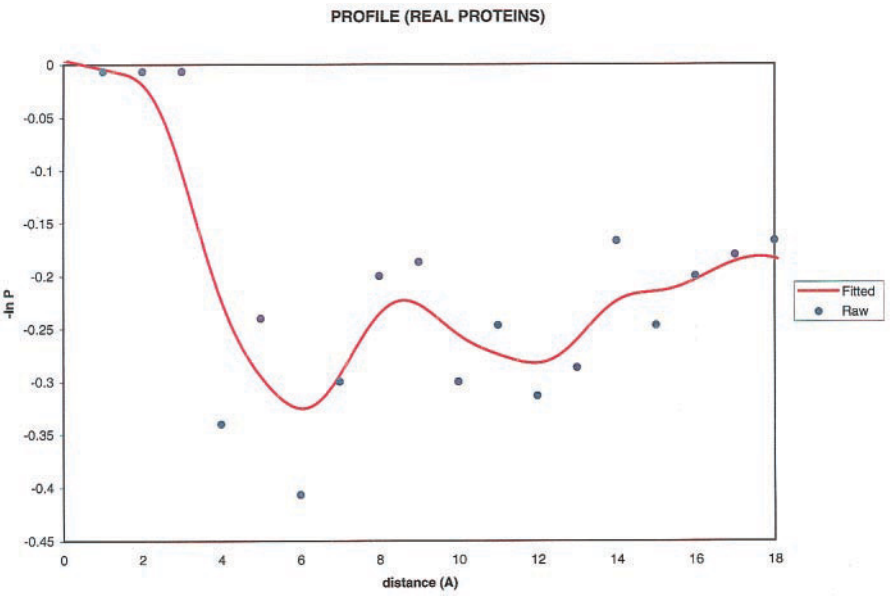

FIGURE 8 Profile of the preference of observing a residue having correlated mutations with respect to the average values observed in protein structures at a given distance. The points represent the raw values obtained from the direct analysis. The line is the nonparametric regression curve (see Methods). The values have been shifted to zero at the origin.

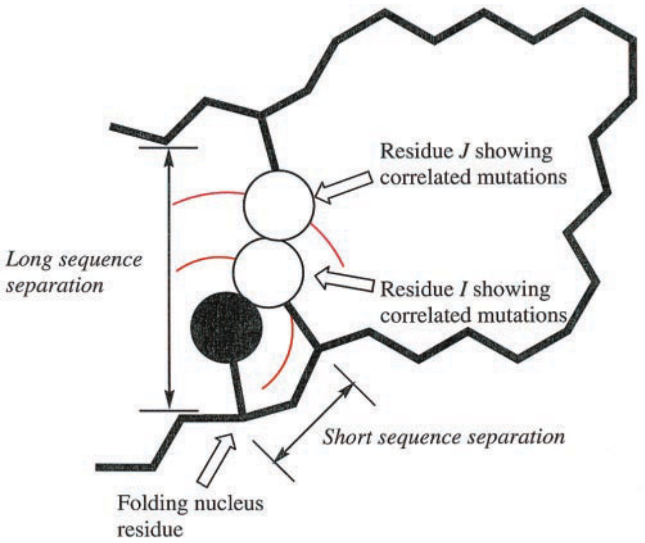
Residue *J* showing correlated mutations
Long sequence separation
Residue *I* showing correlated mutations
Short sequence separation
Folding nucleus residue

FIGURE 9 A schematic picture of the 3D relationship between kinetically hot/folding nucleus residues and residues having correlated mutations in sequence space.

We created a test set including proteins for which experimental data about their folding nucleus were available, as well as proteins for which contacts were predicted blindly, in advance of the knowledge of the structure, during the recent CASP3 contest (*http://PredictionCenter.llnv.gov/casp3*) (see Ortiz et al., 1999).

Contacts are predicted from the MSAs as described (Ortiz et al., 1999) (see Methods). Overall, the prediction accuracy in this small sample is similar to that obtained when a larger number of proteins was used when developing the prediction method. Thus, most of the predicted contacts have correspondence with a real contact within a local sequence window of $\delta = 3$ (data not shown). In contrast, from the topology of the protein, a vibrational analysis with the GNM was conducted as described (Bahar, et al., 1997). In agreement with Demirel et al. (1998), we observe a statistically significant overlap between the experimentally described folding nucleus and the kinetically hot residues (data not shown). Similar to the results obtained with the lattice protein, we also observe a statistically significant short sequence distance between the closest element of the predicted contact from the correlated mutation analysis and the closest kinetically hot residue (Table 4), although this is not the case for the second element of the pair (Table 4). Similar results were obtained when analyzing the relationship in 3D space. Thus, it is found that residues from the closest member of a correlated mutation pair tend to appear in the first coordination shell of a kinetically hot residue more often than expected by chance (see Table 5 and Fig. 8). A somewhat weaker tendency is observed for the second element, which tends to be located in the second solvation shell of the kinetically hot residues (Fig. 8) and in contact with the first element of the pair. Both elements have significantly different radial distribution functions with respect to the kinetically hot residues (Table 5). A qualitative picture of the relationships can be seen in Fig. 9.

CONCLUDING REMARKS

Topology is a main factor determining the identity of the residues forming the folding nucleus, and folding rate is strongly dependent on the stability of a subset of these residues. The requirement of a minimum stability in the folding nucleus appears to create some restrictions in the sequence space of the residues forming the coordination shell around the critical nucleus. One realization of these restrictions is in the form of correlated mutations, which, as a result of these topological constraints, tend to occur with higher frequency between contacting residues. These results are consistent with a nucleation–condensation model for protein folding and have implications in the development of methods for structure prediction.

We thank L. A. Mirny and E. I. Shakhnovich for making their database of model proteins available to us. A.R.O. acknowledges partial support from the Spanish Ministry of Education. A.R.O. also acknowledges Pere Constans for discussions and source code concerning nonparametric regression. Support from National Institutes of Health grant GM37408 is gratefully appreciated.

REFERENCES

Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA.* 96:11305–11310.

Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des.* 2:173–181.

Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. E. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Demirel, M. C., A. R. Atilgan, R. L. Jernigan, B. Erman, and I. Bahar. 1998. Identification of kinetically hot residues in proteins [In Process Citation]. *Protein Sci.* 7:2522–2532.

Dinner, A. R., S. S. So, and M. Karplus. 1998. Use of quantitative structure–property relationships to predict the folding ability of model proteins. *Proteins.* 33:177–203.

Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA.* 96:11299–11304.

Hardle, W., and J. S. Marron. 1995. Fast and simple scatterplot smoothing. *Comput. Stat. Data Analysis.* 20:1–17.

Johnson, R. A., and D. W. Wichern. 1992. Applied Multivariate Statistical Analysis. 3rd ed. Prentice Hall, Upper Saddler River, NJ.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika.* 23:187–200.

Koonin, E. V. 1997. Big time for small genomes. *Genome Res.* 7:418–421.

Koonin, E. V., R. L. Tatusov, and M. Y. Galperin. 1998. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8:355–363.

Mammen, E., and J. S. Marron. 1997. Mass recentered kernel smoothers. *Biometrika.* 84:765–777.

Mirny, L. A., V. I. Abkevich, and E. I. Shakhnovich. 1998. How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. USA.* 95:4976–4981.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.

Munoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA.* 96:11311–11316.

Ortiz, A. R., A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins.* (Suppl.) 3:177–185.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. Numerical Recipes. The Art of Scientific Computing. Cambridge University Press.

Ptitsyn, O. B. 1998. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* 278:655–666.

Reyment, R., and K. G. Joreskog. 1996. Applied Factor Analysis in the Natural Sciences. Cambridge University Press.

Sali, A., E. Shakhnovich, and M. Karplus. 1994. How does a protein fold? *Nature.* 369:248–251.

Sander, C., and R. Schneider. 1991. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins.* 9:56–68.

Shakhnovich, E. I. 1996. Modeling protein folding: the beauty and power of simplicity. *Fold. Design.* 1:R50–R54.

Shakhnovich, E. 1997. Theoretical studies of protein folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* 7:29–40.

Shakhnovich, E., V. Abkevich, and O. Ptitsyn. 1996. Conserved residues and the mechanism of protein folding. *Nature.* 379:96–98.