doi:10.1016/j.jmb.2003.11.007

J. Mol. Biol. (2004) **335**, 869–880

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

COMMUNICATION

Are Residues in a Protein Folding Nucleus Evolutionarily Conserved?

Yan Yuan Tseng and Jie Liang*

Department of Bioengineering
SEO, MC-063, University of
Illinois at Chicago, 851 S.
Morgan Street, Room 218
Chicago, IL 60607-7052, USA

**Corresponding author*

Protein is the working molecule of the cell, and evolution is the hallmark of life. It is important to understand how protein folding and evolution influence each other. Several studies correlating experimental measurement of residue participation in folding nucleus and sequence conservation have reached different conclusions. These studies are based on assessment of sequence conservation at folding nucleus sites using entropy or relative entropy measurement derived from multiple sequence alignment. Here we report analysis of conservation of folding nucleus using an evolutionary model alternative to entropy-based approaches. We employ a continuous time Markov model of codon substitution to distinguish mutation fixed by evolution and mutation fixed by chance. This model takes into account bias in codon frequency, bias-favoring transition over transversion, as well as explicit phylogenetic information. We measure selection pressure using the ratio ω of synonymous *versus* non-synonymous substitution at individual residue site. The ω-values are estimated using the PAML method, a maximum-likelihood estimator. Our results show that there is little correlation between the extent of kinetic participation in protein folding nucleus as measured by experimental ϕ-value and selection pressure as measured by ω-value. In addition, two randomization tests failed to show that folding nucleus residues are significantly more conserved than the whole protein, or the median ω value of all residues in the protein. These results suggest that at the level of codon substitution, there is no indication that folding nucleus residues are significantly more conserved than other residues. We further reconstruct candidate ancestral residues of the folding nucleus and suggest possible test tube mutation studies for testing folding behavior of ancient folding nucleus.

Keywords: protein folding; folding nucleus; ϕ value; continuous time Markov process; ancestral folding nucleus

Are amino acid residues important for rapid folding preferentially conserved during evolution? Does natural selection optimize proteins for folding kinetics? If protein folding involves initially the formation of a small region of native-like folding nucleus, are identities of these residues well conserved during evolution?[1–5] These fundamental questions of molecular biology have received much attention.[1–12] Of direct relevance are experimental ϕ-value studies, which provide information about the role individual amino acid residues play in the formation of folding nucleus.[13–15] By measuring the change $\Delta\Delta G$ in protein stability and the change $\Delta\Delta G^{\ddagger}$ in folding barrier due to mutation of an amino acid residue, ϕ-value (defined as $\phi \equiv \Delta\Delta G^{\ddagger}/\Delta\Delta G$) for the mutated residue can be calculated. ϕ-Value has been used to measure the extent to which the side-chain of a mutated residue participates in native-like interactions. A ϕ-value of 0.0 indicates that the site of mutation is as unfolded as in the denatured state. A ϕ-value of 1.0 indicates that the site of mutation is as folded as in the native state, i.e. this residue is involved

Abbreviation used: PAML, phylogenetic analysis of maximum likelihood.

E-mail address of the corresponding author: jliang@uic.edu in native-like transition state structure, and is a part of the folding nucleus. A ϕ-value between 0 and 1 is interpreted as possessing different degrees of structure in transition state.[13] Folding nucleus can be identified as formed by the set of residues with ϕ values above a threshold (e.g. $\phi \geq 0.5$).[13] Several computational methods have been developed for predicting protein-folding mechanism and ϕ-values of residues. These include the sequential binary collision model,[16] multisegment model,[17] and single-to-triple sequence approximation model.[18] Model conformations of transition-state ensemble have also been generated explicitly by Monte Carlo sampling using Gō-type potential derived from experimental ϕ-values constraints.[19] A lucid statistical mechanistic picture for understanding ϕ-value experiments can be found in Refs. 20 and 21.

The evolutionary conservation of folding nucleus residues is the subject of several recent studies. These studies, however, have come to different conclusions. Plaxco *et al.* and Larson *et al.* showed that there may be little correlation between sequence conservation and participation in the folding transition state.[8,9] Mirny, Shahknovich and others demonstrated that for rapid folding, sequence identity of folding nucleus is more conserved within protein families and across protein superfamilies.[2,7] It is unclear whether the disagreement between these studies is due to the difference in entropy calculations as attributed by Mirny & Shakhnovich,[7] or differences in choice and processing of the data set, in sequence alignments, in definition of folding nucleus, as well as intrinsic sample bias in ϕ-value analysis, as discussed in detail by Larson *et al.*[9]

Here, we examine the conservation of folding nucleus residues using an approach that differs from previous studies in several aspects. First, instead of studying amino acid residue sequences, we examine the evolution of corresponding coding DNA sequences at the codon level. Second, we use an explicit codon evolutionary model based on continuous time Markov process, which has yielded deep insights about the mechanisms of molecular evolution.[22–24] Instead of using entropy or relative entropy as a quantitative measure of sequence conservation, we assess the ratio of mutation rates of synonymous *versus* non-synonymous changes to detect natural selection at each amino acid residue position. Third, a phylogenetic tree is built to encode the closeness between proteins. Following earlier studies,[25,26] we use the maximum likelihood method developed by Yang[27] to estimate values of parameters of the evolutionary model and draw inference about the conservation of folding nucleus residues.

We find that experimental ϕ-values are not correlated with evolutionary conservation for seven proteins studied here. In addition, results using two statistical tests indicate that except possibly one protein, none of these proteins has folding nucleus more conserved than the rest of the proteins, or than the residue with median selection pressure. We have also reconstructed candidate ancestral folding nucleus residues, and have suggested exploratory test-tube mutation studies on the evolution of protein-folding dynamics.

Synonymous and non-synonymous codon substitution

Protein sequences diverge from a common ancestor because mutations occur. Some fraction of these mutations is fixed into the evolving population by selection and some are fixed by chance, resulting in the substitution of one nucleotide for another nucleotide at various locations. Because evolution occurs at DNA level rather than at amino acid level, models of protein evolution based on codon usage are appealing and have been used widely.[25,28–30] Here, we therefore consider substitutions at the codon level. A codon substitution can have two different outcomes for the nucleotide sequence of protein-coding region: synonymous substitution does not change the encoded sequence of amino acid residues, whereas non-synonymous substitution leads to changes in the amino acid residues. Random mutation and selection pressure will have different effects on the rate of these two types of substitutions,[31–33] and this difference can be exploited for detecting selective pressure at the protein level.[25,34–38] Our key problem is to find out the ratio of the synonymous substitution rate d_{s} and the non-synonymous substitution rate d_{n}. That is, we wish to estimate the ratio of the number of synonymous and non-synonymous substitutions at a specific site or a specific position of the amino acid residue. If natural selection offers no advantage, non-synonymous mutations will have the same rate as synonymous mutations ($d_{\mathrm{n}} = d_{\mathrm{s}}$), and the ratio $\omega = d_{\mathrm{n}}/d_{\mathrm{s}}$ will be 1. If non-synonymous mutations are harmful, deleterious or lethal, purifying selection is at play and the rate for non-synonymous mutation will be reduced: we have $d_{\mathrm{n}} < d_{\mathrm{s}}$ and $\omega < 1$. On the other hand, if Darwinian positive selection favors non-synonymous mutation, we have $d_{\mathrm{n}} > d_{\mathrm{s}}$ and $\omega > 1$. Here, ω is used as a measure of selection pressure. Substitution fixed by evolution and substitution fixed by chance are distinguished by examining the ratio ω at various locations of amino acid residues. This technique has been frequently applied in studies of molecular evolution, e.g. in detecting adaptive evolution.[22,38,39]

Continuous time Markov process for codon substitution

Markov model has been used widely in sequence analysis[40] and in evolutionary models.[23] In the current model, the outcome of codon substitution is determined only by the identity of codon in the ancestral sequence separated by divergence time t, and a codon transition probability matrix $\mathbf{P}(t)$. A phylogenetic tree is a key ingredient of this model. The topology and branch lengths of the tree reflect the evolutionary relationship among different proteins, which can model their closeness.[23] We follow the approach of Yang,[22] Nielsen & Yang,[26] and Yang *et al.*,[41] and briefly describe below the model.

For a given phylogenetic tree, the parameters of the evolutionary model are a 61 × 61 rate matrix **Q** for 61 non-stop codons and the sequence divergence time *t*s (or the branch lengths) of the phylogenetic tree. The divergence time represents expected number of changes between sequences, which are nodes in a phylogenetic tree. The entries q_{ij} of matrix **Q** are infinitesimal substitution rates of nucleotides for the set *C* of 61 non-stop codons, and they are parametrized as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion,} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition,} \end{cases}$$

where μ is the basis rate, κ, the transition/transversion rate ratio, ω, the ratio of non-synonymous and synonymous rates, and π_j is the codon frequency, which can be estimated as observed codon frequency in the sequences. In this model, the 61 × 61 rate matrix **Q** is fully determined by two parameters κ and ω, since π_j can be estimated and μ is a constant.[25,26]

For continuous time Markov process, the transition probability matrix of size 61 × 61 after time *t* is:[24]

$$\mathbf{P}(t) = \{p_{ij}(t)\} = \exp(\mathbf{Q}t)$$

The entry $p_{ij}(t)$ represents the probability that codon *i* will mutate into codon *j* after time *t*. It is calculated through diagonalization of matrix **Q**.

ω ratio from likelihood of phylogeny

For node *i* and node *j* in a phylogenetic tree separated by divergence time t_{ij}, the time reversible probability of observing nucleotide x_i in a position *h* at node *i* and nucleotide x_j of the same position at node *j* is:

$$\pi_{x_i} p_{x_i x_j}(t_{ij}) = \pi_{x_j} p_{x_j x_i}(t_{ij}) \qquad (1)$$

For a set *S* of *s* multiple-aligned sequences with *n* amino acid residues, we assume that a reasonably accurate phylogenetic tree $T = (\mathcal{V}, \mathcal{E})$ is given. Here $\mathcal{V}$ is the set of nodes (or vertices), namely, the union of the set of observed *s* sequences $\mathcal{L}$ (leaf nodes), and the set of $s-2$ ancestral sequences $\mathcal{I}$ (internal nodes). $\mathcal{E}$ is the set of edges (or branches) of the tree. Let the vector $\mathbf{x}_h = (x_1, \ldots, x_s)^{\mathrm{T}}$ be the observed codons at position *h* for the *s* sequences. Without loss of generality, we assume that the root of the phylogenetic tree is an internal node *k*. Given the specified topology of the phylogenetic tree **T** and the set of branch lengths (or divergence times), and if the set of codons $\mathcal{C}_{\mathcal{I}}$ of all internal nodes $\mathcal{I}$ is specified, the probability of observing the *s* number of codons $\mathbf{x}_h$ at position *h* is:

$$p(\mathbf{x}_h | \mathcal{C}_{\mathcal{I}}, \mathbf{T}) = \pi_{x_k} \prod_{(i,j)\in\varepsilon} p_{x_i x_j}(t_{ij})$$

Summing over the set $\mathcal{C}$ of all possible codons for the internal nodes $\mathcal{I}$, we have:

$$p(\mathbf{x}_h | \mathbf{T}) = \pi_k \sum_{\substack{i\in\mathcal{I} \\ x_i\in\mathcal{C}}} \prod_{(i,j)\in\varepsilon} p_{x_i x_j}(t_{ij}) \qquad (2)$$

The probability of observing all codons in the coding region of the nucleotide sequences is:

$$P(S|\mathbf{T}) = P(\mathbf{x}_1, \ldots, \mathbf{x}_s|\mathbf{T}) = \prod_{h=1}^{s} p(\mathbf{x}_h|\mathbf{T})$$

To account for the possibility that the rate of non-synonymous substitution can vary among different sites, the model developed[41] allows *M* possible different classes of non-synonymous substitutions with rates $\omega_1, \ldots, \omega_M$. Each amino acid site falls into the *M* class with probabilities $p_1, \ldots, p_M$.[41] The probability of observing $\mathbf{x}_h$ is then modified from equation (2), which gives $p(\mathbf{x}_h|\omega_m, \mathbf{T})$, to the following:

$$p(\mathbf{x}_h|\mathbf{T}) = \sum_{m=1}^{M} p_m \times p(\mathbf{x}_h|\omega_m, \mathbf{T})$$

Repeating this calculation over all amino acid residue sites, we have:

$$P(S|\mathbf{T}) = \prod_{h=1}^{s} p(\mathbf{x}_h|\mathbf{T})$$

and the likelihood function is:

$$\ell(T) = \sum_{h=1}^{s} \log[p(\mathbf{x}_h|\mathbf{T})]$$

To estimate the parameters κ_h, ω_h for each site *h* used in the mutation rate matrix **Q**, we use a maximum likelihood estimator,[26,37,42] the PAML package by Yang.[27] Our goal is to search for parameters κ_h and ω_h such that the likelihood function $\ell(\mathbf{T})$ is maximized. Here the number *M* of different classes of ω is 10, and they take the default values as assigned by PAML.[41]

Once the model parameters are estimated, the empirical Bayes approach can be used to infer the most likely class of ω value at each residue site.[22] In PAML, the posterior probability $p(\omega_m|\mathbf{x}_h)$ that site h with observed codons $\mathbf{x}_h$ is from class m with rate ratio ω_m is calculated as:

$$\begin{aligned} p(\omega_m|\mathbf{x}_h) &= p_m \times p(\mathbf{x}_h|\omega_m, \mathbf{T})/p(\mathbf{x}_h|\mathbf{T}) \\ &= p_m \times p(\mathbf{x}_h|\omega_m, \mathbf{T})/\sum_m p_m \times p(\mathbf{x}_h|\omega_m, \mathbf{T}) \end{aligned}$$

Data collection and computational procedures

We follow[7] and study evolution of the set of proteins taken from Table 1 of Mirny & Shakhnovich,[7] where the folding nucleus residues are defined. We first query with the sequence of each of the proteins against the HSSP database[43] to obtain homologous protein sequences with overall sequence identity $>30\%$ to ensure that they have the same fold. In some cases, we also searched the CE server[44] for structural homologs. Experimentation using PSI-BLAST searching of the NR-database of protein sequences give almost identical sets of sequences. Here, all redundant sequences are removed. Since paralogous sequences in a single species may exist that can be matched to the query DNA sequence, we only take the sequence with the highest identity to the query protein when multiple homologous sequences are found in a single species. With the exception of protein CI2 where sequences of two paralogs are included, only proteins with ≥ 5 known orthologous DNA sequences are kept. We therefore exclude AcP protein and CD2.d1 protein because fewer than five DNA sequences were found. Since paralogs are excluded, the number of sequences used here is smaller than that used in other studies.[7–9] The amino acid residue sequences of the remaining seven proteins are first aligned using CLUSTALW with default parameters[45] and then with manual intervention. Alignment of the nucleotide sequences is generated following the alignment of the protein sequences. A phylogenetic tree **T** is constructed using maximum likelihood method as implemented in the PAUP method.[46] This tree **T** is then used by the PAML package, an implementation of the maximum likelihood method for estimating ω values.[27] In many cases, minor difference in the tree does not affect final results significantly.[47,48] For each protein, we repeatedly estimate ω 20 times using different initial ω value that is assigned to all amino acid sites. The initial ω values range from 0.01 to 2.00, at an interval of 0.1. About 90% of the computation converges. For each protein, all different converged estimations among the 20 calculations give identical ω parameters at individual codon positions.

Natural selection at protein folding nucleus

The estimation of site-specific ω-values can uncover residues important for biological function, for structural stability, and potentially for folding kinetics. Here we focus on the natural selection of folding nucleus residues which are identified by ϕ-value experiments. An example for estimated ω values is shown in Figure 1.

We first examine the patterns of ω-ratio of non-synonymous *versus* synonymous substitutions in the seven proteins. If folding nucleus residues are more conserved than other residues, selection pressure then must be correlated with the extent of participation in folding nucleus.[9] Following Larson *et al.* we examine directly the correlation of the ϕ-values and the ω-values of characterized residues for each protein. This approach helps to circumvent the unavoidable arbitrariness in the assignment of the set of folding nucleus residues.[9,13] Residues with characterized ϕ-values for these proteins are obtained.[9] Following Plaxco

Table 1. The conservation and packing of folding nucleus residues

Protein	PDB	N_{prot}	N_{seq}	$N\phi$	R^2 [a]	p [a]	p_{all} [b]	$p_{50\%}$ [b]	$Z_{\alpha,\text{fn}}$ [c]	$Z_{\alpha,\text{all}}$ [c]
CI2	2ci2I	83	5	37	2.1×10^{-2}	0.39	3.3×10^{-1}	8.6×10^{-1}	3.29	2.81
Tenascin	1ten	2201	5	27	2.1×10^{-2}	0.47	4.1×10^{-2}	2.5×10^{-1}	3.29	3.44
CheY	3chy	128	7	30	9.8×10^{-2}	0.093	2.6×10^{-3}	8.2×10^{-2}	3.60	3.25
ADA2h	1aye	417	6	19	5.9×10^{-6}	0.99	4.3×10^{-1}	9.9×10^{-1}	3.43	2.78
U1A	1urn	282	12	10	2.2×10^{-1}	0.17	1.6×10^{-1}	9.3×10^{-1}	3.35	3.48
ACBP	1aca	86	16	22	5.2×10^{-3}	0.75	6.7×10^{-2}	6.3×10^{-1}	NMR	NMR
FKBP12	1fkj	107	27	22	6.7×10^{-4}	0.91	4.0×10^{-2}	3.6×10^{-1}	3.11	2.99

N_{prot} : number of residues in the protein sequence; N_{seq} : number of sequences; N_ϕ: number of residues with ϕ-value measured.

[a] Correlation of participation in folding nucleus as measured by ϕ-value and selection pressure as measured by ω. R^2: the fraction of variance in the data that can be explained by the linear regression model; p: the two-sided p-value of t-test for the null hypothesis that the slope of the linear regression models is 0.

[b] Randomization tests for assessing statistical significance of conservation of folding nucleus residues. The median ω value of the folding nucleus is tested against the distribution of the median ω value from 10^5 random samples containing the same number of amino acid residues as that of the folding nucleus drawn from the same protein. p_{all}: the p-value that the folding nucleus residues are more conserved than all other residues in the protein; $p_{50\%}$: the p-value that folding nucleus residues are more conserved than the residue at 50% quantile of all residues ranked by ω-value.

[c] Packing analysis of the folding nucleus and of the whole protein. The average alpha coordination number Z_α for all residues in the protein ($Z_{\alpha,\text{all}}$) and for residues in the folding nucleus residues ($Z_{\alpha,\text{fn}}$) are listed, except for structures determined by NMR techniques. Protein CheY has the highest $Z_{\alpha,\text{fn}}$.

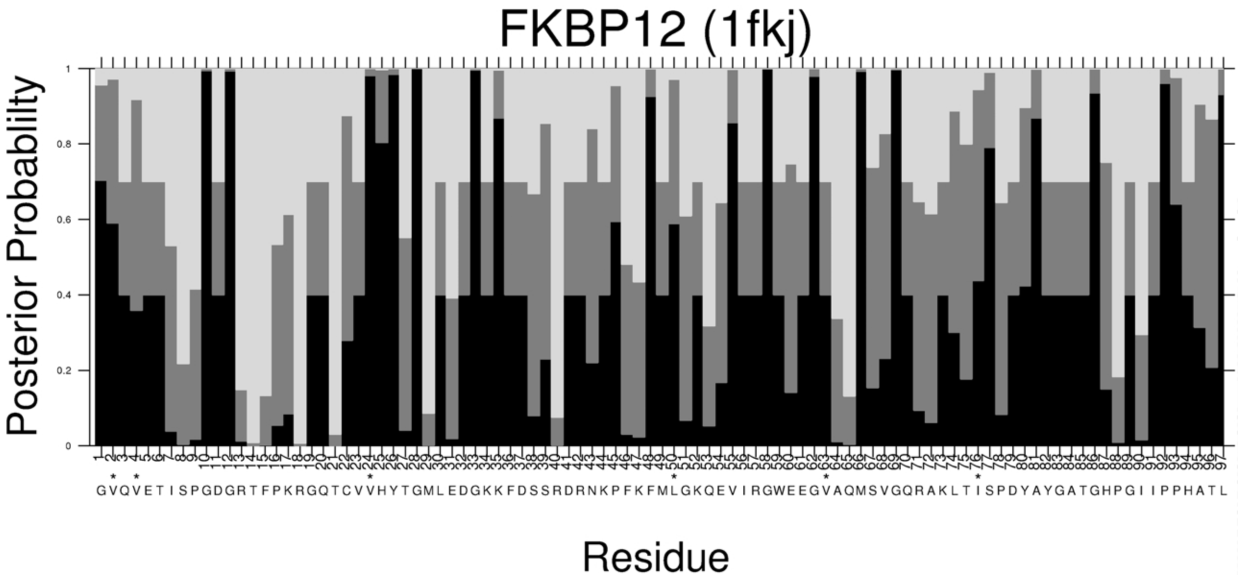


Figure 1. Selection pressure as measured by ω ratio of non-synonymous *versus* synonymous codon substitution rate varies at each amino acid residue site along the sequence of protein FKBP12. The ten possible ω values are grouped into three classes: $\omega_a < 0.12$ (dark), $0.12 \le \omega_b < 0.34$ (gray), and $0.34 < \omega_c$ (light). The x-axis shows the residue number of the protein, the y-axis shows the posterior probability of ω belonging to one of the three classes at each codon position. Residues with large probability for ω_a (dark) are highly conserved residues experiencing strong purifying pressure. Folding nucleus residues as identified by Mirny & Shakhnovich[7] are marked by the symbol "$*$".

et al.,[8] we exclude residues with $\phi < -0.5$ or $\phi > 1.5$, and require all ϕ-values to have standard deviation <1.0, with the exception of protein U1A (1urn), where no data of standard deviations are provided.

Among the set of residues with experimentally characterized ϕ-values, there is little correlation between ϕ-value and ω-value (Figure 2). The R^2 values range between 0.0 and 0.22, and the two-sided p-values of t-test for the null hypothesis that the slope of the linear regression models is 0 range from 9% to 99% (Table 1). That is, there is no indication of significant correlation between the extent of kinetic participation as measured by ϕ-value and selection pressure as measured by ω-value. Our results are similar to those found by Plaxco *et al.*,[8] and Larson *et al.*,[9] where relative entropy instead of ω was used as the measure of evolutionary conservation.

The weighted mean values of estimated ω ratio $\bar{\omega} = \sum_m p_m \omega_m$ at each codon position are plotted in Figure 3. It is clear that for each protein, many folding nucleus residues as defined by Mirny & Shakhnovich[7] have small values of ω, many are often smaller than the median ω-value of all codon positions. This indicates that folding nucleus residues experience purifying selection pressure. However, there are also many other residues with small ω-value, some of which have not been characterized by ϕ-value studies. As discussed,[9] the lower ω-values of folding nucleus as defined by Mirny & Shakhnovich[7] residues could also be a reflection of the experimental bias in choosing conserved protein core residues for ϕ-value experiments. Can we still conclude that experimentally identified folding nucleus residues in general are more conserved than the rest of the protein?

We use a randomization test following the approach first developed,[7] to address this question. The null hypothesis H_0 is that nucleus residues have equal or greater median ω values than that of the whole protein. That is, folding nucleus residues are no more conserved than the whole protein sequence. The alternative hypothesis H_a is that folding nucleus residues have less median ω values than the whole protein sequence and are evolutionarily more conserved. We calculate the median of ω values of the nucleus residues as defined by Mirny & Shakhnovich[7] and compare them with the distribution of median of ω value in random samples containing the same number of residues drawn from the same protein. As defined by Mirny & Shakhnovich[7] we use a sample size of 10^5. The fraction of the random samples with median ω value smaller than that of the folding nucleus provides the p-value that the observed median ω-values of the folding nucleus is due to random chance. Similarly,[7] we use the threshold of $p = 2\%$ to decide whether evolutionary conservation of the folding nucleus is statistically significant. Table 1 shows that p-value ranges between 0.26% (CheY) and 43% (ADA2h), but the majority are between 4.0% (FKBP12) and 43% (ADA2h). With the exception of CheY, the null hypothesis cannot be rejected with statistical significance at the confidence level of $p < 2\%$. That is, except CheY, folding nuclei as defined by Mirny & Shakhnovich[7] in these proteins are not significantly more conserved than the rest of the protein.

To further assess selection pressure on folding

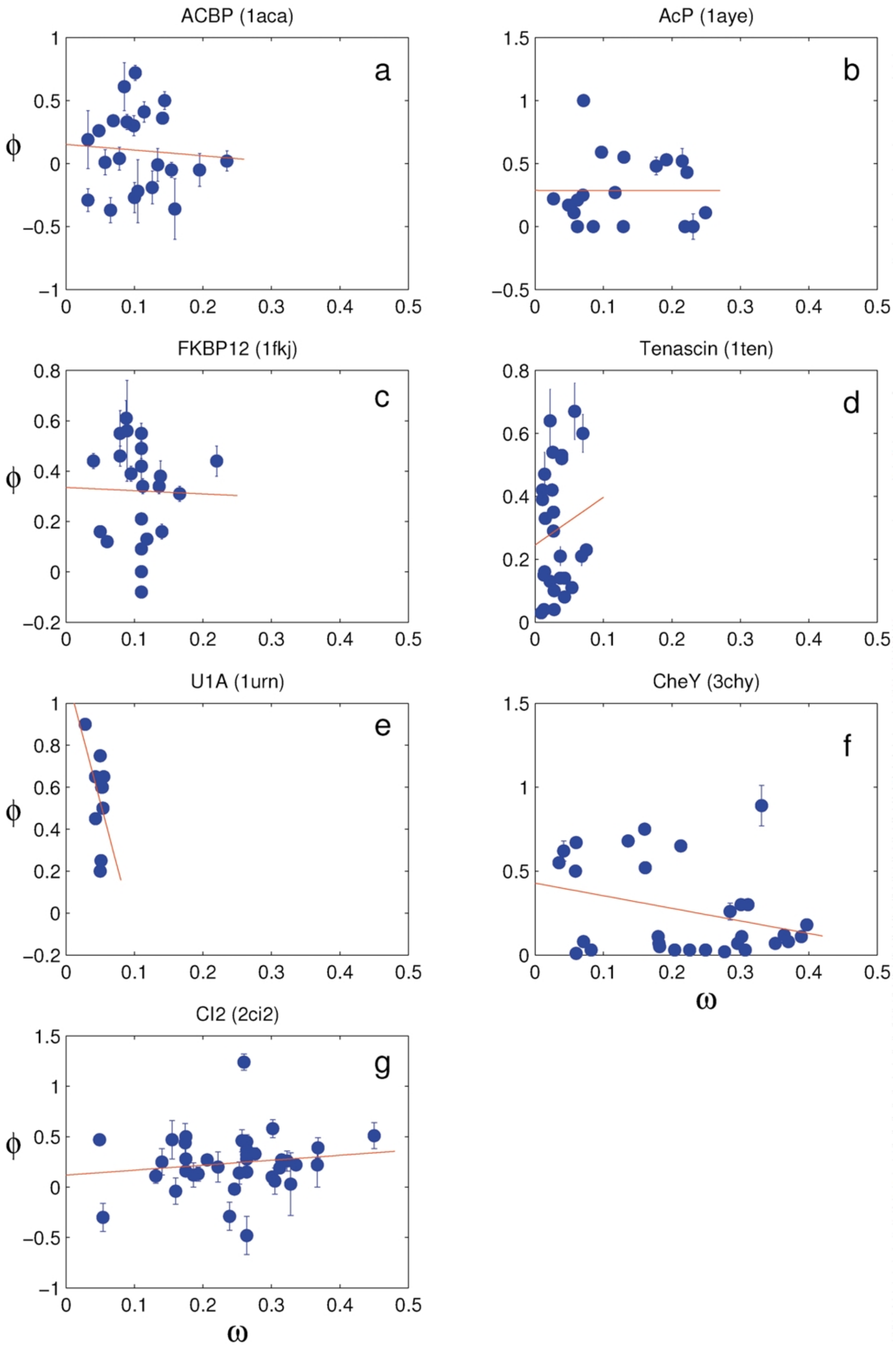


Figure 2. Participation in folding nucleus as measured by experimental ϕ-value and selective pressure as measured by ω-value are poorly correlated.

nucleus residues, we evaluate a different null hypothesis, again using randomization test. The null hypothesis H_0 now is that the folding nucleus residues have equal or greater median ω-values than the residue with median ω-value of the whole protein. That is, folding nucleus as defined by Mirny & Shakhnovich[7] are no more conserved than the residue halfway in the rank ordered list of all residues when sorted by estimated mean ω-value. Table 1 shows that the p-values range from 8.2% to 99%. With the criterion of $p < 2\%$, the null hypothesis cannot be rejected with statistical significance for any of the proteins. That is, folding nucleus for every protein studied here is not significantly more conserved than the residue with median ω-value.

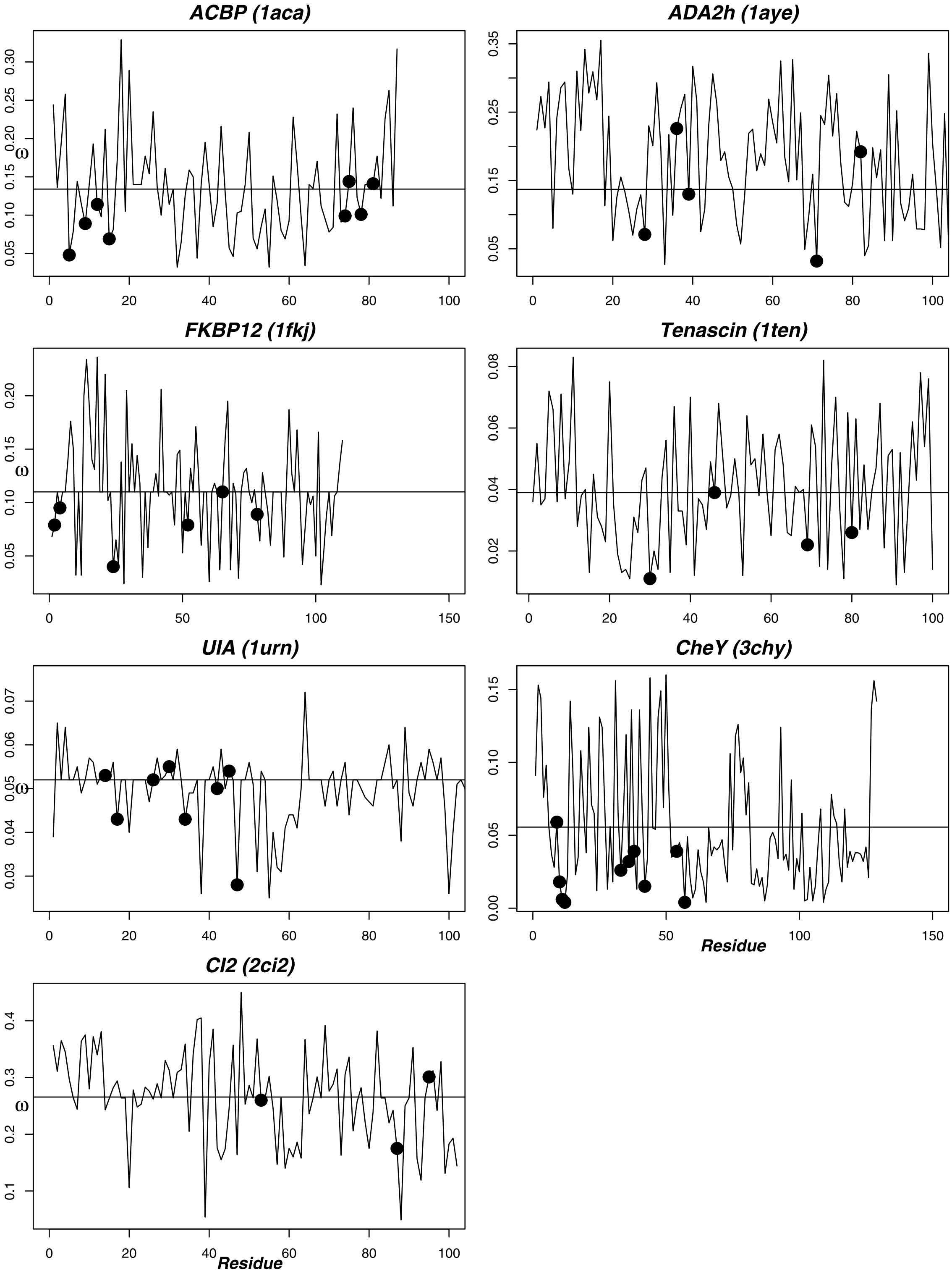


Figure 3. The weighted mean value $\bar{\omega} = \sum_{m=1}^{10} p_m \times \omega_m$ of estimated ω ratio at each residue position of the proteins. The x-axis shows the residue number of the protein, the y-axis shows the estimated $\bar{\omega}$ at each residue position. The horizontal line marks the median $\bar{\omega}$ value of all positions. Folding nucleus residues as identified by Mirny & Shakhnovich[7] are marked by ●. Except protein CheY, randomization tests show that folding nucleus residues are not more conserved than the rest of the protein, and in all cases (including CheY protein) are not more conserved than the residue at 50% quantile of all residues ranked by ω.

Conservation of folding nucleus of CheY

CheY is the only protein among those studied here that may have a well-conserved folding nucleus based on results of the first randomization test. Correlation study of ϕ-value and conservation measured by reduced entropy also suggested that CheY protein has a well-conserved folding nucleus.[9] What are the possible reasons for the strong conservation of folding nucleus in this protein? It was suggested earlier that tightly packed protein interior residues are well conserved and these are often part of the folding nucleus residues.[4,6,49] We use a parameter z_α recently introduced[50] to characterize protein local packing. z_α is defined as $z_\alpha \equiv n_c/n$, where n_c is the number of non-bonding atomic alpha contacts between different residues, and n is the total number of atoms. Two atoms are in alpha contact if they are separated by a weighted Voronoi facet which intersects with the protein.[50] z_α characterizes protein packing more faithfully than other parameters such as radius of gyration.[50]

We calculate z_α for the folding nucleus as defined by Mirny & Shakhnovich[7] and for the whole protein (Table 1). We find that the folding nucleus of CheY has the highest z_α value (3.60) compared to the folding nuclei of other proteins, whereas the whole protein z_α value of CheY has similar values to other proteins. This indicates that the folding nucleus of CheY has significantly larger $z\alpha$ than the rest of CheY protein. The folding nucleus of CheY is packed tighter than folding nuclei in other proteins. This observation can intuitively explain the significant conservation in CheY: tight packing in this case is accompanied by little tolerance to mutation, since the lack of packing defects such as voids reduces the possibility for substitution of different amino acid residues. However, this is a rather tentative hypothesis. It is possible that very tightly packed residues are more conserved, independent of whether they are in folding nucleus or not. It is also possible that if results of additional experimental ϕ-value studies become available, the definition of the folding nucleus might change. To fully resolve the relationship of packing, folding, and evolutionary conservation, more detailed additional studies are required, which is beyond the scope of this work.

Reconstructing ancestral folding nucleus

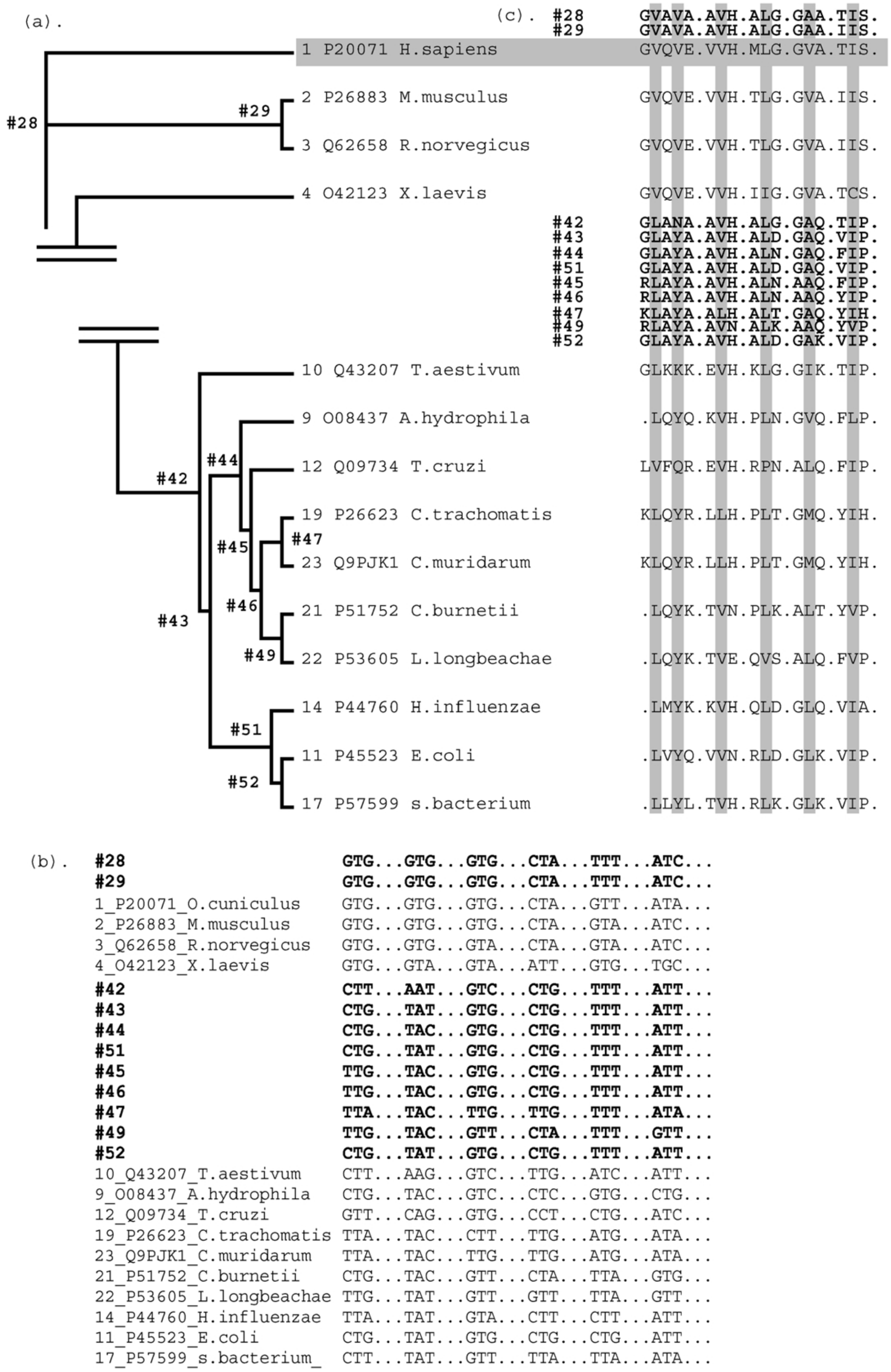
The approach used here can also suggest further experimental exploration of evolution history of protein folding dynamics. With the continuous time Markovian model, we can reconstruct likely candidate sequences of ancestral proteins at different evolutionary times. Specifically, identities of amino acid residues in the folding nucleus of ancient ancestral proteins can be postulated.

As an example, we show in Figure 4 the reconstructed residues of the folding nuclei of FKBP12 as defined by Mirny & Shakhnovich.[7] The six folding nucleus residues are VVVLVI in human FKBP12 protein. The first residue is L in some reconstructed ancestral genes, the second can be Y or N, the third can be L, the fifth can be A and the sixth can be a V instead of I. Based on this simple analysis, an interesting quadruplet mutagenesis study can be suggested to experimentally test the folding dynamics of mutated FKBP12, where the folding nucleus is changed. The reconstructed ancient folding nuclei suggests a combination of residues represented by the pattern $\mathbf{L}\{\mathbf{N},\mathbf{Y}\}\mathbf{L}\mathit{LA}\mathbf{V}$. Here $\{\mathbf{N},\mathbf{Y}\}$ means either a N or a Y residue is drawn.

The fourth residue L in all ancestral genes are the same as that in human FKBP12, but inspection of sequences of other extant species shows that the fourth residue can be any of I, P, or V, and the fifth can be any of I, V, L, and M. A further interesting experiment could be to test the folding behavior of 6-tuple mutants with folding nucleus formed by any combination of residues represented by the pattern $\mathbf{L}\{\mathbf{N},\mathbf{Y}\}\mathbf{L}\{\mathbf{AI},\mathbf{P},\mathbf{V}\}\{\mathbf{I},\mathbf{V},\mathbf{L},\mathbf{M}\}\mathbf{V}$. The recreated proteins then can be assayed for folding behavior, which can be compared with that of proteins present in extant organisms. Such experimental palaeobiochemistry was already envisioned by Pauling & Zuckerkandl many years ago,[51] and the number of such studies is rapidly growing.[52–57] An in-depth study on recreating the full sequence of ancestral proteins will require additional detailed analysis, including choosing the most appropriate detailed evolutionary model.[58–60]

Discussion

Although folding nucleus is under purifying pressure, we fail to observe significant conservation for protein folding nucleus residues. Despite concerns raised by Larson *et al.*[9] about the specific choices of the data by Mirny & Shakhnovich[7] we use exactly the same set of proteins, the same definition of nuclei residues, and follow the same randomization test as that of Mirny & Shakhnovich.[7] It is possible that this would bias our study towards reproducing the results of Mirny & Shakhnovich.[7] Nevertheless, our results are similar to that of Plaxco *et al.* and Larson *et al.*,[8,9] and are different from that of Mirny & Shakhnovich.[7] The different conclusion of this study and that of Mirny & Shakhnovich[7] is likely due to the different evolutionary models employed, namely, the difference between a DNA-codon based continuous-time Markov model *versus* an implicit evolution model implied by entropy calculation. The conclusion that folding nuclei residues are not conserved will likely remain if we were to use the data set and the definitions of folding nuclei by Larson *et al.*[9] Experimental studies in barnase, SH3 domain, chymotrypsin inhibitor 2 suggest that the folding nucleus observed in wild-type protein may not be indispensable, and alternative folding nucleus may arise if residues are mutated.[61–65] Another experimental example is Im9 and Im7 proteins. They are E



```
#28                     GVAVA.AVH.ALG.GAA.TIS.
#29                     GVAVA.AVH.ALG.GAA.IIS.
1 P20071 H.sapiens      GVQVE.VVH.MLG.GVA.TIS.
2 P26883 M.musculus     GVQVE.VVH.TLG.GVA.IIS.
3 Q62658 R.norvegicus   GVQVE.VVH.TLG.GVA.IIS.
4 O42123 X.laevis       GVQVE.VVH.IIG.GVA.TCS.
#42                     GLANA.AVH.ALG.GAQ.TIP.
#43                     GLAYA.AVH.ALD.GAQ.VIP.
#44                     GLAYA.AVH.ALN.GAQ.FIP.
#51                     GLAYA.AVH.ALD.GAQ.VIP.
#45                     RLAYA.AVH.ALN.AAQ.FIP.
#46                     RLAYA.AVH.ALN.AAQ.YIP.
#47                     KLAYA.ALH.ALT.GAQ.YIH.
#49                     RLAYA.AVN.ALK.AAQ.YVP.
#52                     GLAYA.AVH.ALD.GAK.VIP.
10 Q43207 T.aestivum    GLKKK.EVH.KLG.GIK.TIP.
9 O08437 A.hydrophila   .LQYQ.KVH.PLN.GVQ.FLP.
12 Q09734 T.cruzi       LVFQR.EVH.RPN.ALQ.FIP.
19 P26623 C.trachomatis KLQYR.LLH.PLT.GMQ.YIH.
23 Q9PJK1 C.muridarum   KLQYR.LLH.PLT.GMQ.YIH.
21 P51752 C.burnetii    .LQYK.TVN.PLK.ALT.YVP.
22 P53605 L.longbeachae .LQYK.TVE.QVS.ALQ.FVP.
14 P44760 H.influenzae  .LMYK.KVH.QLD.GLQ.VIA.
11 P45523 E.coli        .LVYQ.VVN.RLD.GLK.VIP.
17 P57599 s.bacterium   .LLYL.TVH.RLK.GLK.VIP.
```

```
#28                      GTG...GTG...GTG...CTA...TTT...ATC...
#29                      GTG...GTG...GTG...CTA...TTT...ATC...
1_P20071_O.cuniculus     GTG...GTG...GTG...CTA...GTT...ATA...
2_P26883_M.musculus      GTG...GTG...GTG...CTA...GTA...ATC...
3_Q62658_R.norvegicus    GTG...GTG...GTA...CTA...GTA...ATC...
4_O42123_X.laevis        GTG...GTA...GTA...ATT...GTG...TGC...
#42                      CTT...AAT...GTC...CTG...TTT...ATT...
#43                      CTG...TAT...GTG...CTG...TTT...ATT...
#44                      CTG...TAC...GTG...CTG...TTT...ATT...
#51                      CTG...TAT...GTG...CTG...TTT...ATT...
#45                      TTG...TAC...GTG...CTG...TTT...ATT...
#46                      TTG...TAC...GTG...CTG...TTT...ATT...
#47                      TTA...TAC...TTG...TTG...TTT...ATA...
#49                      TTG...TAC...GTT...CTA...TTT...GTT...
#52                      CTG...TAT...GTG...CTG...TTT...ATT...
10_Q43207_T.aestivum     CTT...AAG...GTC...TTG...ATC...ATT...
9_O08437_A.hydrophila    CTG...TAC...GTC...CTC...GTG...CTG...
12_Q09734_T.cruzi        GTT...CAG...GTG...CCT...CTG...ATC...
19_P26623_C.trachomatis  TTA...TAC...CTT...TTG...ATG...ATA...
23_Q9PJK1_C.muridarum    TTA...TAC...TTG...TTG...ATG...ATA...
21_P51752_C.burnetii     CTG...TAC...GTT...CTA...TTA...GTG...
22_P53605_L.longbeachae  TTG...TAT...GTT...GTT...TTA...GTT...
14_P44760_H.influenzae   TTA...TAT...GTA...CTT...CTT...ATT...
11_P45523_E.coli         CTG...TAT...GTG...CTG...CTG...ATT...
17_P57599_s.bacterium_   CTT...TAT...GTT...TTA...TTA...ATA...
```

Figure 4. Reconstructed ancestral protein sequences of FKBP12 protein. (a) The relevant part of the phylogenetic tree for FKBP12 is shown. Human FKBP12 protein from which experimental data were obtained is shown in shadow. (b) Multiple alignment of DNA sequences of the folding nucleus of FKBP12 protein, including those of reconstructed folding nucleus of ancestral proteins. (c) Multiple alignment of translated amino acid residue of the folding nucleus residues identified by ϕ-value studies (highlighted) and flanking residues.

colicin-binding immunity proteins that are of the same fold with about 60% sequence identity. The folding of Im9 and Im7 are two-state and three-state process, respectively. Although these two proteins have similar folding mechanism, ϕ-value studies reveal that the kinetically important residues are different.[66,67] This is consistent with recent simulation studies which suggest that evolution selection is more robust for residues important for stability than for kinetic accessibility.[68,69] In addition, the definition of a folding nucleus is arbitrary, because it is based on a threshold of ϕ value (e.g. $\phi \geq 0.5$).[13] An earlier study suggested that the critical nucleus may be as large as 10^2 residues, the size of a whole protein domain.[70] The non-uniqueness of folding nucleus was pointed out in a study using an off-lattice model system.[71] The role of protein structure in folding is discussed from the viewpoint of small-world connections.[72] Recent computational studies based on exact enumerable lattice models using master equation showed that there are remarkable heterogeneity in structural contacts underlying macroscopic two-state folding kinetics of model Gō protein.[20,21] The kinetic barrier was shown to result from a reduced number of microroutes near the bottom of the folding funnel.[20,21] If these studies portray accurately the microscopic picture of the folding process, there are likely to be many different native contacts that form folding nuclei for different folding pathways in the free energy landscape. It is reasonable to expect that a large subset of residues is capable of providing critical native contacts, and these contacts vary for different microscopic folding pathways. The roles of these residues in folding are largely interchangeable, and this may be reflected in the lack of extraordinarily strong purifying selection pressure in the current set of folding nucleus residues characterized by ϕ-value studies.

In summary, we use a continuous time Markovian model[25] and apply a maximum likelihood estimator developed[27] to study the evolution of protein-folding dynamics. We examine the coding DNA sequences rather than amino acid residue sequences, and assess selection pressure by estimating the ratio ω of non-synonymous *versus* synonymous codon substitution rate. The position specific rate ratio is used to distinguish substitutions fixed by evolution and by chance. We found that folding nucleus residues experience purifying selection pressure, but they are not significantly more conserved than the rest of the residues of the whole protein. The only exception is CheY protein, where the folding nucleus is significantly more conserved. This may be due to extraordinarily tight packing, which is reflected by the high alpha coordination number Z_α. Results described here provide another confirmation that evolution does not preserve kinetically important residues, which has been a subject of debate in literature.[7–9] We further suggest exploratory palaeobiochemical studies testing the evolution of protein-folding dynamics.

Acknowledgements

We thank Drs Xun Gu, Robie Mason-Gamer, and Clare Woodward for helpful discussions, Dr Ziheng Yang for generous help in using PAML. This work is supported by grants from National Science Foundation (CAREER DBI0133856 and DBI0078270) and National Institute of Health (GM68958).

References

1. Shrivastava, I., Vishveshwara, S., Cieplak, M., Maritan, A. & Banavar, J. R. (1995). Lattice model for rapidly folding protein-like heteropolymers. *Proc. Natl Acad. Sci. USA*, **92**, 9206–9209.
2. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–98.
3. Mirny, L. A., Abkevish, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976–4981.
4. 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.
5. Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in in protein superfamilies. *Fold. Des.* **3**, 239–251.
6. Ptitsyn, O. B. & Ting, K-L. H. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* **291**, 671–682.
7. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
8. Plaxco, K. W., Riddle, D. S., Larson, S., Ruczinski, I., Thayer, E. C., Buchwitz, B. *et al.* (2000). Evolutionary conservation and protein folding kinetics. *J. Mol. Biol.* **298**, 303–312.
9. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxo, K. W. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225–233.
10. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the *src* SH3 domain. *Nature Struct. Biol.* **5**, 714–720.
11. Fulton, K. F., Main, E. R. G., Dagget, V. & Jackson, S. E. (2000). Mapping the interactons present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445–461.
12. Demirel, M. C., Atilgan, A. R., Jernigan, R. L., Erman, B. & Bahar, I. (1998). Identification of kinetically hot residues in proteins. *Protein Sci.* **7**, 2522–2532.
13. Fersht, A. R. (1997). Nucleation mechanism in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9.
14. Matouschek, A., Kellis, J. T., Jr, Serrano, L. & Fersht, A. R. (1990). Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, **346**, 440–445.
15. Matouschek, A. & Fersht, A. R. (1991). Protein engineering in analysis of protein folding pathways and stability. *Methods Enzymol.* **202**, 82–112.
16. Alm, E. & Baker, D. (1999). Prediction of

protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305–11310.

17. Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299–11304.
18. Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
19. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature*, **409**, 641–645.
20. Ozkan, S. B., Bahar, I. & Dill, K. A. (2001). Transition states and the meaning of (-values in protein folding kinetics. *Nature Struct. Biol.* **8**, 765–769.
21. Ozkan, S. B., Dill, K. A. & Bahar, I. (2002). Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.* **11**, 1958–1970.
22. Yang, Z. (2001). *Handbook of Statistical Genetics*, Wiley, New York Chapter 12..
23. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). *Molecular Systematics*, Sinauer, Sunderland, MA.
24. Liò, P. & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Res.* **8**, 1223–1244.
25. Goldman, N. & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.
26. Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.
27. Yang, Z. (1997). PAML: a program for package for phylogenetic analysis by maximum likelihood. *CABIOS*, **15**, 555–556.
28. Schöniger, M., Hofacker, G. L. & Borstnik, B. (1990). Stochastic traits of molecular evolution-acceptance of point mutations in native actin genes. *J. Theoret. Biol.* **143**, 287–306.
29. Muse, S. V. & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **13**, 105–114.
30. Yang, Z., Nielsen, R. & Hasegawa, M. (1998). Models of amino acid substitutions and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611.
31. Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge.
32. Gillespie, J. H. (1994). *The Causes of Molecular Evolution*, Oxford University Press, Oxford.
33. Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Mol. Biol. Evol.* **11**, 715–724.
34. Hughes, A. L. & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.
35. Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99.
36. Messier, W. & Stewart, C. B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, **385**, 151–154.
37. Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573.
38. Yang, Z. & Nielsen, R. (1998). Synonymous and non-synonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**, 409–418.
39. Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. (2000). Positive darwinian selection in the evolution of mammalian female reproductive proteins. *Proc. Natl Acad. Sci. USA*, **98**, 2509–2514.
40. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
41. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.
42. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.
43. Sander, C. & Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–58.
44. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
45. Thompson, J. D., Higgins, D. G. & Gibson, T. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
46. Swofford, D. L. (2002). *PAUP: Phylogenetic Analysis Using Parsimony and Other Methods*, Sinauer, Sunderland, MA.
47. Ford, M. J. (2001). Molecular evolution of transferrin: evidence for positive selection in salmonids. *Mol. Biol. Evol.* **18**, 639–647.
48. Yang, Z. & Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19**, 49–57.
49. Privalov, P. L. (1996). Intermediate states in protein folding. *J. Mol. Biol.* **258**, 707–725.
50. Zhang, J., Chen, R., Tang, C. & Liang, J. (2003). Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. *J. Chem. Phys.* **118**, 6102–6109.
51. Pauling, L. & Zuckerkandl, E. (1963). Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.* **17**, S9–S16.
52. Golding, G. B. & Dean, A. M. (1998). The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**, 355–369.
53. Chang, B. S. & Donoghue, M. J. (2000). Recreating ancestral proteins. *Trends Ecol. Evol.* **15**, 109–114.
54. Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H. & Hutchison, C. A., III (1994). Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl Acad. Sci. USA*, **91**, 1569–1573.
55. Jermann, T. M., Opitz, J. G., Stackhouse, J. & Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, **374**, 57–59.
56. Chandrasekharan, U. M., Sanker, S., Glynias, M. J.,

Karnik, S. S. & Husain, A. (1996). Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science*, **271**, 502–505.

57. Dean, A. M. & Golding, G. B. (1997). Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proc. Natl Acad. Sci. USA*, **94**, 3104–3309.
58. Schulter, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, **51**, 1699–1712.
59. Cunningham, C. W., Omland, K. E. & Oakley, T. H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.* **13**, 361–366.
60. Zhang, J. & Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44**, S139–S146.
61. Matthews, J. M. & Fersht, A. R. (1995). Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry*, **34**, 6805–6814.
62. Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874–880.
63. Viguera, A. R. & Serrano, L. (2002). Unspecific hydrophobic stabilization of folding transition states. *Proc. Natl Acad. Sci. USA*, **99**, 5349–5354.
64. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
65. Neira, J. L., Davis, B., Ladurner, A. G., Buckle, A. M., Gay Gde, P. & Fersht, A. R. (1996). Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. direct evidence for a nucleation-condensation mechanism. *Fold. Des.* **1**, 189–208.
66. Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2002). Im7 folding mechanism: misfolding on a path to the native state. *Nature Struct. Biol.* **9**, 209–216.
67. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003). Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* **326**, 293–305.
68. Dokholyan, N. V. & Shakhnovich, E. I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* **312**, 289–307.
69. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
70. Bryngelson, J. D. & Wolynes, P. G. (1990). A simple statistical field-theory of heteropolymer collapse with application to protein folding. *Biopolymers*, **30**, 1–2.
71. Guo, Z. & Thirumalai, D. (1996). The nucleation-collpse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Fold. Des.* **2**, 377–391.
72. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.*, 061910.

Edited by C. R. Matthews

(Received 3 September 2003; received in revised form 24 October 2003; accepted 5 November 2003)