Universal distribution of protein evolution rates as a consequence of protein folding physics

Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Edited* by William A. Eaton, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, and approved December 24, 2009 (received for review September 14, 2009)

The hypothesis that folding robustness is the primary determinant of the evolution rate of proteins is explored using a coarse-grained off-lattice model. The simplicity of the model allows rapid computation of the folding probability of a sequence to any folded conformation. For each robust folder, the network of sequences that share its native structure is identified. The fitness of a sequence is postulated to be a simple function of the number of misfolded molecules that have to be produced to reach a characteristic protein abundance. After fixation probabilities of mutants are computed under a simple population dynamics model, a Markov chain on the fold network is constructed, and the fold-averaged evolution rate is computed. The distribution of the logarithm of the evolution rates across distinct networks exhibits a peak with a long tail on the low rate side and resembles the universal empirical distribution of the evolutionary rates more closely than either distribution resembles the log-normal distribution. The results suggest that the universal distribution of the evolutionary rates of protein-coding genes is a direct consequence of the basic physics of protein folding.

correct folding probability | evolutionary rate distribution | off-lattice models | common structure networks | network average evolution rate

Each protein-coding gene evolves at a characteristic rate (molecular clock), although the clock is substantially overdispersed (1, 2). However, evolutionary rates (ER) across the entire set of protein-coding genes in an evolving organismal lineage span 3 to 4 orders of magnitude (1). Strikingly, the shape of the ER distribution across the sets of orthologous genes remains nearly the same (approximately log-normal) across the entire diversity of cellular life forms, from bacteria and archaea to mammals (3, 4).

The identification of the features of proteins that determine their characteristic ER and the factors behind the universality of the ER distribution is a major unsolved problem in evolutionary biology. For approximately 40 years after the broad variability of ER was discovered, it was generally assumed that the ER depends on some combination of the specific functional constraints affecting a protein and its general importance for the survival of the respective organism (5, 6). As intuitively appealing as it might be, this "functional hypothesis" seems to be poorly compatible with the results of genome-wide studies on correlates of the ER. Indeed, there is very little correlation between functional characteristics of genes or their biologic importance, measured through the knockout effect, and the ER (7–10). In contrast, gene expression level and protein abundance, characteristics that do not seem to be directly related to the specific functions or biologic importance of a gene, show a consistent and strong negative correlation with the ER (11–15). These findings prompted the development of the mistranslation-induced misfolding (MIM) hypothesis, according to which mRNA translation incurs a cost proportional to the number of misfolded protein molecules, which are not only a waste but are often toxic to the cell (15, 16). The cost is particularly high for the most abundant proteins, hence strong selection for folding robustness is deemed to be the primary determinant of the ER (13–15).

The universality of the ER distribution of protein-coding genes, together with the equally universal negative correlation between ER and protein abundance, suggest a purely mechanistic (devoid of references to function) and fundamental (common to all proteins in all life forms) explanation of the source of the ER variability. Here, we explore the implications of the idea that folding robustness is the sole determinant of proteins fitness, for the ER distribution in the context of simple yet diverse sequence and structure spaces derived from a simple off-lattice folding model. Many important results on protein folding (17–19) and evolution (20, 21) have been obtained using coarse-grained lattice and off-lattice models of proteins. Off-lattice models manifest a higher level of complexity and are therefore more appropriate for the study of protein evolution (22–29).

Here we show that, under the robustness-rate assumption, our model closely reproduces the empirically observed ER distribution as well as the relationship between the ER and translation rate. The results suggest that the universal ER distribution is a direct consequence of fundamental principles of protein folding.

Results and Discussion

Rationale and Outline. Our goal is to compute the ER with as few assumptions as possible. We postulate that a protein's fitness is a linear function of the number of misfolded copies produced to reach a required abundance. To exploit this idea, it is necessary to compute the probability that a given sequence folds to a particular structure and to construct a network of sequences that can fold to this structure (30, 31). The fitness of each sequence in the network is derived from its folding robustness, so the theory of population dynamics (32, 33) can be used to assign transition rates (fixation probabilities) among the members of the network connected by point substitutions. The transition rates can be used to compute the instantaneous ER of the network members as well as the steady-state, network-averaged ER.

The key aspect of this approach is the computation of the correct folding probability (CFP) for an arbitrary sequence. We constructed a simple model in which the CFP could be computed efficiently. We do not aim at a truly realistic representation of proteins but seek to create sufficiently rich and diverse sequence and structure spaces in the hope that the precise architecture of these spaces is not critical for the qualitative features of the ER distribution.

Folding Model and CFP. To compute folding probabilities, we chose a coarse-grained off-lattice model in which the protein is represented by a flexible chain with pairwise interactions between

Author contributions: A.E.L., Y.I.W., and E.V.K. designed research; A.E.L. performed research; A.E.L., Y.I.W., and E.V.K. analyzed data; and A.E.L. and E.V.K. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1] To whom correspondence should be addressed. E-mail: koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0910445107/DCSupplemental.

monomers (34, 35). Clementi et al. (24) explored interaction potentials and designed sequences for reaching a known native state. Alm and Baker (36) found that folding mechanisms are relatively insensitive to the details of the model. This conclusion is further supported by the studies of the folding funnel (37–39). Simplified models have elucidated a wide range of phenomena from the fundamental aspects of protein folding, such as the shape of protein folding landscapes, the nature of the transition state ensemble (39–41), and the importance of "gatekeeper" residues for misfolding and aggregation (42), to the folding mechanism of a large multidomain protein with complex topology (43).

Model sequences consist of four monomer types: hydrophobic (H), polar (P), positively charged (+) and negatively charged (−). Each monomer represents a run several amino acids long corresponding to the persistence length of the protein. Therefore, the angles between successive bonds are unrestricted, and there is no bending rigidity. Steric repulsion and entropic effects within the region described by our coarse-grained monomers are modeled via a spring between pairs of nearest neighbor monomers of rest length a and spring constant proportional to temperature. Next nearest neighbors and beyond interact via a pairwise potential U_{ij}, which consists of a soft-core repulsion (to model excluded volume interactions), a long-range attraction or repulsion when appropriate, and a screened Coulomb interaction between charged monomers:

$$U_{ij} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} + \frac{q_i q_j e^{-Dr_{ij}}}{r_{ij}}, \qquad [1]$$

where r_{ij} is the distance between monomers i and j, q_i is the monomer's charge, and D is the Debye-Hückel screening length. Coefficients A_{ij} and C_{ij}, which set the strength of the soft-core repulsion and the long-range attraction or repulsion, respectively, are selected separately for all 10 pair types. To emulate the effects of solvent, we assign a stronger attraction to HH pairs than to pairs that involve a P, +, or −. The precise choices of parameters make little difference in the statistics of the ER (see *SI Materials and Methods* for details). The dynamics of the flexible chain are implemented via a Brownian dynamics method, which is appropriate when inertial and hydrodynamic effects can be neglected (44).

The native structure of a sequence and its CFP are computed starting with an ensemble of high-temperature configurations and rapidly quenching each unfolded configuration to a temperature below the folding transition. The resulting folded configuration is held at the low temperature to exclude the possibility of a low barrier and then quenched to zero temperature to find the exact energy minimum and the corresponding structure.

The folded structures are compared using a 3D alignment, with rotations and reflections taken into account. When an alignment score is smaller than a certain threshold (see *Materials and Methods* for details), the two structures are considered identical. The folded configuration that is attained most often is taken as the native structure of a sequence. Its CFP is the fraction of the unfolded configurations that fold to the native structure.

This procedure for finding the native structure does not conform to Anfinsen's hypothesis (45), according to which the native structure is the global minimum of free energy. Instead, the native structure is selected via a rapid quench procedure that finds the folded state with the largest basin of attraction in configuration space (46). The kinetically selected native state must be separated by an energy barrier from the rest of the configuration space to prevent spontaneous structural transitions during the quench. The justification of the rapid quench selection of the native state is 2-fold. First, this method is computationally feasible in an off-lattice model whereby the native structure is not known a priori. Second, rapid protein folding is possible only when the folded state can be reached kinetically from a wide range of starting configurations. The presence of a barrier is also required for a reasonable lifetime of the folded state. Although, under kinetic control, sequence–structure combinations might evolve toward greater stability (47), the native state is not required to be the global energy minimum for the rapid folding and the reasonable lifetime conditions to be met. We explore the violation of the Anfinsen conjecture in *SI Materials and Methods*.

Robust Folders. Equipped with a method to compute native structures and the CFP, we explored the distribution of CFPs among all sequences of length N. The end monomers are fixed to be of the + and − types to model, respectively, the charged N- and C-termini of proteins, thus leaving $N-2$ changeable monomers. When N is small enough, an exhaustive sampling of the sequence space is possible. Fig. 1 shows the distribution of the logarithmic ratio of the probability CFP of reaching the native state and the misfolding probability 1 − CFP. If the polymer were in thermal equilibrium, this ratio would be proportional to the free energy of folding ΔG_{fold}. In our case, it is merely a convenient way of plotting the distribution of CFPs. The features of the distribution depend weakly on the details of the folding model, such as the number of monomer types, length of the chain, and the range of repulsive interactions. In the remainder of this work we use parameters that yield more robust folders. As we demonstrate in *SI Materials and Methods*, this selection does not seem to affect the distribution of ER qualitatively. For the chosen folding model (HP+− with long-range repulsion), the total number of robust folders (sequences with CFP >0.8) does not seem to decline with N. Because for $N > 12$, exhaustive sampling is not feasible, we find robust folders using a simulated annealing search (ref. 48 and references therein) (see *Materials and Methods* for details).

Common Structure Networks. To construct common structure networks of sequences connected by point substitutions (30), we start with a robust folder and examine all sequences that differ by a single substitution. If the mutant has a probability >0.2 of folding to the same structure it is added to the network, and its mutants are examined in turn. This process is repeated until all sequences that fold with probability >0.2 to the original structure are found. It turns out that some robust folders and even some unconnected networks share the same structure. Table 1 summarizes the diversity of robust folders, their structures, and

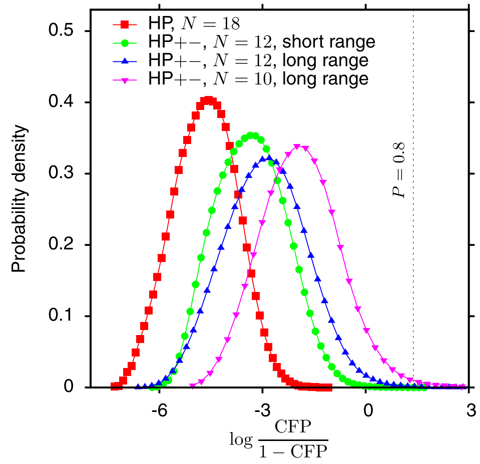

Fig. 1. Distribution (probability density function) of the logarithmic ratio of the probability of a sequence to fold into the native structure and the probability to fold to an alternative structure. See *SI Materials and Methods* for the values of parameters used to compute these distributions.

Table 1. Classification of robust folders (CFP >0.8) by native structure and by membership in connected networks

Sequence length N	Robust folders	Distinct structures	Distinct networks	Networks with multiple members
12	423	264	304	171
15	310	146	229	114
18	833	90	145	134

associated networks. The decline with N of the number of robust folders with distinct native structures can be attributed to the increasing size of the sequence space. *SI Materials and Methods* indicates that small networks may be missing from our $N = 18$ sample. Not surprisingly, it is more difficult to find a small network with the simulated annealing search. The statistics of the architecture of the common structure networks of robust folders are given in the *SI Materials and Methods*.

Misfolding and Fitness. A number of protein fitness criteria have been proposed: thermodynamic stability (49–51), foldability (52), folding rate (53), exposure of functional residues (54), and the strength of ligand binding (55). In contrast, we select the folding robustness as the primary determinant of fitness. Because the fitness of an organism is a function of the total number of misfolded protein molecules (of all expressed proteins), the impact of a single protein's expression on the organism's fitness varies linearly with the number of misfolded copies of this particular protein. Let Q be the misfolding probability in a single translation and folding event. Then the average number M of misfolded copies produced to reach an abundance level A of correctly folded protein molecules is $M = AQ/(1 - Q)$. We postulate that the fitness of a protein is:

$$w = \quad -kM \qquad [2]$$

We set $k = 1$ without loss of generality. A different k is equivalent to rescaling the abundance level by a constant factor. Different relationships between w and M do not affect the ER distribution qualitatively (see *SI Materials and Methods* for details).

There are two misfolding pathways (16): first, a correctly translated protein can misfold, and second, the protein can be mistranslated and the mistranslated sequences can misfold with the probability that depends on the type and position of the misincorporated monomer(s). Here we assume that there is a fixed per-monomer mistranslation probability r and that all mistranslations are equally likely. The misfolding probability due to mistranslation is then the sum over all possible mistranslation events of the product of the mistranslation event probability and the misfolding probability of the mistranslated sequence. When r is small, misfolding of the correctly translated sequence is the dominant source of misfolding; conversely, when r is large, misfolding of mistranslated sequences dominates. Therefore, for each sequence, there exists a characteristic per-monomer mistranslation probability R_{eq} for which the contributions of the two sources of misfolding probability are equal. The distribution of R_{eq} for our model (presented in *SI Materials and Methods*) peaks at $r \approx 0.015$. We chose a range of r values centered on 0.015 so that both the case when misfolding of the correctly translated sequences dominates and the opposite situation when mistranslation-induced misfolding is most important are covered. *SI Materials and Methods* presents evidence that the composition of the pool of misfolded proteins does not qualitatively affect the ER distribution.

ER Distribution: Comparison of the Model and Comparative-Genomic Results. We assume that evolution proceeds by fixation attempts of random point substitutions (other types of mutations such as insertions, deletions, and recombinations are ignored). Mutations between the network members can be fixed, whereas mutations that take the sequence out of the network are eliminated. Fixation probabilities are assigned using the theory of population dynamics. When the mutation rate is small so that at most a single mutant is present in a population of effective size N_e, the fixation probability of a sequence j that arises in the sea of i is (1)

$$\pi(i \to j) = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}}, \qquad [3]$$

where $s = w_j - w_i$. Note that this formulation is equivalent to the one used by Wilke and Drummond (56).

Evolution is a Markov process on the common structure network governed by the transition probabilities between the network members. Two measures of the ER can be defined. Instantaneous ER, defined for each sequence in the network is the probability that a random substitution in that sequence is accepted. Instantaneous ER varies substantially across the network. Fold-averaged ER is the average of the instantaneous ER over all sequences in the network weighted by the probability of finding each sequence in the stationary ensemble. Fold-averaged ER is equivalent to the long time average of the instantaneous ER and accordingly is a more robust measure of the characteristic rate of evolution of a given fold. The stationary distribution of the Markov process is found from the left eigenvector of the transition probability matrix with unit eigenvalue (57). Other fold-averaged properties, such as the fold-averaged CFP, can be computed in a similar fashion.

Because fitness is proportional to the required abundance A, it serves as the evolutionary pressure. When A is so small that the typical fitness gap between sequences is smaller than the inverse effective population size, the representation of each sequence in the stationary ensemble is determined solely by the topology of the network (30), with highly connected sequences having higher representation. In the limit of large A, only the fittest sequence (one with the highest CFP) will have an appreciable representation in the stationary ensemble and therefore will determine the fold-averaged ER and fold-averaged CFP.

A typical distribution of the fold-averaged ER across distinct networks with more than two members is shown in Fig. 2. The distribution of the logarithm of the ER has an asymmetric peak and a long tail on the low ER side. As we show in *SI Materials and Methods*, the location of the peak scales as the fixation probability of a neutral mutation. Thus, in contrast to the mean and the median of the distribution, which scale linearly with N_e (cf. Fig. 5), the location of the peak depends on log N_e.

Comparisons of the model ER distribution with the empirical distributions of ER of proteins of diverse organisms using a quantile-quantile plot (Fig. 3) and a normal probability plot (Fig. 4) reveal a remarkable similarity in the shapes of the distributions. The model and the empirical distributions similarly deviate from the log-normal distribution (all are convex upward in the middle third of the distribution in Fig. 4), the only appreciable difference being the somewhat sparser populated tails at low and high ER for the model distribution, a difference that should be expected given the smaller number of model data points.

Dependence of the ER Distribution on Protein Abundance, Mistranslation Rate, and Population Size. In the present model, protein abundance A, per-monomer mistranslation rate r, and effective population size N_e must be specified to compute fixation

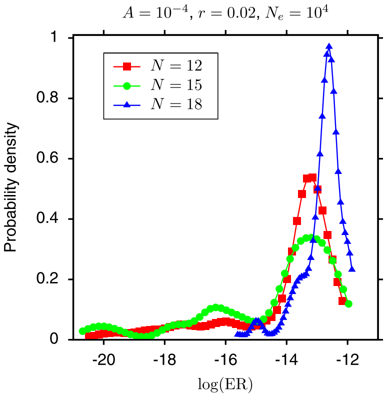

Fig. 2. Distributions (probability densities) of the logarithm of the fold-averaged ER for three chain lengths N and typical values of abundance ($A = 10^{-4}$), effective population size ($N_e = 10^4$), and per-monomer mistranslation probability ($r = 0.02$).

probabilities and therefore the fold-averaged ER. The effect of these parameters on the ER distribution can be examined by plotting the mean and median of the distribution against the parameter values (Fig. 5). Because the exponent in the denominator of the fixation probability (3) contains the product AN_e, the shape of the ER distribution depends weakly on A and N_e separately when their product is held fixed.

The Pearson's skewness Sk is an indicator of the relative contributions of the peak and the tail of the distribution. $Sk = 0$ for a normal distribution, and as the low ER tail becomes heavier, Sk decreases. An increase in protein abundance leads to a decrease of the mean ER and an increase of the weight of the tail (Fig. 5). This behavior is in qualitative agreement with the universally observed anticorrelation between protein abundance (expression level) and ER (11, 12, 15, 58). The increase in N_e has a similar effect on the ER distribution, underscoring the observed proportionality between N_e and the intensity of purifying selection (33). The effect of changing the mistranslation probability is more subtle. For relatively high r values, for which mistranslation is the dominant source of misfolding, the mean and median of ER decrease. The magnitude of the skewness, on the other hand, increases for small r (Fig. 5). Conceivably, increasing r imposes stronger demands on the CFP of variant sequences and hence leads to a decrease of the fold-averaged ER.

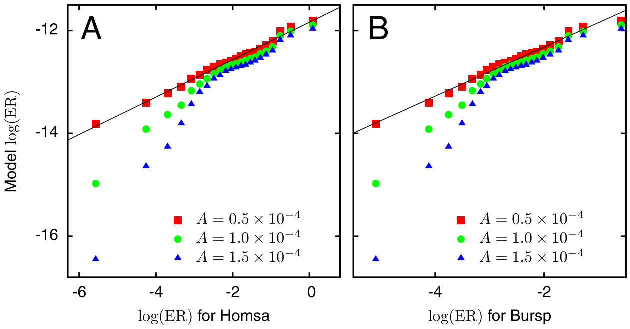

Fig. 3. Comparison of the model-derived ($N = 18$, $r = 0.02$, $N_e = 10^4$) and observed distributions of ER. The quantile-quantile plot compares the empirical ER distributions for the sets of orthologous proteins taken from ref. 4 for (*A*) *Homo sapiens* and *Macaca mulatta* (denoted Homsa, 16603 proteins) and (*B*) two strains of the bacterium *Burkholderia sp* (Bursp, 4014 proteins) with the distributions yielded by the model with three different abundance levels. Solid lines are linear fits to the $A = 0.5 \times 10^{-4}$ data.

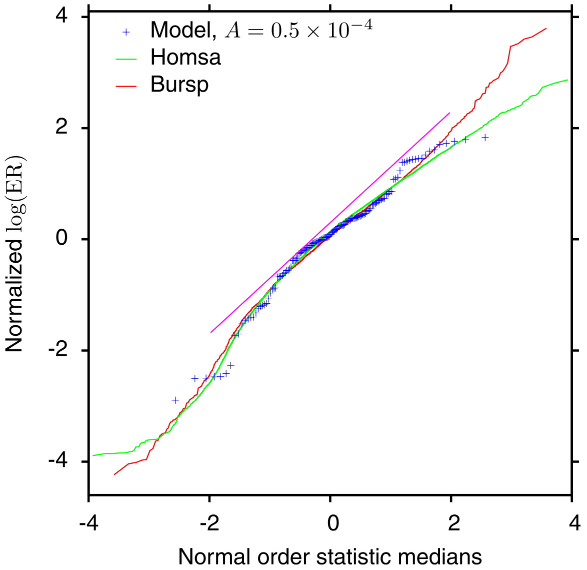

Fig. 4. Normal probability plot comparing the normalized model-derived and empirical ER distributions with the normal distribution. The distributions are the same as used for Fig. 3. The empirical ER distributions are more similar to each other and to the model distribution (all are convex upward) than to the normal distribution. A normal distribution would appear in a normal probability plot as a 45° line, shown here for reference.

Conclusions. The universal ER distribution across sets of orthologous protein-coding genes from all forms of cellular life and the equally ubiquitous anticorrelation between protein abundance and ER suggest that protein evolution might be largely governed by simple physical principles rather than function-specific adaptive processes. Here we investigate this possibility using a simple yet flexible model of protein folding together with the assumption that protein misfolding rate is the primary determinant of fitness. This assumption is at the core of the MIM hypothesis (15, 59) and is plausible given the stronger correlation between ER and protein abundance compared with other correlations between evolutionary and phenomic variables (8, 14) and the partial homogenization of ER values for domains fused in multidomain proteins

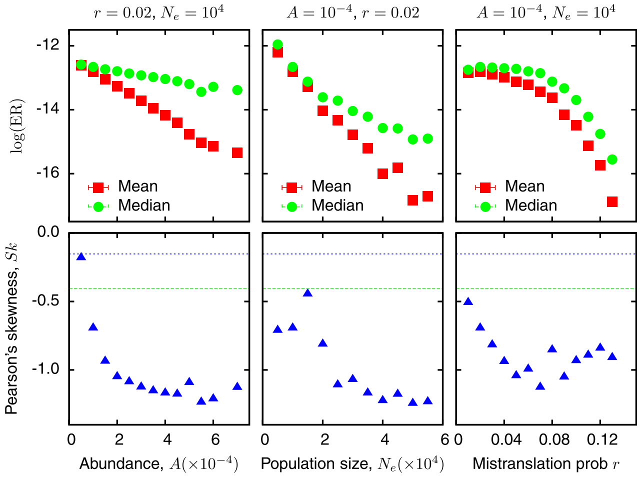

Fig. 5. Dependence of the mean, median, and Pearson's skewness, defined as 3(median − mean)/(standard deviation), of the ER distribution derived from the model on abundance, effective population size, and per-monomer mistranslation probability. The chain length is $N = 18$. The dashed lines are the empirical Sk values computed for the Bursp (blue) and Homsa (green) samples.

(59). The ER distribution derived from the model closely resembles the empirically observed universal distribution. The negative correlation between ER and protein abundance is reproduced as well. Given the stochastic rather than deterministic nature of the simulated folding in the present model, these results do not seem to be trivial consequences of the misfolding-fitness assumption. The model used is simple enough to be tractable but seems to be sufficiently rich to mimic the major forces that govern protein folding. Therefore the results suggest that the course of protein evolution is determined primarily by the fundamental principles of protein folding, which are the same across all life forms, hence the universal shape of the ER distribution and the anticorrelation between ER and protein abundance. Under this view, protein evolution is constrained by purifying selection against misfolding-inducing mutations but is much less dependent on selection for specific functions.

Materials and Methods

Calculation of the Native State. The energy of the chain is

$$E = \sum_{|i-j|>1} U_{ij} + \frac{bT}{2}\sum_{i=1}^{N-1}(r_{i,i+1}-a)^2, \qquad [4]$$

where the first term is the sum of the energies in Eq. **1** over nonnearest neighbor pairs, and the second term reflects the springs connecting nearest neighbors. The spring constant is proportional to temperature *T*. The folding kinetics of the chain are simulated via the Brownian dynamics algorithm. Units are chosen so that each component α of the *i*'th monomer's coordinates $x_{\alpha i}$ is updated according to

$$x_{\alpha i}(t+\Delta t) = x_{\alpha i}(t) - \frac{\Delta t}{T}\frac{\partial E}{\partial x_{\alpha i}}(t) + W_{\alpha i}(t), \qquad [5]$$

where $W_{\alpha i}(t)$ is a random variable with zero mean, variance 2Δ*t*, and uncorrelated with *W* for other times, monomers, and directions (60).

A large number (usually 3,000) of uncorrelated configurations is selected from the high *T* ensemble. Each unfolded configuration is quenched below the folding transition and held at that temperature to ensure that the folded configuration is surrounded by an energy barrier. Then, the exact minimum of the total energy and the corresponding 3D chain configuration are found using a conjugate gradient energy minimization algorithm.

To compare two folded configurations a mutual alignment score is computed as follows. After superimposing the centers of mass of the two configurations, the minimum of the sum of the squared distance between pairs of nearest neighbor monomers is computed over all possible mutual rotations and reflections. When folded structures of the same sequence are compared, the nearest neighbor monomers must be of the same type. This restriction is relaxed when comparing mutants. When the alignment score is below $0.1a^2$, the structures are considered identical. After classifying all folded configurations the native structure of a sequence is identified as the configuration most often attained by the quenching procedure. The CFP is the fraction of the high-temperature configurations that fold to the native structure.

When exhaustive sampling of the sequence space is computationally unfeasible, robust folders (CFP >0.8) are identified using a simulated annealing algorithm. Starting with a random sequence, mutations are chosen at random from a set of all possible substitutions and insertion–deletion pairs. A mutation is accepted when it increases the CFP. If ΔCFP <0 the mutation is accepted with a probability exp(ΔCFP/annealing temperature). By tweaking the starting and ending annealing temperature, cooling algorithm, and the number of attempted mutations, an appreciable fraction of robust folders can be identified. We confirmed that all robust folders are found by the simulated annealing search for *N* = 12 when they can be independently identified via exhaustive sampling of the sequence space. Robust folders presented in Table 1 were found using the following parameter values: *a* = 1, *b* = 50, and charge $q_{\pm} = \pm 2$. We used two sets of Lennard-Jones parameters and the Debye-Hückel screening length for comparing the effect of the long-range repulsion on the ER distribution. The values of the parameters are listed in Table S1, Table S2, and Table S3.

Identification of Orthologs and Estimation of Evolutionary Distances. Putative orthologs were identified as bidirectional best hits in an all-against-all BLASTP comparison of the protein sequences encoded in the respective pairs of genomes (61). Protein sequences of orthologs were aligned using MUSCLE (62). Maximum likelihood estimates of the amino acid distances between the aligned sequences of orthologous proteins were calculated using the PROTDIST program of the PHYLIP package (63) with the Jones-Taylor-Thornton evolutionary model (64) and gamma-distributed site rates with shape parameter 1.0. The distributions of the ER among orthologous genes were normalized by computing the geometric mean of all rates and dividing the original rates by this mean value, thus bringing the geometric mean of the normalized distribution to 1.

ACKNOWLEDGMENTS. The authors' research is supported by the intramural funds of the Department of Health and Human Services (National Institutes of Health, National Library of Medicine).

1. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
2. Gillespie JH (1994) *The Causes of Molecular Evolution* (Oxford Univ Press, Oxford).
3. Grishin NV, Wolf YI, Koonin EV (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res* 10:991–1000.
4. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 106:7273–7280.
5. Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J Mol Evol* 7:167–183.
6. Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. *Annu Rev Biochem* 46: 573–639.
7. Koonin EV, Wolf YI (2006) Evolutionary systems biology: Links between gene evolution and function. *Curr Opin Biotechnol* 17:481–487.
8. Wolf YI, Carmel L, Koonin EV (2006) Unifying measures of gene function and evolution. *Proc Biol Sci* 273:1507–1515.
9. Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.
10. McInerney JO (2006) The causes of protein evolutionary rate variation. *Trends Ecol Evol* 21:230–232.
11. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
12. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 13:2229–2235.
13. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102:14338–14343.
14. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327–337.
15. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
16. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10:715–724.
17. Dill KA, et al. (1995) Principles of protein folding—a perspective from simple exact models. *Protein Sci* 4:561–602.
18. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14: 70–75.
19. Shakhnovich E (2006) Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chem Rev* 106:1559–1588.
20. Xia Y, Levitt M (2004) Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol* 14:202–207.
21. Bornberg-Bauer E (2002) Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. *Z Phys Chem* 216:139–154.
22. Thirumalai D, Klimov DK (1999) Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr Opin Struct Biol* 9:197–207.
23. Clementi C (2008) Coarse-grained models of protein folding: Toy models or predictive tools? *Curr Opin Struct Biol* 18:10–15.
24. Clementi C, Maritan A, Banavar JR (1998) Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys Rev Lett* 81:3287–3290.
25. Nelson ED, Onuchic JN (1998) Proposed mechanism for stability of proteins to evolutionary mutations. *Proc Natl Acad Sci USA* 95:10682–10686.
26. Hoang TX, Cieplak M (2000) Molecular dynamics of folding of secondary structures in Go-type models of proteins. *J Chem Phys* 112:6851–6862.
27. Miller J, Zeng C, Wingreen NS, Tang C (2002) Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins* 47:506–512.
28. Wingreen NS, Li H, Tang C (2004) Designability and thermal stability of protein structures. *Polymer* 45:699–705.
29. Schnabel S, Bachmann M, Janke W (2007) Identification of characteristic protein folding channels in a coarse-grained hydrophobic-polar peptide model. *J Chem Phys* 126:105102.

30. Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA* 96: 10689–10694.
31. Bornberg-Bauer E (1997) How are model protein structures distributed in sequence space? *Biophys J* 73:2394–2403.
32. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.
33. Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA).
34. Hermans J, Berendsen HJC, van Gunsteren WF, Postma JPM (1984) A consistent empirical potential for water-protein interactions. *Biopolymers* 23:1513–1518.
35. Jorgensen WL, Tirado-Rives J (1988) The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 110:1657–1666.
36. Alm E, Baker D (1999) Matching theory and experiment in protein folding. *Curr Opin Struct Biol* 9:189–196.
37. Clementi C, Vendruscolo M, Maritan A, Domany E (2000) Folding Lennard-Jones proteins by a contact potential. *Proteins* 37:544–553.
38. Srinivas G, Bagchi B (2003) Study of the dynamics of protein folding through minimalistic models. *Theor Chem Acc* 109:8–21.
39. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem* 48:545–600.
40. Das P, Matysiak S, Clementi C (2005) Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc Natl Acad Sci USA* 102:10141–10146.
41. Lazaridis T, Karplus M (1997) “New view“ of protein folding reconciled with the old through multiple unfolding simulations. *Science* 278:1928–1931.
42. Matysiak S, Clementi C (2006) Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J Mol Biol* 363:297–308.
43. Das P, et al. (2005) Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proc Natl Acad Sci USA* 102:14569–14574.
44. Noguchi H, Yoshikawa K (2000) Folding path in a semiflexible homopolymer chain: A brownian dynamics simulation. *J Chem Phys* 113:854–862.
45. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181: 223–230.
46. Honeycutt JD, Thirumalai D (1990) Metastability of the folded states of globular proteins. *Proc Natl Acad Sci USA* 87:3526–3529.
47. Govindarajan S, Goldstein RA (1998) On the thermodynamic hypothesis of protein folding. *Proc Natl Acad Sci USA* 95:5545–5549.
48. Press WH, Flannery BP, Teukolsky SA, Vettering WT (1986) *Combinatorial Minimization: Method of Simulated Annealing* (Cambridge Univ Press, Cambridge, UK), 1st Ed, pp 326–334.
49. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104: 16152–16157.
50. Mélin R, Li H, Wingreen NS, Tang C (1999) Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. *J Chem Phys* 110:1252–1262.
51. Cui Y, Wong WH, Bornberg-Bauer E, Chan HS (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA* 99:809–814.
52. Govindarajan S, Goldstein RA (1997) Evolution of model proteins on a foldability landscape. *Proteins* 29:461–466.
53. Gutin AM, Abkevich VI, Shakhnovich EI (1995) Evolution-like selection of fast-folding model proteins. *Proc Natl Acad Sci USA* 92:1282–1286.
54. Saito S, Sasai M, Yomo T (1997) Evolution of the folding ability of proteins through functional selection. *Proc Natl Acad Sci USA* 94:11324–11328.
55. Bloom JD, Wilke CO, Arnold FH, Adami C (2004) Stability and the evolvability of function in a model protein. *Biophys J* 86:2758–2764.
56. Wilke CO, Drummond DA (2006) Population genetics of translational robustness. *Genetics* 173:473–481.
57. Norris JR (1997) *Markov Chains, Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge Univ Press, Cambridge, UK).
58. Schrimpf SP, et al. (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* 7:e1000048.
59. Wolf MY, Wolf YI, Koonin EV (2008) Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biol Direct* 3:40.
60. Turq P, Lantelme F, Friedman HL (1977) Brownian dynamics: Its applications to ionic solutions. *J Chem Phys* 66:3039–3044.
61. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
62. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
63. Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 266:418–427.
64. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.