doi:10.1016/j.jmb.2004.08.086 J. Mol. Biol. (2004) 343, 1451–1466

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Genomic Determinants of Protein Folding Thermodynamics in Prokaryotic Organisms

Ugo Bastolla[1]*, Andrés Moya[2], Enrique Viguera[1,3] and Roeland C. H. J. van Ham[1,4]

[1]*Centro de Astrobiología (CSIC-INTA), E-28850 Torrejón de Ardoz, Spain*

[2]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de Valencia, E-46071 Valencia Spain*

[3]*Departamento de Biología Celular y Genética, Universidad de Málaga, E-29071 Málaga Spain*

[4]*Plant Research International PO Box 16, 6700AA Wageningen, The Netherlands*

**Corresponding author*

Here we investigate how thermodynamic properties of orthologous proteins are influenced by the genomic environment in which they evolve. We performed a comparative computational study of 21 protein families in 73 prokaryotic species and obtained the following main results. (i) Protein stability with respect to the unfolded state and with respect to misfolding are anticorrelated. There appears to be a trade-off between these two properties, which cannot be optimized simultaneously. (ii) Folding thermodynamic parameters are strongly correlated with two genomic features, genome size and G+C composition. In particular, the normalized energy gap, an indicator of folding efficiency in statistical mechanical models of protein folding, is smaller in proteins of organisms with a small genome size and a compositional bias towards A+T. Such genomic features are characteristic for bacteria with an intracellular lifestyle. We interpret these correlations in light of mutation pressure and natural selection. A mutational bias toward A+T at the DNA level translates into a mutational bias toward more hydrophobic (and in general more interactive) proteins, a consequence of the structure of the genetic code. Increased hydrophobicity renders proteins more stable against unfolding but less stable against misfolding. Proteins with high hydrophobicity and low stability against misfolding occur in organisms with reduced genomes, like obligate intracellular bacteria. We argue that they are fixed because these organisms experience weaker purifying selection due to their small effective population sizes. This interpretation is supported by the observation of a high expression level of chaperones in these bacteria. Our results indicate that the mutational spectrum of a genome and the strength of selection significantly influence protein folding thermodynamics.

Keywords: protein folding; molecular evolution; orthologous proteins; intracellular bacteria; mutational bias

Introduction

Orthologous proteins expressed in different organisms share similar structure and function, but how similar are their thermodynamic properties? In recent years, experimental, computational and statistical studies have provided important insights into the process of protein folding.[1] The main focus in these studies has been the role of the native state topology, which is known to be highly conserved through evolution. Comparatively little attention has been paid to the question how sequence evolution influences the folding properties of proteins. A better understanding of this issue would be very useful both for testing theories of evolutionary change and for improving our knowledge of protein folding. Genomic projects now provide a wealth of evolutionary data that can be used to address this question.

Ohta proposed that the major cause of molecular evolution in naturally evolving populations is the fixation of slightly deleterious mutations in small populations through random genetic drift.[2] Bacteria with obligatory endosymbiotic or parasitic lifestyle in particular are subject to this process because

Abbreviations used: PDB, Protein Data Bank; PCA, principal component analysis; KD, Kyte & Doolittle.

E-mail address of the corresponding author: bastollau@inta.es

transmission bottlenecks during infection of new hosts result in small effective population sizes. Moreover, these bacteria reproduce asexually and lack genetic recombination, factors thought to exacerbate the process of fixation of deleterious mutation.[3] In this light, Moran compared the substitution rate of proteins of the aphid endosymbionts of the species *Buchnera aphidicola* with those of its free-living relative *Escherichia coli*, and found that the former tend to evolve at a faster rate. She interpreted this finding as evidence of the reduced efficacy of selection in endosymbiotic bacteria.[4] Furthermore, Lambert & Moran[5] showed through a computational analysis that 16 S rRNAs of obligatory endosymbionts have accumulated deleterious mutations, resulting in thermodynamically less stable molecules than those of related free-living bacteria.

The recent study by Itoh *et al.*[6] confirmed the acceleration of the substitution rate on a genomic scale in intracellular bacteria, but the authors attributed this effect primarily to higher mutation rates. However, their study was based on a selection of genes that yielded tree topologies in which *B. aphidicola* forms a sister group with *E. coli*, with *Haemophilus influenzae* as an outgroup. Genes supporting the alternative topology with *B. aphidicola* as the outgroup were attributed to lateral gene transfer and excluded from the analysis. A subsequent phylogenetic analysis by Canbäack *et al.*[7] strongly suggested that the latter tree topology was in fact produced as an artifact of tree reconstruction methods, using genes with an increased evolutionary rate and strong compositional bias in the *B. aphidicola* lineage, and that the genes selected for the analysis by Itoh *et al.*[6] were the ones most strongly constrained by natural selection. Without this bias in the genes examined, the acceleration of the substitution rate in endosymbiotic bacteria is best explained by relaxation of purifying selection and host level selection, since genes which are essential for the host metabolism evolve more slowly and appear to be more constrained.[7]

In a recent computational study, in the framework of the sequencing of the genome of *B. aphidicola* from *Baizongia pistacea*,[8] our group found that the normalized energy gap, a crucial indicator of efficient and fast folding, is systematically lower for proteins encoded in obligate intracellular bacteria than for the orthologous proteins of their free-living relatives. In light of the statistical theory of protein folding, this implies that slow folding, possible misfolding and aggregation can dramatically reduce protein-folding efficiency in intracellular bacteria. That such problems may indeed occur is suggested by the observation of exceptionally high expression levels of chaperones in these bacteria,[9,10] proteins that help other proteins to fold properly and reduce the risk of misfolding. Furthermore, a recent study has demonstrated that over-expression of the GroELS chaperone produced a fitness recovery in an experimental population of *E. coli* that had experienced accumulation of deleterious mutations by passage through a series of populational bottlenecks.[11]

Here, we provide a quantitative relationship between genomic traits and protein thermodynamics. To address this issue, it is necessary to adopt a statistical approach and to examine a large sample of organisms and proteins. We extend our previous computational study to a total of 21 protein families from 73 prokaryotic species, and perform thermodynamic calculations with a new method. Since not enough experimental data are available for addressing this problem, a computational approach, like the one described here, can give very valuable insights. We are confident that our results will stimulate experimental verification of the evolutionary relationships disclosed here.

The proteins were selected on the basis of the following criteria: (i) family members must be present in intracellular bacteria; (ii) they must be soluble globular proteins; (iii) they must have at least one experimentally known structure; and (iv) they cannot be too large in order to yield reliable results. These requirements considerably reduced the number of protein families that could be included. However, since each individual family showed the same correlations as observed for the entire set of proteins, further increasing the number of proteins would not have modified our results qualitatively.

The computational method is based on a fold recognition algorithm that uses an effective free energy function, without relying on sequence similarity. For most globular proteins considered, the effective free energy that we use takes its lowest value on the native structure, when this is available or on structures of proteins homologous to the query sequence. Moreover, the effective native energy correlates strongly with the unfolding free energy measured experimentally for proteins with two-states folding thermodynamics. Therefore, the correlations presented here are expected to remain valid if experimental quantities are used instead of computational estimates.

To circumvent the limitations of predicted protein folding thermodynamics properties, we have also correlated genomic and folding thermodynamic properties with a selection of ten amino acid properties related to hydrophobicity. The two amino acid properties showing the strongest correlations, however, also take into account other types of interactions, besides the hydrophobic effect. We therefore sometimes refer to the set of ten properties by the term "interactivity" to stress that the hydrophobic effect plays a central role, but not the only one. Amino acid properties are strongly correlated both with genomic properties and with experimental and calculated thermodynamic properties, thus supporting the correlations discovered through our computational approach.

Previously, Gu *et al.*[12] and D'Onofrio *et al.*[13] considered the relationship between protein

hydrophobicity, a proteomic property, and the G+C content of the corresponding gene, a genomic property (variation in G+C content is much larger between genomes than within a genome). Since these two studies produced contradicting results, we reconsidered this issue here, using several hydrophobicity scales. Our main conclusion is that there is a positive correlation between protein hydrophobicity (in a general sense) on one side, and G+C content and genome size on the other.

Results

Fold recognition

For 94% of the data set (908 proteins out of 965) one homologous protein was recognized as the best scoring model, even when sequence identity between target and template was as low as 15%. The remaining 6% of proteins were discarded from the analysis. For these proteins, at least one homologous model obtained an effective energy very close to the one of the best scoring model, so that the normalized energy gap α was extremely small. The discarded proteins mostly belonged to one of five families (AckA; Ddl, Dyr, Efts and RnpA) and to organisms characterized by a small value of the normalized energy gap α. Therefore, their inclusion in the statistics would have made the average normalized energy gap for these species even smaller, thus strengthening the observed correlations.

Our alignments compared favorably with those stored in the PFAM database of aligned protein families,[14] in the sense that the sequence identities correlated very strongly with those obtained from the PFAM alignments and in several cases even surpassed these.

Correlations of protein thermodynamic parameters

We have characterized protein folding thermodynamics through two variables: the unfolding free energy per residue, measuring stability with respect to the unfolded state, and the normalized energy gap α, measuring stability with respect to misfolded states.

These two quantities are negatively correlated ($R=-0.40$, student-t -3.7, 71 degrees of freedom, $P<5\times10^{-4}$; Figure 1). A similar correlation was found in an analysis of a large database of non-redundant protein structures.[15]

Protein thermodynamics correlates with hydrophobicity

The correlation between unfolding free energy and the normalized energy gap was further examined by considering a key property of a protein sequence, the mean hydrophobicity. More hydrophobic proteins tend to have larger unfolding free energy (they are more stable with respect to unfolding), but also have a smaller energy gap (many alternative states have low effective free energy). A maximally hydrophobic protein would behave like a homopolymer, with a vanishing free energy gap, and without a unique native state.

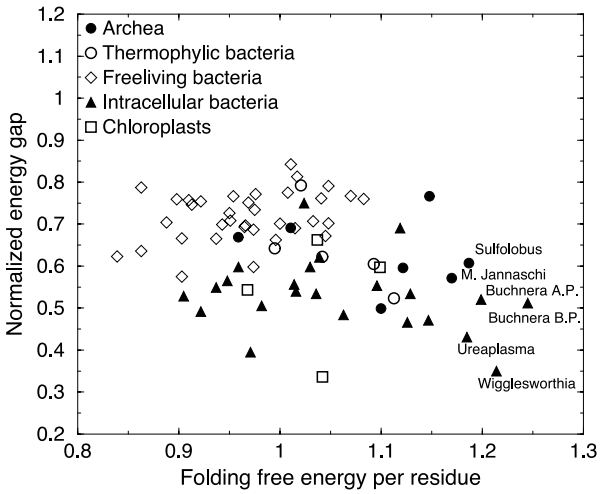

Figure 1. Average normalized energy gap *versus* average free energy per residue. Each point represents a prokaryotic species.

Although hydrophobicity plays a central role in the protein folding literature, there is no agreement on how to measure it. Hundreds of empirical hydropathy scales have been proposed. We used a selection of eight hydropathy scales from the literature plus two amino acid properties recently studied by one of us and co-workers,[16] denoted by IH and CH, which correlate strongly with other hydropathy scales. For each organism, we have calculated the mean hydrophobicity of its proteins, using all ten scales, and correlated these with thermodynamic and genomic properties.

For each hydropathy scale, the average hydrophobicity is positively correlated with the average of the calculated unfolding free energy per residue, and negatively correlated with the predicted normalized energy gap (Table 1). Correlations vary in extent, but not in sign, depending on the scale chosen. We found the weakest correlation for the Kyte & Doolittle (KD) scale (correlation coefficients $R=0.22$, -0.37, respectively). This is not unexpected, given that KD is a hydropathy scale mainly used for identification of transmembrane helices that attributes small or even negative values to aromatic residues, which interact very strongly. The KD scale also correlates least strongly with genomic properties (see below). The strongest correlations with protein folding thermodynamic properties were found for the novel scales IH ($R=0.83$, -0.68, respectively) and CH ($R=0.83$, -0.72, respectively). In addition, this result is not unexpected, since the IH scale was derived from the energy function that we used for computational estimates, and the CH scale correlates very strongly

Table 1. Correlation coefficients of different hydrophobicity scales with genomic and proteomic properties

	CH	IH	G98	FP	AV	MP	WW	R88	L76	KD
α	−0.72	−0.68	−0.61	−0.52	−0.58	−0.45	−0.46	−0.39	−0.35	−0.37
F/N	0.83	0.83	0.54	0.57	0.58	0.67	0.49	0.32	0.35	0.22
GC_{12}	−0.72	−0.59	−0.48	−0.42	−0.40	−0.32	−0.34	−0.32	−0.26	−0.14
GC_3	−0.56	−0.43	−0.39	−0.35	−0.30	−0.20	−0.25	−0.21	−0.20	−0.06
Size	−0.61	−0.55	−0.48	−0.42	−0.41	−0.38	−0.35	−0.27	−0.23	−0.16
PC-1	−0.90	−0.83	−0.69	−0.67	−0.66	−0.61	−0.57	−0.47	−0.43	−0.33

With 73 organisms, a 0.05 significance level is achieved for $|R|>0.20$ and a 0.01 significance level for $|R|>0.27$. The variables considered are: (1) normalized energy gap, α; (2) unfolding free energy per residue, F/N; G+C content at first and second (3) and third (4) codon position; (5) genome size; (6) principal component coefficient, PC-1. Hydropathy scales are ranked according to the value of PC-1. Notice that the scales that correlate more strongly with thermodynamic properties also tend to correlate more strongly with genomic properties. The hydropathy scales considered are: (1) CH, Connectivity scale (U. Bastolla *et al.*, unpublished results), (2) IH, Interaction scale (U. Bastolla *et al.*, unpublished results), (3) G98, hydrophobic-polar classification;[12] (4) FP, Fauchere & Pliska transfer free energies;[68] (5) AV, average hydrophobicity scale;[72] (6) MP, Manavalan, & Ponnuswamy;[70] (7) WW,[71] (8) R88, transfer free energies calculated by Roseman;[69] (9) L76, Levitt;[66] (10) KD, Kyte & Doolittle.[67]

with it. More surprisingly, these two scales also show the strongest correlations with genomic properties like genome size and composition (see below).

For three out of ten scales (KD, R88, L76), the correlations were not significant unless the properties of each protein were normalized with respect to the representative sequence of each protein family. This normalization reduces the effect of protein topology and chain length relative to the effect of sequence changes.

The average thermodynamic parameters *versus* average hydrophobicities measured with IH parameters are plotted in Figure 2. Each point represents one organism. Besides prokaryotic species, we also show proteins from chloroplast and eukaryotic nuclear genomes. They show a similar pattern as proteins from prokaryotic genomes, except that the proteins from metazoans (*Homo sapiens* and *Caenorabditis elegans*) appear to be more stable than expected on the basis of the general trend. This is probably an effect of the fact that the templates for structural modeling were in this case mostly human proteins, which increases the predicted stability.

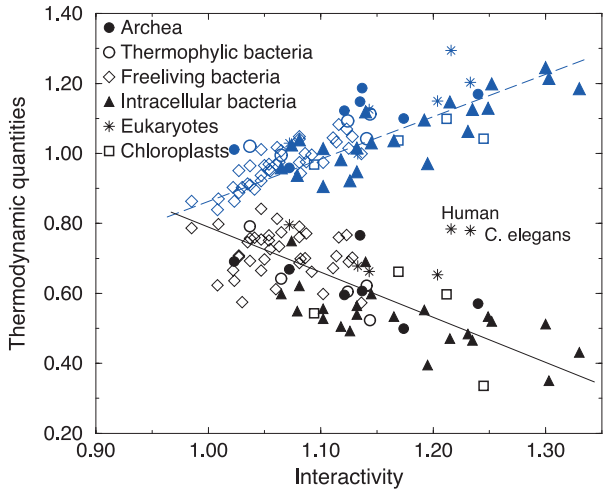

Figure 2. Average thermodynamic properties of proteins of different organisms as a function of the IH interactivity of their sequences. The symbols fitted by the continuous line (decreasing with hydrophobicity) represent the normalized energy gap, and the symbols fitted by the broken line (increasing with hydrophobicity) represent the unfolding free energy per residue. The correlation coefficient is $R=0.83$ between interactivity and calculated unfolding free energy and $R=-0.68$ between interactivity and normalized energy gap.

Protein thermodynamics correlates with genome size

Genome size was found to be strongly correlated with the average protein thermodynamic parameters. It correlated positively with the normalized energy gap with correlation coefficient $R=0.65$, $P<10^{-6}$ (Figure 3) and negatively with the calculated unfolding free energy per residue, $R=-0.49$, $P<10^{-5}$. These correlations can be understood through the negative correlation between genome size and mean hydrophobicity (see below). Proteins in smaller genomes, like those of intracellular bacteria, tend to be more hydrophobic and hence tend to have larger unfolding free energy (they are more stable with respect to the unfolded state) and smaller normalized energy gap (they are less stable with respect to misfolded states). The latter condition is expected to cause less efficient folding.

Results look qualitatively similar for each of the 21 protein families which were included in the

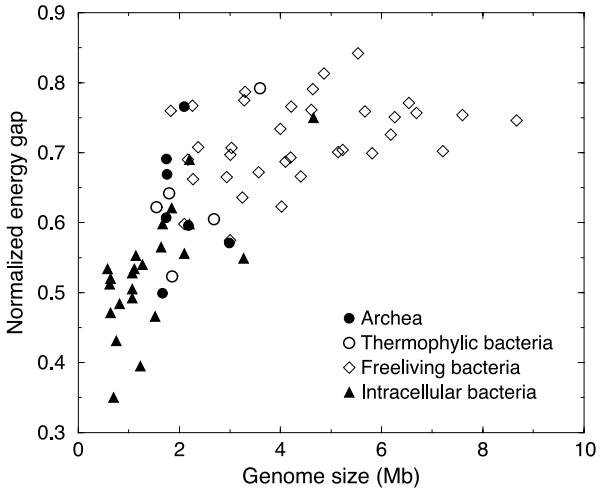

Figure 3. Average of the normalized energy gap *versus* genome size.

calculation of species averages. A different pattern was observed for the chaperone DnaK, which was not used for species averages. No correlation between thermodynamic and genomic properties was found for this protein. As mentioned above, chaperones buffer the effects of detrimental mutations, and they appear to be important for the maintenance of fitness in intracellular bacteria. The aberrant pattern for DnaK therefore reinforces the proposed scenario of folding problems for proteins from small bacterial genomes.

Compositional bias

In the analysis of G+C composition of genes, third codon positions and first plus second codon positions were considered separately. The third codon position is almost neutral with respect to selection at the protein level, although it can be subject to selection for optimal codon usage.[16] This position is thought to reflect mostly mutational forces, at least in prokaryotic genomes.[17,18] As originally noticed by Sueoka,[19] the G+C composition of a gene strongly influences the amino acid composition of its coded protein. Bernardi & Bernardi were the first to notice that the GC content at first and second position, GC_{12}, is strongly correlated with the one at third position, GC_3,[20] and both are correlated with the genomic G+C content.

The G+C content at third position, GC_3, varied broadly from 0.09 to 0.95 in the species examined. In contrast, the G+C content and first plus second codon positions, GC_{12}, only varied from 0.27 to 0.59, due to selection at the protein level. In our data set, the correlation coefficient between GC_{12} and GC_3 is $R=0.87$ (Figure 4), and both values are strongly correlated with the genomic GC content (correlation coefficient $R=0.98$ for GC_3 and $R=0.93$ for GC_{12}). Similar results were reported by several authors, but we show the plot in Figure 4 in order to highlight that the GC_{12} *versus* GC_3 curve is concave downwards. This means that bacteria with low GC_3 have a lower GC_{12} than would be expected from extrapolation of the pattern of bacteria with high GC_3. All of the former have obligatory intracellular lifestyles.

Intracellular bacteria are characterized by a strong compositional bias towards nucleotides A and T and by reduced genomes. This association is quantitative. The G+C content correlates with genome size, as previously observed by Moran.[21] For the genomes considered here, the correlation coefficient is $R=0.65$, $t=7.3$, $P<10^{-6}$ (Figure 5). Genome size and G+C content appear to be uncorrelated in Archaea and thermophilic Eubacteria, while they are significantly correlated both in free-living bacteria ($R=0.52$, $P<10^{-3}$) and, more strongly, in intracellular bacteria ($R=0.72$, $P<10^{-4}$).

Consistently with the correlation between genome size and composition, the GC_3 was found to be negatively correlated with the predicted folding free energy per residue ($R=-0.52$, $t=-5.15$, $P<10^{-5}$) and positively correlated with the normalized energy gap ($R=0.42$, $t=3.9$, $P<10^{-3}$). For GC_{12}, these correlations were even stronger. $R=-0.77$ ($t=-6.4$, $P<10^{-6}$) and $R=0.58$ ($t=6.1$, $P<10^{-6}$), respectively. This is explained by the fact that first and particularly second codon position affect the encoded amino acid much more than the third codon position, whose changes are often synonymous.

Genomic properties and hydrophobicity

We have reconsidered the relationship between protein hydrophobicity and the G+C content of the corresponding gene. Previous results were contradictory. Gu *et al.*, using a simple classification of amino acid residues in three classes from hydrophobic to polar, showed that the G+C content of a

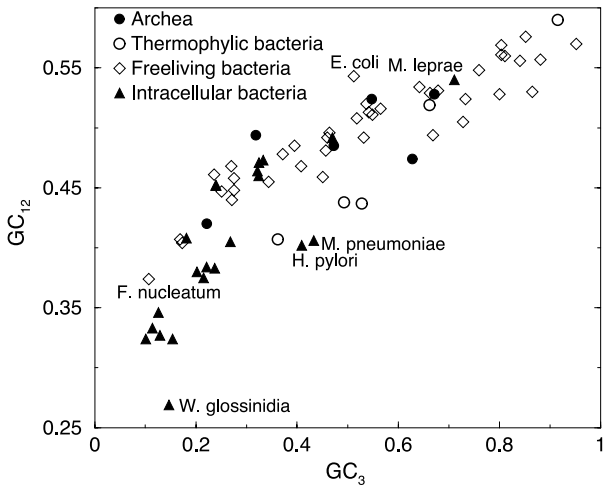

Figure 4. G+C content at first and second (vertical) *versus* third (horizontal axis) codon position. The correlation coefficient is $R=0.87$.

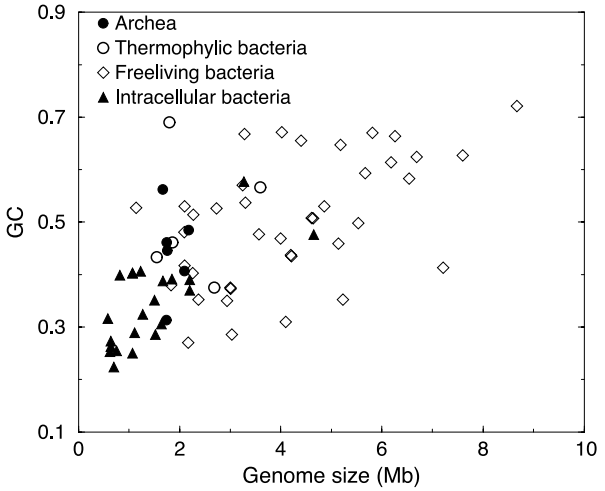

Figure 5. Genome size *versus* genomic G+C content. Intracellular bacteria cluster in the lower left corner of this plot.

gene is negatively correlated with the hydrophobic content of the corresponding protein.[12] In their study, the G98 classification of amino acid residues was used and each protein family was examined separately. D'Onofrio *et al*.[13] came to the opposite conclusion using the KD hydropathy scale and averaging over sequences that corresponded to different protein families. Here, we correlated average hydrophobicities with three genomic properties: GC_{12}, GC_3 and genome size; we considered ten different hydrophobicity scales, and we averaged hydrophobicities of proteins of the same organism in two different ways.

In the first calculation, we averaged hydrophobicities without normalizing them. This is analogous to the procedure followed by D'Onofrio *et al*.[13] We found significant ($P<0.05$) negative correlations between GC_{12} and hydrophobicity for six hydrophobicity scales (CH, IH, WW, MP, FP, AV), non-significant negative correlation for three scales (R88, L76, G98) and a weak but significant positive correlation for the scale KD. The latter was used by D'Onofrio *et al*.,[13] whose results are here qualitatively reproduced. The GC_3 showed significant negative correlations only for three scales (CH, IH, WW) and significant positive correlations for the scale KD, while the genome size showed significant negative correlations for four scales (CH, IH, WW, MP) and no case of significant positive correlations.

In the second calculation, hydrophobicities were normalized before averaging, as explained in Materials and Methods. By doing so, the between-genome differences due to variation in the fold composition are strongly reduced. This calculation is thus analogous to the work of Gu *et al*.,[12] who considered each protein family separately. Almost all hydrophobicity scales yielded highly significant ($P<0.01$) negative correlation with the GC_{12}, a significant but weaker correlation ($P<0.05$) with the GC_3, and a correlation of intermediate value ($P<0.01$) with the genome size. The only exceptions were the L76 scale, which showed correlations only at the 5 % level, and the KD scale, which showed non-significant (but still negative) correlations. The strongest correlations were observed for the scales CH ($R=-0.72$, -0.56 and -0.61 for GC_{12}, GC_3 and genome size, respectively), IH ($R=-0.59$, -0.43 and -0.55), G98 ($R=-0.48$, -0.39 and -0.48) and FP ($R=-0.42$, -0.35 and -0.42). It is interesting that almost the same ranking is observed for the correlation between hydrophobicity and thermodynamic properties: the hydrophobicity scales that correlate strongest with G+C content are also those which correlate strongest with protein thermodynamic parameters.

The relationship between genome size and hydrophobicity, calculated using the novel scale CH, is shown in Figure 6.

Results concerning normalized protein properties are summarized in Table 1. We conclude that protein hydrophobicity, as measured by nine of the ten scales considered, is negatively correlated with both G+C content and genome size, but for most scales correlations are only significant when protein properties are normalized in order to reduce their dependence on family-specific properties like length and topology. The KD scale is the only one that does not show significant correlations with any genomic property even after averaging. The use of this scale, and the fact that proteins with different length and topology were averaged together, explains the apparent contradiction between our results and the results obtained by D'Onofrio *et al*.[13]

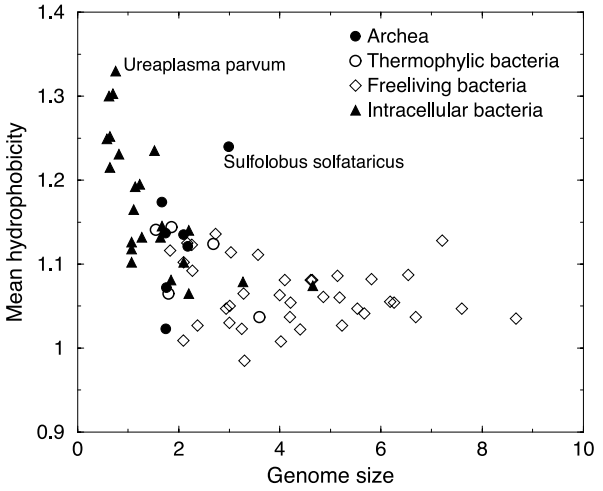

Figure 6. Mean protein hydrophobicity measured with the scale CH *versus* genome size. Each point represents a prokaryotic species.

Sequence similarity

Sequence similarity between the protein examined and the protein used as structural template influenced positively the estimated protein stability: it correlated positively with the energy gap ($R=0.55$) and also, although not significantly, with the unfolding free energy. Therefore, the reported negative correlation between energy gap and unfolding free energy would become stronger if we correct for this artifact.

We observed a positive correlation between genome size and sequence similarity with the best structural model. This is due to the fact that proteins are mostly crystallized from a few model organisms, which happen to be free-living bacteria with large genomes. Very few proteins from pathogenic bacteria have been crystallized, and none from endosymbionts. In this case, the extent of the correlation between genome size and energy gap would be lower after correcting for sequence similarity. This artifact, however, does not affect the hydrophobicity, which does not depend on template structure, and only slightly affects the unfolding free energy. As explored hereafter, the extent of the artifact can be significantly reduced through principal component analysis (PCA).

Principal component analysis

The properties that we have described so far provide a coherent quantitative characterization of the intracellular lifestyle. Intracellular bacteria are characterized by reduced genome size, AT rich genes, and hydrophobic (interactive) proteins with large unfolding free energy but a low normalized energy gap. These properties were studied together through PCA (see Table 2). The first principal component obtained distinguished very effectively between intracellular and free-living bacteria. Interesting outliers included *Yersinia pestis* and *Mycobacterium leprae*, which are pathogens of recent origin that still retain moderately large genomes and a high G+C content, and *Fusobacterium nucleatum*, a bacterium that lives in dental plaque and has a very strong compositional bias towards A+T. The first PC correlated strongly with genomic properties (genome size, $R=0.81$; GC_3, $R=0.77$; GC_{12}, $R=0.88$) and with proteomic properties (unfolding free energy, $R=-0.75$; normalized energy gap, $R=0.81$; hydrophobicity or interactivity, correlation ranging from $R=-0.33$ for the KD scale to $R=-0.90$ for the CH scale) Its correlation with the sequence similarity with the representative sequence was much weaker ($R=0.39$). The first principal component explained 58 percent of the total variance.

By contrast, the second component correlated strongest with sequence similarity ($R=0.85$), then with the effective free energy ($R=0.47$) and the normalized energy gap ($R=0.42$), and did not significantly correlate with the remaining variables. Its value can therefore be interpreted as the excess of calculated stability, both for the free energy and for the normalized energy gap, attributed to proteins with high similarity to the best model. The second component explained only 17% of the total variance, and it was particularly large in the clade of enterobacteria, since many of the proteins crystallized are derived from *E. coli* (see Figure 7). Note that the endosymbiotic species *B. aphidicola* and *Wigglesworthia*, which are closely related to *E. coli*, also have large value of the second principal component. The stability of their proteins is therefore overestimated, and nevertheless they present the smallest predicted values of the normalized energy gap.

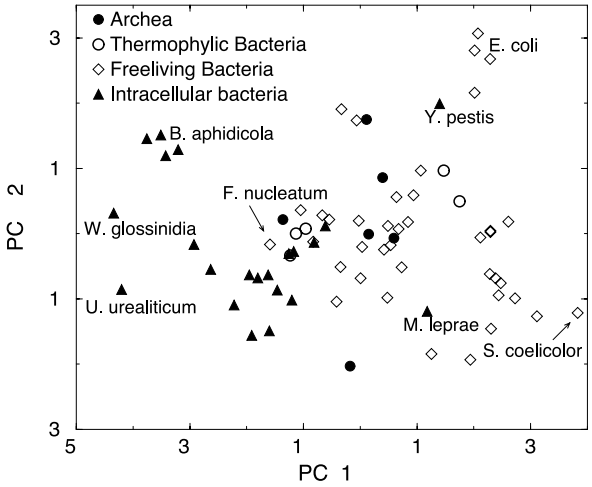

Figure 7. First and second principal components.

Discussion

We have shown that the computational thermodynamic properties of orthologous prokaryotic proteins, sharing the same structure, function and evolutionary origin, but encoded in different organisms, are quantitatively correlated with two traits of the genomes in which they evolved: genome size and the G+C content of its genes.

Computational approach

In this study, we calculated folding thermodynamic properties using an effective free energy function and a fold recognition algorithm. The validity of our computational approach is demonstrated by a strong correlation between calculated and experimental unfolding free energies for proteins folding with two-states thermodynamics (correlation coefficient $R=0.95$; U.B., unpublished results) and for a database of more than thousand mutants (correlation coefficient $R=0.65$; U.B., unpublished results). The validity of the energy function is also supported by its success as a scoring function in our fold recognition algorithm. The structure of a homologous

Table 2. Matrix of correlation coefficients

	α	F/N	Hydro	size	GC_{12}	GC_3	Seq.id.
F/N	−0.41						
Hydro	−0.72	0.83					
Size	0.65	−0.49	−0.61				
GC_{12}	0.58	−0.60	−0.72	0.65			
GC_3	0.42	−0.47	−0.56	0.61	0.89		
Seq.id.	0.55	0.05	−0.35	0.33	0.27	0.18	
PC-1	0.81	−0.75	−0.90	0.81	0.88	0.77	0.39
PC-2	0.42	0.47	0.06	0.07	−0.16	−0.21	0.85

The columns and rows represent three protein related quantities: normalized energy gap α, folding free energy per residue F/N, hydrophobicity (CH scale); three genomic quantities: genome size, G+C content at first and second codon position (GC_{12}), G+C content at third codon position (GC_3); percent identity with the sequence of the representative protein, and first (PC-1) and second (PC-2) principal component of the correlation matrix. With 73 organisms, a 0.05 significance level is achieved for $|R|>0.20$ and a 0.01 significance level for $|R|>0.27$.

protein always appeared in the top three scores and got highest scores in 94% of the cases, despite we did not score sequence similarity. A result that further supports our computational analysis is the existence of correlations between hydrophobicity and genomic properties. These correlations do not rely on our method for predicting folding thermodynamic properties and thus constitute an independent line of evidence for a correlation between genomic and proteomic properties.

As mentioned above, the similarity between the sequence of the target protein and the sequence of the best structural template can influence the estimated thermodynamic parameters. However, PCA showed that the excess of stability attributed to proteins that are very similar to their structural template only contributes to the second principal component. Since principal components are orthogonal, interpretation of the second component as the unwanted influence of sequence similarity on estimated stability implies that the first component is largely free from this effect. The first component can thus safely be interpreted as a quantitative characterization of the intracellular lifestyle. Proteins from endosymbiotic enterobacteria present properties typical of intracellular organisms (large unfolding free energy, small normalized energy gap) despite their high similarity with template proteins from *E. coli*.

The notion that hydrophobicity is one of the main forces in protein folding is a long-standing concept[22] and has been subject to intense experimental study.[23] Mutational studies have pointed at the importance of hydrogen bonds for providing stability against unfolding, comparable to that of the hydrophobic effect.[24] In this study, we did not explicitly consider hydrogen-bonding. However, since orthologous proteins share very similar and often identical secondary structure, differences in stability between them rarely arise from differences in the network of hydrogen bonds of their main-chains. Hydrogen bonds involving side-chains are implicitly taken into account by our energy function. Electrostatic interactions are another essential source of stability for folded proteins that our energy function takes into account implicitly. Together with other factors, they are thought to be crucial for the thermostability of proteins in thermophilic organisms.[25–27] Interactions of aromatic rings between themselves[28,29] and with positively charged residues[30–32] have also been proposed as an important source of thermostability and can be effectively reproduced through our energy function.

Despite the importance of these energy terms, hydrophobicity provides the largest contribution to our calculated unfolding free energy and it provides a simple key for the interpretation of our results. Among the hydropathy scales used here, the largest explanatory power for both genomic and proteomic properties was displayed by the scales CH and IH. The IH scale was derived from our interaction matrix, thus it is not a surprise that it correlates very strongly with computational estimates of thermodynamic properties, but it is surprising that this scale is also the second in the rank of the correlations with genomic properties. The CH scale, derived from the connectivity properties of amino acid residues in solved crystal structures (U. Bastolla *et al*. unpublished results), correlates strongest both with genomic and with thermodynamic properties. Since this scale was obtained independently of the effective energy function, it gives strong independent support to the correlations highlighted here.

The KD scale shows the weakest correlation both with genomic and with thermodynamic properties. This scale attributes small or negative hydrophobicity to aromatic amino acid residues (Phe, Trp, Tyr), which have some of the largest hydrophobicity values in other empirical scales. They have also very high values in the novel IH and CH scales, since they tend to form rather strong inter-residue interactions and they tend to be very connected. Enhanced aromatic interactions have been found in proteins of thermophilic organisms.[29] For these reasons, the aromatic content is expected to correlate strongly with folding thermodynamic properties. Failure to attribute large values to hydrophobic amino acid residues explains the weak correlation between the KD scale and thermodynamic properties. It also explains the weak correlation between the KD scale and the A+T content, which is strongly correlated with the aromatic content (Phe=TTY, Tyr=TAY, W=TGG).

Unfolding *versus* misfolding

In this study, we observed that the unfolding free energy per residue and the normalized energy gap, a computational measure of folding efficiency, are negatively correlated. This implies that orthologous proteins that are more stable with respect to unfolding are less stable with respect to misfolding. We consider this an instance of frustration in protein sequence space: natural selection for both thermodynamic properties acts in opposite directions, thus it cannot optimize both at the same time and has to trade-off between them.

At first sight this result seems counterintuitive, but it can easily be understood considering the effect of hydrophobicity. Mean protein hydrophobicity correlates positively with the calculated unfolding free energy and negatively with the normalized energy gap. The latter correlation is expected, since a highly hydrophobic sequence will have several compact conformations with effective energies comparable to the native state. The limit of a maximally hydrophobic sequence corresponds to a homopolymer, in which the normalized energy gap shrinks to zero. No folding is expected in this limit, also according to theory and simulation of lattice polymers.[33,34] All the hydrophobicity scales that we considered are correlated with calculated thermodynamic properties. For some scale these correlations are so strong that they fully explain the negative correlation between unfolding free energy and normalized energy gap.

The role of the hydrophobic effect on the stability of the native state is demonstrated by several experimental studies.[23] It has been recognized only more recently that hydrophobicity also affects the stability of misfolded states.

Intermediate states of protein folding can be characterized as metastable, partially folded states. They may occur in the pathway between the unfolded and the native state (for instance, intermediate states where one domain is folded and another one is not yet) or off-pathway, in which case they act as kinetic traps that slow down the folding process. Folding intermediates are not an intrinsic characteristic of a protein. Several experiments have shown that mutations, as well as physico-chemical changes in the solvent, like pH and temperature changes, can stabilize or destabilize folding intermediates. Often, mutations stabilizing folding intermediates also result in more hydrophobic sequences.[35] Uversky,[36] analyzing a large sample of proteins with and without intermediate states, demonstrated that the existence of folding intermediates is strongly dependent on the content of hydrophobic and charged amino acid residues in the protein sequence. Intermediate states were only found in proteins with relatively high hydrophobicity and low net charge.

The role of hydrophobicity in folding kinetics has recently been a subject of intense study. It has been shown that non-native hydrophobic interactions that stabilize on-pathway intermediate states or folding transition states can speed up the folding rate considerably.[37–40] This appears to be in qualitative agreement with a simple theoretical model that demonstrates that a little increase in non-native interactions can accelerate protein folding.[41] However, a too large extent of non-native interactions will increase the ruggedness of the free-energy landscape, with the consequence of slowing down folding and kinetically trapping the protein in off-pathway conformations.[41–43] In natural proteins this extreme behavior is not observed, although off-pathway intermediates have been characterized for several proteins.[44–48]

Intermediate states may induce irreversible aggregation, often mediated by hydrophobic interactions or by the formation of an amyloid beta structure.[49,50] Molecular chaperones assist folding in vivo by sequestrating proteins in misfolded and intermediate states, thus preventing aggregation and speeding up the folding process.[51] Therefore, very hydrophobic globular proteins are expected to experience problems of slow folding and aggregation. Our results concerning the small value of the normalized energy gap for very hydrophobic proteins strongly support this view. Consistent with this is the observation of very high expression of chaperones in insect endosymbionts,[9,10] whose proteins are characterized by very hydrophobic proteins with a very low normalized energy gap.

It is well known that natural proteins are marginally stable, i.e. their unfolding free energy is of the order of k_BT. Two explanations have been proposed for this common feature. The first one states that marginal stability is functionally important because proteins need flexibility in order to function. The other point of view asserts that marginal stability is an unavoidable consequence of molecular evolution, with the selective advantage of improved stability being balanced by the mutational pressure.[52] Our result does not directly apply to this debate, but it may suggest another factor of the observed marginal stability: since we have shown that an improved stability against unfolding is often attained at the expense of stability against misfolding, not only the mutational pressure, but also negative selection against misfolding and potential aggregation problems might prevent protein sequences from reaching very large unfolding free energies.

Genomic *versus* proteomic properties

The main result of this paper is that the normalized energy gap of homologous prokaryotic proteins is positively correlated with the genome size and the G+C composition of the corresponding genes, whereas the unfolding free energy per residue is negatively correlated with both genomic properties. The present analysis confirms our previous finding that the normalized energy gap tends to be lower for proteins encoded in obligatory intracellular bacteria, whose proteins are therefore expected to fold less efficiently than proteins encoded in their free-living relatives.[8]

There are two complementary explanations of the relationship between protein thermodynamics and genomic properties, based on mutational bias and on natural selection, respectively.

Mutational bias

Following Muto & Osawa,[18] we interpret the G+C content at third codon positions as an indicator of the genomic mutational bias rather than of codon usage bias†. Due to the structure of the genetic code, a mutational bias towards A+T at the DNA level

† There are several reasons for this interpretation: First, selection for codon usage is only important for highly expressed proteins and large effective populations. Recent studies found extremely weak indications of selection for codon usage in genes of *B. aphidicola*.[54,73] Second, a recent study has attributed the differences in codon usage between bacterial species to genomic mutational bias.[18] Third, the effect of codon usage on the GC_3 is very reduced when averaging over all possible codons. Fourth, the strong correlation between the base composition at third codon positions and at first plus second codon positions[17,20] is hard to explain by selection, because selection at third positions acts on codon usage while at the other positions it acts on amino acid usage. This correlation is most simply explained by the hypothesis that base composition reflects the mutational bias. Last, the G+C content at third codon positions correlates strongly with the G+C content at pseudogenes and intergenic spacers.

translates into a mutational bias toward more hydrophobic residues at the protein level. This association is confirmed by the correlation between the hydrophobicity of the translated protein and the A+T content at first and second codon position. This is strong for several hydropathy scales and significant for all those we tested, with the only exception of the KD scale that attributes negative or low hydrophobicity to aromatic residues.

Selection strength

For reduced genomes, genome size may be regarded as a quantitative characterization of the extent or age of the intracellular lifestyle of the microbial organism. Bacteria that became intracellular recently, as was probably the case for *Y. pestis* and *M. leprae*, have a larger genome than bacteria that became obligatory intracellular much longer ago. The intracellular lifestyle implies a reduced effective population size due to bottlenecks that occur during transmission to new hosts. Therefore, the efficacy of selection is expected to be reduced,[2,3] particularly in asexual populations lacking effective recombination. The observed correlation between genome size and normalized energy gap is consistent with a weaker purifying selection in intracellular bacteria. This interpretation is also consistent with a computational study of the stability of the 16 S rRNA of aphid endosymbionts[5] and with the extremely high expression level of chaperones observed in these bacteria.[9,10]

These two interpretations are not mutually exclusive. We believe that both of them have to be invoked to explain the observed patterns. In fact, the influence of mutational bias on protein composition ultimately depends on natural selection.

It has been observed in several studies that the amino acid usage is strongly influenced by genomic G+C content,[19,20,50] and this, in the case of endosymbionts of the genus Buchnera, has been mainly attributed to mutational bias.[54,55] The relationship between mutational bias and amino acid composition is best expressed in the scatter plot of GC_{12} *versus* GC_3,[20] Figure 4. The linear correlation is quite strong, $R=0.87$. Through GC_{12}, the GC_3 also influences hydrophobicity (the correlation coefficient goes from $R=-0.20$ to -0.56 for nine out of ten hydropathy scales).

Since hydrophobicity is expected to influence protein folding thermodynamics, as our calculations indicate, one wonders about the selective effect of these global changes in proteomic properties. Lobry,[53] from comparative analysis of bacterial genomes, has suggested the existence of a selective pressure towards "optimal" amino acid frequencies. This selective pressure is visible in Figure 4, in the fact that the variance of the GC_{12} is much narrower than the variance of the GC_3. Nevertheless, purifying selection is not strong enough to remove the influence of the mutational bias on amino acid frequencies. We believe that the balance between unfolding free energy and the normalized energy gap that we have described before plays a crucial role in this respect. Proteins more hydrophobic than average will have larger unfolding free energy but a smaller normalized energy gap. For moderate hydrophobicity, this displacement is probably almost selectively neutral, since it is advantageous for one property but not for the other one.

We expect that very hydrophobic protein sequences bear a selective disadvantage. This is indicated in our calculation by the small value of the normalized energy gap in hydrophobic proteins. Experimentally, very hydrophobic sequences are more likely to have folding intermediate states and are more prone to aggregation. The high expression level of chaperonins in intracellular organisms also suggests the existence of potential folding problems for proteins of these organisms. Therefore, one should expect that the GC_{12} *versus* GC_3 curve is concave upwards for small GC_3, due to selection against highly hydrophobic sequences. Figure 4, however, indicates the contrary: the GC_{12} of proteins of organisms with low GC_3 is even smaller than one would expect from extrapolation of proteins with high and moderate GC_3. We consider this fact a strong indication that selection efficacy is reduced in organisms with low GC_3, as it is expected, since these organisms are mostly obligate pathogens or endosymbionts characterized by reduced population size.

We stress that the interpretation discussed above does not involve a monotonic accumulation of deleterious mutations that unavoidably leads to extinction, which would be in disagreement with the fact that endosymbionts survived for several hundred millions years in their respective hosts. Protein thermodynamics results from a balance between selection, where unfolding free energy and normalized energy gap pull in opposite directions, and mutation. The less stable the protein, the less likely it is that mutations will have a detrimental effect on its stability. Thus, a reduction in selective strength may be compatible with stationary (on the average) values of the folding parameters, even if at a reduced level of stability. This view is shared with a recent population genetic model of compensatory nearly neutral mutations,[56] and it will be the subject of a future investigation.

It has recently been reported that the rates of *in vitro* refolding of orthologous proteins in prokaryotes and eukaryotes correlate with their differential rates of biosynthesis (U. Bastolla & O. Demetrius, unpublished results). In this experimental work, the authors concluded that faster folding in prokaryotes, in which chain elongation is faster, minimizes the occurrence of unfolded nascent proteins. In endosymbiotic bacteria protein synthesis is thought to be slower than in free-living bacteria. An indication of this is the weak codon usage bias towards more frequent tRNA species in endosymbiotic bacteria.[54] Another indication for this is their very slow growth rate. This suggests an

alternative explanation for our observation of reduced protein folding efficiency in intracellular bacteria: it is possible that it is due to weaker selection on folding efficiency when the protein chain is translated very slowly. In order to contrast this possible interpretation, we have examined eukaryotic proteins, which are also synthesized slower than proteins of free-living bacteria, and we have noticed that their normalized energy gap is larger than that of intracellular bacterial proteins. Therefore, the interpretation that relates the low folding efficiency of proteins of intracellular bacteria to their slow biosynthesis rate does not appear to be supported by the data.

Conclusion

In summary, both mutational pressure and variable selective strength appear responsible for the systematic differences between orthologous proteins of different bacteria, sharing the same structure and function but having different thermodynamic properties. A mutational bias towards A+T at the DNA level translates into a bias towards more hydrophobic proteins, which are characterized by larger unfolding free energies but lower stability against misfolding. Probably these two opposite effects almost balance for moderately hydrophobic proteins, so that purifying selection cannot avoid the influence of the mutational bias on the amino acid frequencies. We expect, however, that a too large hydrophobicity has a negative effect on protein folding efficiency and on fitness, particularly due to the increased risk of protein aggregation. This expectation is confirmed by the very low value of the normalized energy gap. In contrast, the influence of the mutational bias on the protein composition, as indicated by the slope of the GC_{12} *versus* GC_3 curve, is even larger for large bias than for moderate bias, as if natural selection were less able to counteract the effect of mutational bias. The most plausible explanation seems that the efficacy of purifying selection is reduced in the organisms characterized by a strong AT bias. These organisms are mostly intracellular bacteria, they have a small effective population size and their genomes do not recombine. They are also characterized by a very reduced genome size, which can be regarded as another quantitative characterization of the extent of their intracellular lifestyle.

It remains to be explained why intracellular lifestyle and low G+C content always tend to go together. The influence that mutational bias and reduced selection exert on protein thermodynamics is probably part of the answer.

Our results are also relevant for the studies of protein folding. They confirm the importance of the normalized energy gap, a crucial parameter in the statistical mechanical models of protein folding. Through natural selection, naturally occurring proteins are different from random heteropolymers, for which the normalized energy gap vanishes, but species in which selection is weaker have proteins that seem to dangerously approach random heteropolymers. Thus this computational approach provides a bridge between population genetics and molecular evolution on the one hand and the statistical mechanics of biological macromolecules on the other.

Materials and Methods

Protein families

We selected a total of 21 families of small homologous proteins for which at least one structure is known, and which are also present in the reduced genomes of obligatory intracellular bacteria. They are listed in Table 3. In addition, we studied the Chaperone DnaK, which was not used to calculate average properties, since its function in assisting protein folding sets it apart from the other 21 families (see below).

For each protein family we included sequences from the PFAM database,[14] complemented with sequences from the TIGRFAM database of homologous families of completed prokaryotic genomes†. Only the best match to the family from each organism was considered. In case of almost equivalent matches, we chose the sequence that provided the largest value of the energy gap.

Prokaryotic species

The following 73 prokaryotic species were studied. Archaea: *Aeropyrum pernix*, *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Pyrococcus furiosus*. Thermophylic bacteria: *Aquifex aeolicus*, *Bacillus stearothermophilus*, *Thermotoga maritima*, *Thermus aquaticus*, *Thermoanaerobacter tengcongensis*. Free-living bacteria: *Agrobacterium tumefaciens*, *Bacillus anthracis*, *B. subtilis*, *B. halodurans*, *Brucella melitensis*, *Caulobacter crescentus*, *Chlorobium tepidum*, *Clostridium acetobutylicum*, *C. perfringens*, *Corynebacterium glutamicum*, *Deinococcus radiodurans*, *Escherichia coli* K12, *E. coli* O157, *Enterococcus faecalis*, *Fusobacterium nucleatum*, *Haemophilus influenzae*, *Lactococcus lactis*, *Listeria innocua*, *L. monocytogenes*, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Nostoc sp.*, *Pasteurella multocida*, *Pseudomonas aeruginosa*, *P. putida*, *P. syringae*, *Ralstonia solanacearum*, *Rhizobium loti*, *R. meliloti*, *Salmonella typhimurium*, *Shewanella oneidensis*, *Shigella flexneri*, *Staphylococcus aureus*, *Streptomyces coelicolor*, *Synechocystis sp.*, *Vibrio cholerae*, *Xylella fastidiosa*, *Xanthomonac axonopodis*, *Zymomonas mobilis*. Obligatory endosymbionts and parasites: *Borrelia burgdorferi*, *Campylobacter jejuni*, *Chlamydia pneumoniae*, *C. muridarum*, *C. trachomatis*, *Coxiella burnettii*, *Helicobacter pylori*, *Mycoplasma capricolum*, *M. genitalium*, *M. pneumoniae*, *Mycobacterium leprae*, *Treponema pallidum*, *Ureaplasma parvum*, *Yersinia pestis*, *Wigglesworthia glossinidia*, *Wolbachia sp.*, *Buchnera aphidicola* from the aphid hosts *Baizongia pistacea*, *Schizaphis graminum* and *Acyrthosiphon pisum*. We note that several of the free-living bacteria are opportunistic parasites.

In addition, homologous proteins from five algal chloroplasts (*Cyanidium caldarium*, *Guillardia theta*, *Euglena gracilis*, *Porphyra purpurea*) and from six

† The Institute for Genomic Research, url: http://www.tigr.org/TIGRFAMs/index.shtml

Table 3. Protein families studied, PFAM code and representative proteins in the PDB

Protein	Gene	PFAM	Representative structures	Length
Acetate kinase (dom.2)	ackA	PF00871	1g99 (ACKA_METTE)	241
ATP synthase ε	AtpC	PF00401	1aqt (ATPE_ECOLI)	138
Chaperone protein dnaK	dnaK	PF00012	1dkz(DNAK_ECOLI)	603
D-ala D-ala ligase	Ddl	PF07478 PF01820	1iov (DDLB_ECOLI), 1e4eB (VANA_ENTFC), 1ehiA (DDL-LEUME)	305–377
Citrate synthase	gltA	PF00285	1k3p (CYSY_ECOLI), 1a59 (CISY_ABDS2), 1iom (Q72J03), 1aj8 (CISY_PYRFU), 1o7× (CISY_SULSO), 6csc (CISY_CHICK), 1cts (CISY_PIG)	371–433
3-Dehydroquinase	Aroq	PF001220	1d0iE (AROQ_STRCO), 2dhqA (AROD_MYCTU)	146–156
Dihydrofolate reductase	folA	PF00186	1ra9 (DYR ECOLI), 3dfr (DYR_LACCA), 1df7 (DYR_MYCTU), 1d1g (DYR_THEMA), 1vdr (DYR_HALVO), 1dyr (DYR_PNECA), 1drf (DYR_HUMAN), 8dfr (DYR_CHICK)	159–206
DUTPase	dut	PF00692	1euw (DUT_ECOLI)	151
Elongation factor TS	efts	PF00889	1efuB (EFTS_ECOLI), 1tfe (EFTS_THETH)	142–207
Flavodoxin	flav	PF00258	1ag9 (FLAV_ECOLI), 1f4p (FLAV_DESVH), 1czn (FLAV_SYNP7), 1rcf (FLAV_ANASP), 5nul (FLAV_CLOBE), 1fue (FLAV_HELPJ)	138–175
Peptide deformylase	def	PF01327	1g2a (DEF_ECOLI)	168
Peptidyl-tRNA hydrolase	pth	PF01195	2pth (PTH_ECOLI)	194
Phosphocarrier protein H	PtsH	PF00381	1opd (PTHP_ECOLI), 2hpr (PTHP_BACSU), 1ptf (PTHP_STRFE), 1pch (PTHP_MYCCA)	85–88
Phosphopantetheine adenyltransferase	coad	PF01467	1b6tA (COAD_ECOLI)	159
50 S ribosomal protein L14	rl14	PF00238	1whi (RL14_BACST), 1jj2J (RL14_HALMA)	122–132
Ribosomal methyltransferase J	ftsJ	PF01728	1ej0 (RRMJ_ECOLI)	180
RNase H	rnhA	PF00075	2rn2 (RNH_ECOLI), 1ril (RNH_THETH)	155–166
RNase P	RnpA	PF00825	1a6f (RNPA_BACSU), 1d6t (RNPA_STAAW)	116–117
Thioredoxin I	trxA	PF00085	2trx (THIO_ECOLI), 1thx (THI2_ANASP), 1fb6 (THIM_SPIOL), 1dby (THIM_CHLRE), 1ep7 (THIH_CHLRE), 1erv (THIO_HUMAN)	105–140
Thioredoxin	trxB	PF00070	1trb (TRXB_ECOLI), 1fl2 (AHPF_ECOLI), 1vdc (TRB1_ARATH)	320–330
Triosephosphate isomerase	Tpi	PF00121	1tre (TPIS_ECOLI), 1aw2 (TPIS_VIBMA), 2btm (TPIS_BACST), 1b9b (TPIS_THEMA), 1hg3 (TPIS_PYRWO), 1ydv (TPIS_PLAFA), 1tpf (TPIS_TRYBB), 1tcd (TPIS_TRYCR), 1amk (TPIS_LEIME), 7tim (TPIS_YEAST), 1hti (TPIS_HUMAN), 1tph (TPIS_CHICK)	225–255
Triptophan synthase α chain	TrpA	PF00290	1qopA (TRPA_SALTY), 1geq (TRPA_PYRFU)	248–268

eukaryotic nuclear genomes (*Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Spinacia oleracea*, *Homo sapiens*, *Caenorhabditis elegans*) have also been examined when available.

Protein model and effective free energy

The protein model that we adopt is described in detail in the paper by Bastolla *et al*.[57] Briefly, protein structures are represented as contact matrices such that C_{ij} equals one if residues at i and j are in contact and zero otherwise. Given a sequence $\mathbf{S}=\{S_1,\ldots,S_N\}$ and its contact matrix $\mathbf{C}$, the configurational free energy, including the entropy of the solvent, but not chain entropy, is assumed to have the form of the sum of contact interactions, $E(\mathbf{C};\mathbf{S})/k_BT=\sum C_{ij}U(S_i,S_j)$, where $U(a;\ b)$ is the 20×20 interaction matrix in the work done by Bava *et al*.[58]

The calculated unfolding free energy of the protein folded in the native configuration $\mathbf{C}_{nat}$, with respect to the unfolded state, is estimated as $\Delta G(\mathbf{C}_{nat};\mathbf{S})=k_BT=-E(\mathbf{C}_{nat},\mathbf{S})/k_BT+Ns$, where N is chain length and s is a term representing the configurational entropy per residue minus the stabilization due to the hydrogen bonding network of secondary structure. A fit of the above equation to the unfolding free energies of 44 globular proteins with two-states folding thermodynamics and no disulphide bonds from the database Protherm[58] suggests that s is almost constant for different protein topologies and indicates that $\Delta G(\mathbf{C}_{nat},\mathbf{S})$ is strongly correlated with the experimental free energy, with correlation coefficient $R=0.95$. The correlation is still rather strong ($R=0.85$) if the two quantities are divided by chain length (UB, unpublished data). Therefore we use the quantity $\Delta G(\mathbf{C}_{nat};\mathbf{S})=Nk_BT$ as an estimate of the unfolding free energy per residue.

Fold recognition

For most of the sequences studied the native structure was not known, but the structure of one or more orthologous proteins were available in the Protein Data Bank (PDB). Alignments were generated by threading sequences on the more than 6000 structures present in the last release of the PDBselect90 database.[59] The best candidate native structure was identified by minimizing the sum of the effective energy plus a gap penalty term. Alternative low energy structures were then used to calculate the energy gap. We applied two different alignment strategies, which gave the same qualitative results.

Fixed sequence-structure alignment

For each protein family, we chose a representative

sequence of known structure and aligned all other sequences to it. For each sequence, we then forced structures in the PDB to follow the same pattern of alignment with predetermined gaps, which were not penalized. In nearly all cases, the representative structure was correctly identified as the native state. In a few cases, manual improvement of the alignments was necessary to obtain this result.

Optimized sequence-structure alignment

In this approach, for each possible sequence structure alignment the placement of gaps was optimized by minimizing the effective free energy function plus a gap penalty term, using the program Protfinder†. In most cases, the structure of one of the orthologous proteins was recognized as the best model. The few sequences for which this did not happen were discarded. These were mostly sequences with many gaps in the alignment with structures of homologous proteins and sequences belonging to genomes characterized by low energy gaps in encoded proteins. Despite the fact that our alignment score is not based on sequence identity, the alignments that we obtained were often coincident, and sometimes better than the ones obtained by sequence based methods. Because the method does not require manual inspection of the alignments, it could be applied to a larger number of proteins. We report here only results obtained through this method. The fixed alignment method gave qualitatively identical results, which were partially reported in the work done by van Ham.[8]

We discarded proteins for which a gap in the alignment obtained by the fold recognition algorithm was larger than 15% of the sequence. Such gaps correspond to insertion or deletion of long loops, terminal regions or even entire domains. We also discarded proteins for which the fold recognition algorithm provided a structural model not derived from a homologous protein.

Normalized energy gap

Statistical mechanical studies of protein folding have shown that fast folding and thermodynamic stability of the native state require that the free energy landscape of the protein chain is smooth.[59–64] We use the normalized energy gap α[64] as a quantitative measure of the smoothness of the free energy landscape. A large value of α implies that all low energy structures are very similar to the native one. This is a measure of the stability with respect to misfolded states. If α is very small, limited thermodynamic stability of the native state, limited stability against mutations, slow folding, misfolding and aggregation problems are expected. As a measure of structural similarity we use the overlap $q(\mathbf{C};\mathbf{C}')$, which counts the number of common contacts between two structures $\mathbf{C}$ and $\mathbf{C}'$, normalized so that $q(\mathbf{C};\mathbf{C}')$ equals one if and only if $\mathbf{C}$ and $\mathbf{C}'$ share all of their contacts. The free energy landscape is said to be smooth if configurations very different from the native one, $\mathbf{C}_{\mathrm{nat}}$, all have high energy. This is a prerequisite for stability and fast folding. The normalized energy gap α is defined through the set of inequalities:

$$\frac{E(\mathbf{C};\mathbf{S}) - E(\mathbf{C}_{\mathrm{nat}},\mathbf{S})}{|E(\mathbf{C}_{\mathrm{nat}},\mathbf{S})|} \geq \alpha(\mathbf{S})(1 - q(\mathbf{C};\mathbf{C}_{\mathrm{nat}})) \qquad (1)$$

where $\mathbf{C}$ is any alternative configuration. A large value of α is also a prerequisite for successful fold recognition.

† http://www.cab.inta.es/~CAFASP

Sequence similarity

The estimates of thermodynamic parameters depend in part on the protein giving the best structural match in the fold recognition. They are inadequately estimated if a considerable fraction of the sequence cannot be aligned to the model structure. We therefore discarded all proteins for which less than 85% of the sequence could be aligned. Sequence similarity with the best structural model also has a strong effect. Since the difference between the structures of two homologous proteins increases with their sequence dissimilarity,[65] the less similar the sequences, the worse the structural model will be. This is an unavoidable problem. However, despite the fact that our results are not fully reliable for every individual protein, we expect that the correlations that they allow to discover remain valid when experimentally measured quantities are used instead of our computational estimates.

Interactivity and hydrophobicity

We have correlated genomic and folding thermodynamic properties with the mean values of ten amino acid properties $h(\alpha)$; $a=1,\ldots,20$. All of the studied properties are related to hydrophobicity. Eight of them represent empirical or computational hydrophobicity scales, the last two are related to the typical interaction strength and the typical connectivity of each amino acid type. These last two scales yield the strongest correlations both with folding thermodynamics and with genomic properties. We refer to these ten properties collectively with the term interactivity, because they try to measure how strongly the amino acid residues in the protein sequence interact with each other.

The properties that we considered are: (1) the L76 hydropathy scale derived by Levitt in 1976 using experimental data and theoretical calculations;[66] (2) the KD hydropathy scale, derived by KD in 1982 to identify trans-membrane helices using diverse experimental data;[67] (3) the FP hydropathy scale derived by Fauchere & Pliska in 1983 from the experimental measurement of octanol/water partition coefficients;[68] (4) the R88 hydropathy scale derived by Roseman in 1988 based on the transfer of solutes from water to alkane solvents;[69] (5) the MP hydropathy scale, derived by Manavalan & Ponnuswamy in 1978 from statistical properties of globular proteins;[70] (6) the augmented Whilmey White hydropathy scale WW, derived by Jayasinghe *et al.* in 2001 to improve recognition of transmembrane helices;[71] (7) the AV hydropathy scale derived by Palliser & Parry in 2001 by averaging 127 normalized hydropathy scales published in the literature;[72] (8) the G98 classification of amino acid residues into polar, hydrophobic and amphiphilic classes, adopted by Gu *et al.*[12] to investigate the relationship between the hydrophobicity of a protein and the nucleotide composition of the corresponding gene; (9) the interaction scale IH obtained from the main eigenvector of the interaction matrix $U(a;b)$ used here. It is known in fact that the main eigenvector of contact interaction matricies is strongly related to hydrophobicity.[74] (10) Last, the connectivity scale CH, that maximizes the correlation with the principal eigenvectors of protein contact matrices for a non-redundant set of PDB structures (U. Bastolla *et al.*, unpublished results).

All these scales are positively correlated with each

other, with correlation coefficients ranging from a minimum of 0.68 between the KD and L76 scales to a maximum of 0.95 between IH and CH scales.

For each protein, we calculated the mean interactivity as $H(S) = \sum h(S_i)/N$ and correlate it to genomic and thermodynamic properties.

Averages

The effective free energy per residue in our model depends strongly on the number of contacts per residue N_c/N and on chain length, two structural properties that are almost constant in proteins of the same family. Prior to averaging the properties of proteins belonging to different families and contained in the same genome, properties of individual proteins were normalized by dividing them by the corresponding properties of the protein representative of the family. This allowed to underline the effects of sequence evolution with respect to the effects of protein topology. The representative protein was chosen as the sequence with largest normalized energy gap. Other choices (for instance proteins from model organisms) yielded the same qualitative results. The normalized properties of the proteins of the same organism were then averaged. Only organisms for which we found at least five suitable proteins were considered in the study.

For each organism, we calculated the average of the following normalized protein properties: (1) unfolding free energy per residue; (2) normalized energy gap; (3) mean interactivity (in this case we added an offset so that all mean interactivities result positive, as explained below); (4 and 5) G+C content of the corresponding gene, distinguishing between first and second codon position on one hand and third on the other hand; (6) sequence similarity with the representative structure.

The normalization was to be performed with caution in the case of interactivity, since some hydropathy scales designed to identify transmembrane helices attribute zero mean hydrophobicity to globular proteins. In this case, different proteins within the same family may have interactivities with different sign, and the normalization may completely obscure the results. To overcome this problem, we add a constant offset to all hydrophobicity scales and rescale them in such a way that the hydrophobicity is positive for all proteins in our set. We choose the offset and the scale factor so that the average of the mean hydrophobicity and the mean hydrophobicity squared over all proteins in our set are the same for all of the scales considered.

Principal component analysis

PCA selects the combinations of variables explaining most of the variance in multivariate data sets. For each organism, we considered seven variables: the variables (1–6) listed above plus genome size. The use of PCA allowed to reduce the unwanted effect of sequence similarity on our results.

Acknowledgements

U.B. thanks Javier Tamames for introducing him to this subject. During this work, U.B., E.V. and R.C.H.J.vH. have been supported through grants from INTA (Spain). U.B. has been partly supported through the I3P Network on Bioinformatics of the CSIC (Spain), financed by the European Social Fund. A.M. has been supported through grant BMC2003-00305 from Ministerio de Ciencia y Tecnologia (MiCyt), Spain.

References

1. Grantcharova, V., Alm, E. J., Baker, D. & Horwich, A. L. (2001). Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **11**, 70–82.
2. Ohta, T. (1976). Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Popul. Biol.* **10**, 254–275.
3. Muller, H. J. (1964). The relation of the recombination to mutational advance. *Mutat. Res.* **1**, 2–9.
4. Moran, N. A. (1996). Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA*, **95**, 4458–4462.
5. Lambert, D. J. & Moran, N. A. (1998). Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA*, **95**, 4458–4462.
6. Itoh, T., Martin, W. & Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular bacteria. *Proc. Natl Acad. Sci. USA*, **99**, 12944–12948.
7. Canbäack, B., Tamas, I. & Andersson, S. G. E. (2004). A phylogenetic study of endosymbiotic bacteria. *Mol. Biol. Evol.* In the press.
8. van Ham, R. C. H. J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U. *et al*. (2003). Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA*, **100**, 581–586.
9. Ishikawa, H. (1984). Characterization of the protein species synthetized *in vivo* and *in vitro* by an aphid endosymbiont. *Insect Biochem.* **14**, 417–425.
10. Aksoy, S. (1995). Molecular analysis of the endosymbionts of tsetse flies: 16S rDNA locus and over-expression of a chaperonin. *Insect Mol. Biol.* **4**, 23–29.
11. Fares, M. A., Ruiz-Gonzalez, M. X., Moya, A., Elena, S. F. & Barrio, E. (2002). GroEL buffers against deleterious mutations. *Nature*, **417**, 398.
12. Gu, X., Hewett-Emmett, D. & Li, W. H. (1998). Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*, **102–103**, 383–391.
13. D'Onofrio, G., Jabbari, K., Musto, H. & Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, **238**, 3–14.
14. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000). The PFAM contribution to the annual NAR database issue. *Nucl. Acids Res.* **28**, 263–266.
15. Widmann, M. & Christen, P. (2000). Comparison of folding rates of homologous prokaryotic and eukaryotic proteins. *J. Biol. Chem.* **275**, 18619–18622.
16. Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.* **151**, 389–409.
17. Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA*, **84**, 166–169.

18. Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. & McAdams, H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl Acad. Sci. USA*, **101**, 3480–3485.
19. Sueoka, N. (1961). Correlation between base composition of the deoxyribonucleic acid and amino acid composition of proteins. *Proc. Natl Acad. Sci. USA*, **47**, 469–478.
20. Bernardi, G. & Bernardi, G. (1985). Codon usage and genome composition. *J. Mol. Evol.* **24**, 1–11.
21. Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.
22. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–63.
23. Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.* **226**, 29–35.
24. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J.* **10**, 75–83.
25. Jaenicke, R. & Bohm, G. (1998). The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**, 738–748.
26. Kumar, S., Tsai, C. J. & Nussinov, R. (2001). Thermodynamic differences among homologous thermophilic and mesophilic proteins. *Biochemistry*, **40**, 14152–14165.
27. Zhou, H. X. (2002). Toward the physical basis of thermophilic proteins: linking of enriched polar interactions and reduced heat capacity of unfolding. *Biophys. J.* **83**, 3126–3133.
28. Hunter, C. A., Singh, J. & Thornton, J. M. (1991). Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. *J. Mol. Biol.* **218**, 837–846.
29. Kannan, N. & Vishveshwara, S. (2000). Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng.* **13**, 753–761.
30. Gallivan, P. G. & Dougherty, D. A. (1999). Cation–π interactions in structural biology. *Proc. Natl Acad. Sci. USA*, **96**, 9459–9464.
31. Zaric, S. D., Popovic, D. M. & Knapp, E. W. (2000). Metal ligand aromatic cation–pi interactions in metalloproteins: ligands coordinated to metal interact with aromatic residues. *Chemistry*, **6**, 3935–3942.
32. Gromiha, M. M., Thomas, S. & Santhosh, C. (2002). Role of cation–pi interactions to the stability of thermophilic proteins. *Prep. Biochem. Biotechnol.* **32**, 355–362.
33. Shakhnovich, E. I. & Gutin, A. M. (1989). Formation of unique structure in polypeptide chains. *Biophys. Chem.* **34**, 187.
34. Bastolla, U., Frauenkron, H. & Grassberger, P. (2000). Phase diagram of random heteropolymers: replica approach and application of a new Monte Carlo algorithm. *J. Mol. Liq.* **84**, 111–129.
35. Spudich, G. M., Miller, E. J. & Marqusee, S. (2004). Destabilization of the *Escherichia coli* RNase H kinetic intermediate: switching between a two-state and three-state folding mechanism. *J. Mol. Biol.* **335**, 609–618.
36. Uversky, V. N. (2002). Cracking the folding code Why do some proteins adopt partially folded conformations, whereas other don't? *FEBS Letters*, **514**, 181–183.
37. Viguera, A. R., Vega, C. & Serrano, L. (2001). Unspecific hydrophobic stabilization of folding transition states. *Proc. Natl Acad. Sci. USA*, **99**, 5349–5354.
38. Northey, J. G., Di Nardo, A. A. & Davidson, A. R. (2002). Hydrophobic core packing in the SH3 domain folding transition state. *Nature Struct. Biol.* **9**, 126–130.
39. Calloni, G., Taddei, N., Plaxco, K. W., Ramponi, G., Stefani, M. & Chiti, F. (2003). Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. *J. Mol. Biol.* **330**, 577–591.
40. Feng, H., Takei, J., Lipsitz, R., Tjandra, N. & Bai, Y. (2003). Specific non-native hydrophobic interactions in a hidden folding intermediate: implications for protein folding. *Biochemistry*, **42**, 12461–12465.
41. Plotkin, S. S. (2001). Speeding protein folding beyond the Go model: how a little frustration sometimes helps. *Proteins: Struct. Funct. Genet.* **45**, 337–345.
42. Bringelson, J. D. & Wolynes, P. G. (1987). Spin-glasses and the statistical-mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.
43. Klimov, D. K. & Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins: Struct. Funct. Genet.* **26**, 411–441.
44. Bilsel, O., Zitzewitz, J. A., Bowers, K. E. & Matthews, C. R. (1999). Folding mechanism of the alpha-subunit of tryptophan synthase, an alpha/beta barrel protein: global analysis highlights the interconversion of multiple native, intermediate, and unfolded forms through parallel channels. *Biochemistry*, **38**, 1018–1029.
45. Bhattacharyya, A. M. & Horowitz, P. M. (2001). The aggregation state of rhodanese during folding influences the ability of GroEL to assist reactivation. *J. Biol. Chem.* **276**, 28739–28743.
46. Fernandez-Recio, J., Genzor, C. G. & Sancho, J. (2001). Apoavodoxin folding mechanism: an alpha/beta protein with an essentially off-pathway intermediate. *Biochemistry*, **40**, 15234–15245.
47. Hoyer, W., Ramm, K. & Pluckthun, A. (2002). A kinetic trap is an intrinsic feature in the folding pathway of single-chain Fv fragments. *Biophys. Chem.* **96**, 273–284.
48. Baldwin, R. L. (1996). On-pathway *versus* off-pathway folding intermediates. *Fold. Des.* **1**, R1–R8.
49. King, J., Haase-Pettingell, C., Robinson, A. S., Speed, M. & Mitraki, A. (1996). Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates. *FASEB J.* **10**, 57–66.
50. Chow, M. K., Lomas, D. A. & Bottomley, S. P. (2004). Promiscuous beta-strand interactions and the conformational diseases. *Curr. Med. Chem.* **11**, 491–499.
51. Agashe, V. R. & Hartl, F. U. (2000). Roles of molecular chaperones in cytoplasmic protein folding. *Semin. Cell Dev. Biol.* **11**, 15–25.
52. Taverna, D. M. & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins: Struct. Funct. Genet.* **46**, 105–109.
53. Lobry, J. R. (1997). Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**, 309–316.
54. Rispe, C., Delmotte, F., van Ham, R. C. H. J. & Moya, A. (2004). Mutational and selective pressure on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res.* **14**, 44–53.
55. Palacios, C. & Wernegreen, J. J. (2002). A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high-expression genes. *Mol. Biol. Evol.* **19**, 1575–1584.
56. Hartl, D. L. & Taubes, C. H. (1998). Towards a theory of evolutionary adaptation. *Genetica*, **102–103**, 525–533.

57. Bastolla, U., Farwer, J., Knapp, E. W. & Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins: Struct. Funct. Genet.* **44**, 79–96.
58. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucl. Acids Res.* **32**, D120–D121.
59. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structure. *Protein Sci.* **3**, 522–524.
60. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA*, **89**, 4918–4922.
61. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Free energy landscapes for protein folding kinetics—intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062.
62. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995). Evolution-like selection of fastfolding model proteins. *Proc. Natl Acad. Sci. USA*, **92**, 1282–1286.
63. Bastolla, U., Frauenkron, H., Gerstner, E., Grassberger, P. & Nadler, W. (1998). Testing a new Monte Carlo algorithm for protein folding. *Proteins: Struct. Funct. Genet.* **32**, 52–66.
64. Bastolla, U., Roman, H. E. & Vendruscolo, M. (1999). Structurally constrained protein evolution: results from a lattice simulation. *Eur. Phys. J. B*, **15**, 385–397.
65. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
66. Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
67. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
68. Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters of amino acid side chain from the partitioning N-acetyl amino acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
69. Roseman, M. A. (1988). Hydrophobicity of polar amino-acid side chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* **200**, 513–522.
70. Manavalan, P. & Ponnuswamy, P. K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature*, **275**, 673–674.
71. Jayasinghe, S., Hristova, K. & White, S. H. (2001). Energetics, stability, and prediction of transmembrane helices. *J. Mol. Biol.* **312**, 927–934.
72. Palliser, C. C. & Parry, D. A. (2001). Quantitative comparison of the ability of hydropathy scales to recognize surface beta-strands in proteins. *Proteins: Struct. Funct. Genet.* **42**, 243–255.
73. Wernegreen, J. J. & Moran, N. (1999). Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes. *Mol. Biol. Evol.* **16**, 83–97.
74. Li, H., Tang, C. & Wingreen, N. S. (1997). Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Letters*, **79**, 765–768.

Edited by J. Thornton

(Received 4 May 2004; received in revised form 24 August 2004; accepted 27 August 2004)