OPEN ACCESS Freely available online

Mutation Bias Favors Protein Folding Stability in the Evolution of Small Populations

Raul Mendez[1☯], Miriam Fritsche[2☯¤a], Markus Porto[2*¤b], Ugo Bastolla[1*]

1 Centro de Biología Molecular "Severo Ochoa", Consejo Superior de Investigaciones Científicas and Universidad Autónoma de Madrid, Madrid, Spain, **2** Institut für Festkörperphysik, Technische Universität Darmstadt, Darmstadt, Germany

Abstract

Mutation bias in prokaryotes varies from extreme adenine and thymine (AT) in obligatory endosymbiotic or parasitic bacteria to extreme guanine and cytosine (GC), for instance in actinobacteria. GC mutation bias deeply influences the folding stability of proteins, making proteins on the average less hydrophobic and therefore less stable with respect to unfolding but also less susceptible to misfolding and aggregation. We study a model where proteins evolve subject to selection for folding stability under given mutation bias, population size, and neutrality. We find a non-neutral regime where, for any given population size, there is an optimal mutation bias that maximizes fitness. Interestingly, this optimal GC usage is small for small populations, large for intermediate populations and around 50% for large populations. This result is robust with respect to the definition of the fitness function and to the protein structures studied. Our model suggests that small populations evolving with small GC usage eventually accumulate a significant selective advantage over populations evolving without this bias. This provides a possible explanation to the observation that most species adopting obligatory intracellular lifestyles with a consequent reduction of effective population size shifted their mutation spectrum towards AT. The model also predicts that large GC usage is optimal for intermediate population size. To test these predictions we estimated the effective population sizes of bacterial species using the optimal codon usage coefficients computed by dos Reis *et al.* and the synonymous to non-synonymous substitution ratio computed by Daubin and Moran. We found that the population sizes estimated in these ways are significantly smaller for species with small and large GC usage compared to species with no bias, which supports our prediction.

Citation: Mendez R, Fritsche M, Porto M, Bastolla U (2010) Mutation Bias Favors Protein Folding Stability in the Evolution of Small Populations. PLoS Comput Biol 6(5): e1000767. doi:10.1371/journal.pcbi.1000767

Editor: Eugene I. Shakhnovich, Harvard University, United States of America

Received September 25, 2009; **Accepted** March 30, 2010; **Published** May 6, 2010

Funding: UB acknowledges financial support from the Spanish Science and Innovation Ministry through the Ramón y Cajal program and through the projects BIO2008-04384 and CSD2006-00023, and a stay at the Aspen Center for Physics where a first version of this work was written. Our collaboration was facilitated through the program "Acciones Integradas Espana-Alemania" of the Spanish Science and Innovation Ministry, project HA2006-0044, and of the Deutscher Akademischer Austauschdienst project D/06/12848. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ubastolla@cbm.uam.es; porto@thp.uni-koeln.de

¤a Current address: Institut für Theoretische Physik, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany
¤b Current address: Institut für Theoretische Physik, Universität zu Köln, Köln, Germany

☯ These authors contributed equally to this work.

Introduction

The quantitative modeling of molecular evolution is of key importance for reconstructing evolutionary histories, as well as for understanding how the properties of natural macromolecules are influenced by their evolution. Already for a long time population size has been recognized as a crucial factor that influences both the evolutionary process and the stability that macromolecules can attain. On the other hand, even if mutation bias in prokaryotes varies from extreme GC rich to extreme AT rich, its influence on the evolutionary process, the stability of evolving macromolecule, and on the fitness of the population has received much less attention. Here, we simulate an evolutionary model that combines population size, GC mutation bias, and protein folding stability, and we show the deep interplay between these variables.

Kimura's neutral model [1,2] is still one of the most influential models of molecular evolution. This model considers all viable macromolecules as equally fit and all the others as nonviable. Within this neutral model, the functional properties of the evolving macromolecules, in particular their folding stability, are independent of population size and, by entropy arguments, they are expected to coincide with the minimal properties compatible with viable molecules [3]. If mutations with small fitness effects are included in the model, population size N becomes a key variable of the evolutionary process, since slightly deleterious mutations are more likely to be fixed in small populations [4–6]. This study has been pioneered by Ohta, who showed that population size can provide a possible explanation for empirical observations such as the generation time effect [7,8]. Obligate intracellular lifestyle, such as that of endosymbiotic or parasitic bacteria, implies a strong reduction in effective population size due to bottlenecks upon transmission from one host to another. Inspired by Ohta's theory, computational studies have compared bacterial species displaying an obligate intracellular lifestyle with their free living relatives, suggesting that the genes of intracellular bacteria evolve faster as a result of relaxed selection [9] (but Itoh *et al.* [10] give a different

Author Summary

The Guanine plus Cytosine (GC) content of bacterial genomes varies from 20% to 80%. This variation is attributed to the mutation bias produced by replication and repair machinaries. However, the evolutionary forces that act on these very different machinaries have remained elusive. It is known that the GC content of genes strongly influences the resulting proteins' hydrophobicity, which is the main determinant of folding stability. This may lead to expectation that the GC content is strongly selected at its optimal value, since proteins that are too hydrophylic face unfolding problems and proteins that are too hydrophobic face misfolding and aggregation problems. In this work, using a realistic model of genotype (DNA sequence) to phenotype (protein folding stability) to fitness mapping and a standard population genetics model, we find that the optimal GC usage depends on population size. In particular, very small populations prefer small GC usage, intermediate populations prefer large GC usage, and large populations prefer no bias. Our results may explain why most intracellular bacteria, evolving with small effective populations, tend to adopt small GC usage. To test this hypothesis, we estimated the effective population size of several bacterial species, finding that those that evolve with 50% GC usage are characterized by significantly larger populations, although several exceptions exist.

interpretation) and that their structural RNAs [11] and their proteins [12] are less stable than the orthologous macromolecules of free living bacteria. Evolution experiments with virus and bacteria confirm the influence of small population size, demonstrating fitness loss in populations evolving under repeated bottlenecks [13,14], and show that such a loss can be partly compensated by over-expressing chaperones that assist protein folding [15]. These findings support the idea that fitness is reduced in small populations as a consequence of the reduction of protein folding stability. Recent theoretical work has shown that, in the appropriate limits, the statistical properties of population genetics are formally equivalent to a statistical mechanical system, so that there is an exact analogy between the reduction of fitness for small populations and the increase of entropy for large temperature [16,17]. In the present study, we will exploit this correspondence to get analytic insight into non-neutral evolution.

Another key evolutionary variable, which however has received little attention, is the nucleotide spectrum. In prokaryotic genomes, it varies from extreme adenine plus thymine (AT) content in obligatory intracellular bacteria to extreme guanine plus cytosine (GC) content, for instance in actinobacteria. These differences in GC content are prevalently thought to be due to mutation bias [18,19]. They are strongest at the third codon position, where GC content barely affects the amino acid composition of the protein, but also influence the coding positions [20,21]. Due to the structure of the genetic code, a mutation bias favoring thymine at the nucleotide level favors the incorporation of hydrophobic amino acids in the translated protein [12,22]. Hydrophobicity is a key property for protein folding [23]. Proteins that are too hydrophylic tend to be naturally unfolded, whereas proteins that are too hydrophobic tend to misfold and aggregate [24]. This qualitative trade-off between unfolding and misfolding was confirmed by a computational study of the properties of homologous proteins in the proteomes of several bacterial species, using a model of protein folding stability that correlates well with experimentally measured unfolding stabilities [12]. In previous work, two of us and colleagues investigated the relationship between unfolding stability, misfolding stability and mutation bias using a protein evolution model with a realistic genotype (DNA sequence) to phenotype (folding stability) mapping in a neutral fitness landscape in which all proteins with stabilities above thresholds have the same fitness. We found that the mutation bias modulates the trade-off between the two kinds of stability, making proteins evolving under AT mutation bias more stable against unfolding but less stable against misfolding [25].

Interestingly, the two aspects discussed above, small population size and mutation bias towards AT, are strongly correlated in nature. In fact, most bacterial and eukaryotic lineages that adopted an intracellular lifestyle, with consequent reduction of their effective population size, also shifted their mutation spectrum towards AT [26], as indicated by the strong correlation between reduced genome size, which is a signature of intracellularity, and the AT bias [9,12]. In this work, we investigate the association between population size and mutation bias, studying its consequences through a model that takes into account all of the relevant features of protein evolution discussed above: folding stability with respect to both unfolding and misfolding, population size, mutation bias, and neutrality, i.e. the relationship between folding stability and fitness.

Results

Model

We adopt the Moran model [27], which describes an evolving haploid population with N individuals that reproduce asexually and stochastically under mutation and selection. The model can be easily extended to diploid populations. We assume here that the product of population size times mutation rate is small, $N\mu \ll 1$, so that the population is monomorphic, i.e. the time scale for appearance of a new mutant in the population is large and at most one single mutant genotype is competing with the wild-type for fixation each time. This assumption is justified for small and intermediate populations when considering an individual protein coding gene, but not an entire genome (see Discussion). However, for large populations the assumption $N\mu \ll 1$ is violated even for an individual gene, and we can not apply the model to this case. In this monomorphic limit, the probability that a mutation arising as a single individual is fixed in the whole population can be exactly computed as [27]

$$P_{\mathrm{fix}}(i \to j) = \frac{1 - \frac{f_i}{f_j}}{1 - \left(\frac{f_i}{f_j}\right)^N}, \qquad (1)$$

where f_i is the exponential growth rate of the phenotype associated to sequence i, which will be called fitness in the following. This analytic result enormously simplifies the numeric study of the system allowing the systematic exploration of its parameter space. In our simulations, we randomly generate a mutated sequence, evaluate its fitness with respect to the wild type, and accept the new mutation according to the above probability.

We model mutations at the DNA level through the HKY process [28], whose only parameters are the equilibrium frequencies of the four bases A,T,G,C in the absence of selection, and the transition/transversion ratio k, whose influence is very weak and which we set to $k = 2$ [8]. In order to reduce the number of parameters, we assume that Chargaff's second parity rule holds, so that $\pi(\mathrm{A}) = \pi(\mathrm{T})$ and $\pi(\mathrm{G}) = \pi(\mathrm{C})$. Thus, the mutation model only depends on the GC usage, $\mathrm{GC} = \pi(\mathrm{G}) + \pi(\mathrm{C})$. GC usage different from 0.5 determines a mutation bias towards AT or

towards GC, therefore we sometimes refer to the GC usage variable as the mutation bias. In our model, the GC usage variable very strongly correlates with the GC content of the evolving gene in the stationary state of the evolutionary dynamics. The same correlation is thought to exist between the GC content of bacterial genomes, in particular at third codon position, and the GC usage of the mutations arising in bacterial replication. Therefore, we will compare the variable GC usage in our model with the variable GC content at third codon position in bacterial genomes.

Folding stability. In our model the fitness of an individual carrying a particular gene depends on the folding properties of the translated protein, which are estimated through a simple protein folding model. This model was used in our previous works [25,29,30] and it is similar to those used by others [31–39]. A characteristic of our model that distinguishes it from similar ones is that we consider two types of stability, with respect to misfolding and with respect to unfolding. Stability with respect to unfolding is estimated through the folding free energy F of a protein sequence $\mathbf{A}$, calculated with a simple contact interaction model (see Methods). Free energies estimated in this way correlate well with experimental measures (correlation coefficient $r=0.92$ over a test set of 20 proteins, UB, unpublished result). Stability with respect to misfolding is estimated through the normalized energy gap α (see Methods), which is the normalized difference between the effective energy of the native state and the minimum effective energy predicted through a Random Energy Model, representing the energy of compact intermediate structures very different from the native one. These misfolded structures can trap the folding process, and they can expose hydrophobic patches and promote aggregation.

Interestingly, these two kinds of stability respond in an opposite way to an increased mutation pressure towards hydrophobicity: while $-F$ increases for increasing mean hydrophobicity, meaning that proteins become more stable with respect to unfolding, the normalized energy gap decreases. This is due to the fact that the maximum stability of all potential misfolded structures increases more than the stability of the native structure, thus making misfolding and aggregation problems potentially more serious [12]. This trade-off between the two stabilities has a deep influence on the evolutionary dynamics.

Fitness. We adopt a fitness function that depends on the normalized stabilities $x_\alpha(\mathbf{A})=\alpha(\mathbf{A})/\alpha_{\rm thr}$ and $x_F(\mathbf{A})=F(\mathbf{A})/F_{\rm thr}$ and on the neutrality exponent S,

$$f(x_\alpha,x_F,S)=\begin{cases}\dfrac{1}{1+x_\alpha^{-S}+x_F^{-S}} & \alpha(\mathbf{A})>0\wedge F(\mathbf{A})<0,\\ 0 & \text{otherwise.}\end{cases} \quad (2)$$

The neutral thresholds $\alpha_{\rm thr}>0$ and $F_{\rm thr}<0$ define the scale of acceptable stabilities and they are kept fixed throughout the simulation. With this definition the fitness takes values between 0 and 1, vanishing if the protein does not fold correctly, which means that it is considered essential. Two plots of fitness versus stability for $S=1$ and $S=20$ are represented in Fig. 1 for illustration purposes. The fitness becomes a binary variable, either 0 or $f_{\rm max}$, if the neutrality exponent S is either zero (in this case all sequences satisfying $\alpha>0$ and $F<0$ are equally fit) or infinite (in this case all sequences overcoming the neutral thresholds $\alpha>\alpha_{\rm thr}$ and $F<F_{\rm thr}$ have fitness 1 and all other sequences are not viable). These limits are equivalent to Kimura's neutral model [2], which we studied previously [25,29,30], in which it is assumed that mutations that maintain stabilities above the neutral thresholds have no fitness effect, while all the others are lethal. This motivated us to name the parameter S the neutrality exponent. Notice that the term neutrality is sometimes defined as the fraction of proteins that retain wild-type structure under mutations [40]. This definition assumes a neutral model where the wild-type structure is either stable ($f=1$) or unstable ($f=0$). We prefer to call this quantity the fraction of neutral neighbors [29], and to call neutrality exponent the exponent S that determines the smoothness of the relationship between stability and fitness.

We choose the two neutral thresholds proportional to the values of α and F for the reference protein in the Protein Data Bank (PDB), multiplied with coefficients B_α and B_F. In simulations of neutral evolution, B_α and B_F have to be smaller than one so that the reference protein is viable. We present results with $B_\alpha=B_F=0.95$. We tested the robustness of our results with respect to both changes in the analytical form of the fitness function and the values of parameters, as discussed in the following.

Analytic results

We can analytically predict how the population size N and the neutrality exponent S influence stability and fitness by exploiting the formal analogy between population genetics and statistical mechanics demonstrated by Berg and coworkers [16] and by Sella and Hirsh [17]. These authors noticed that, in the monomorphic limit $N\mu\ll 1$ mentioned above and that we assume throughout this work, the Moran process, as well as other evolutionary processes studied in population genetics, tends to a stationary distribution of the form $\exp(N\log f)$. This distribution is equivalent to a Boltzmann distribution where population size N plays the role of inverse temperature and the logarithm of fitness, $\varphi(x_\alpha,x_F,S)=\log(f)$ plays the role of minus energy. This result implies that the probability to find a protein with stability values α and F in the stationary state of an evolving population is proportional to $\exp(N\log f(x_\alpha,x_F,S))$ multiplied by a factor that depends on the mutation process. The bias arising in the mutation process was treated as a "chemical potentia" by Sella and Hirsh [17] or as a mutational entropy by Berg et al. [16]. These two formalisms are qualitatively equivalent. We find the name mutational entropy more intuitive, and we will use it in the following. We define $P_{\rm mut}(\alpha,F)$ the probability to find stability parameters α and F under mutation alone, and we introduce the quantity $\sigma(x_\alpha,x_F,\mathrm{GC})=\log(P_{\rm mut}(x_\alpha,x_F,\mathrm{GC}))$, which we call the mutational entropy compatible with stabilities x_α and x_F under the given mutation process (notice that strictly speaking σ is not an entropy, however we find this name intuitive for indicating the mutational force that opposes protein stability). As discussed above, the mutational entropy depends on the GC usage, which can favor one kind of stability with respect to the other. Taking all this into account, the stationary distribution of stability that results from mutation and selection is

$$P(\alpha,F)\propto P_{\rm mut}(\alpha,F)\exp(N\log f(\alpha,F)). \quad (3)$$

The logarithm of the above probability can be interpreted as minus an evolutionary free energy divided by temperature $1/N$, and it is given by

$$G(x_\alpha,x_F)\propto\sigma(x_\alpha,x_F,\mathrm{GC})+N\varphi(x_\alpha,x_F,S), \quad (4)$$

where $\varphi(x_\alpha,x_F,S)=\log(f)$ is called the additive fitness [17]. The distribution Eq. (3) is peaked around the values $\overline{x}_\alpha$ and $\overline{x}_F$ that maximize the exponent G, i.e. minimize the evolutionary free energy. The equations that define these most likely values read

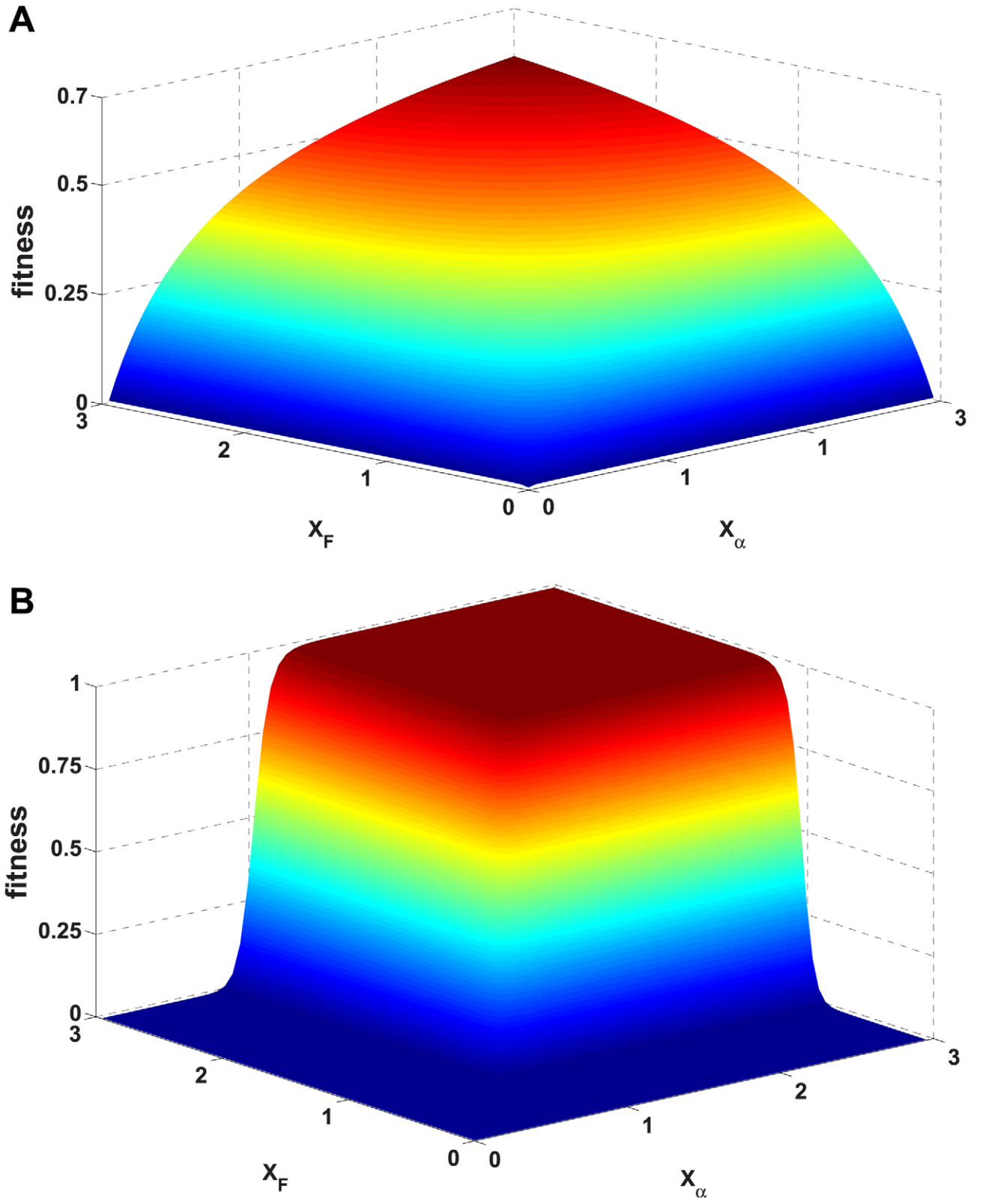

Figure 1. Fitness versus stabilities for $S=1$ (top) and $S=20$ (bottom).
doi:10.1371/journal.pcbi.1000767.g001

$$\left[\frac{\partial\sigma}{\partial x_i}\right]_{x_i=\overline{x}_i} = -N\left[\frac{\partial\varphi}{\partial x_i}\right]_{x_i=\overline{x}_i} = -NS\frac{\overline{x}_i^{-S-1}}{1+\overline{x}_\alpha^{-S}+\overline{x}_F^{-S}}, \quad (5)$$

where $i=\alpha,F$. We call the above the maximum-likelihood (ML) equations. Notice that the maximum likelihood values $\overline{x}_\alpha$ and $\overline{x}_F$ depend on the parameters N, S and GC. We can study this dependence analytically, assuming that Eq. (3) is narrowly peaked around these values, so that averages can be calculated as $\langle x_i\rangle=\overline{x}_i$ and $\langle\varphi(x_\alpha,x_F)\rangle=\varphi(\overline{x}_\alpha,\overline{x}_F)$. This approximation is justified by the fact that the mutational entropy σ is expected to be proportional to protein length L, which is of the order of 10^2, and the selective term is proportional to population size, which is also large, so that the exponent G is large and the distribution very narrow. The condition that G has a maximum at $\overline{x}_\alpha,\overline{x}_F$ requires that its Hessian matrix H, consisting of its second derivatives, is negative definite,

$$H_{ij}=\frac{\partial^2\sigma}{\partial x_i\partial x_j}+N\frac{\partial^2\varphi}{\partial x_i\partial x_j}. \quad (6)$$

This Hessian is the sum of the Hessian of $\varphi(x_\alpha,x_F)$, which is negative by construction, as it is easy to verify, and the Hessian of $\sigma(x_\alpha,x_F)$, which is the logarithm of a probability. We assume that the mutational entropy $\sigma(x_\alpha,x_F)$ has a single maximum at stabilities $(x_\alpha^{\rm mut},x_F^{\rm mut})$, so that its Hessian is negative. The values $(x_\alpha^{\rm mut},x_F^{\rm mut})$ that represent the most likely values of x_α and x_F in the absence of selection depend on GC. By definition of α, $x_\alpha^{\rm mut}$ is always negative, which is not a viable stability ($f=0$). However, our numerical results show that $x_F^{\rm mut}$ is positive for small GC usage, corresponding to hydrophobic sequences. The mutational entropy σ decreases for $x_F>x_F^{\rm mut}$ and for $x_\alpha>x_\alpha^{\rm mut}$, which implies that the corresponding derivatives are negative, as required for the existence of the solution of the ML equations.

We can go beyond the maximum-likelihood approximation writing the exponent G at second order as $G(x_\alpha,x_F)\approx G(\overline{x}_\alpha,\overline{x}_F)+\frac{1}{2}\sum_{ij}H_{ij}(x_i-\overline{x}_i)(x_j-\overline{x}_j)$, which is equivalent to approximating the distribution Eq. (3) as a Gaussian with covariance matrix $-H^{-1}$. Therefore, negativity of the Hessian matrix is equivalent to requiring the covariance matrix to be positive.

Influence of population size. We can calculate how $\overline{x}_\alpha$ and $\overline{x}_F$ depend on population size by taking the derivatives of the ML equations with respect to N (see Text S1). In this way, we find that both stabilities must increase with population size, as expected. The mean fitness $f(\overline{x}_\alpha,\overline{x}_F,S)$ is therefore an increasing function of N, whereas the mutational entropy $\sigma(\overline{x}_\alpha,\overline{x}_F,S)$ is a decreasing function of N.

Influence of the neutrality exponent. Stabilities are not monotonic functions of the neutrality exponent S. At $S=0$ all stabilities above the lethal threshold $x_i=0$ at which fitness drops to zero are selectively equivalent, and the ML equations imply that the stabilities with the largest mutational entropy fulfilling these conditions will prevail. As mentioned above, the most likely value of x_α in the absence of selection is negative for all GC usages, so that $\overline{x}_\alpha\approx 0$ for $S\to 0$. On the other hand, the most likely value of x_F in the absence of selection $x_F^{\rm mut}$ is positive for hydrophobic sequences, corresponding to small GC usage. The ML equations thus predict that $\overline{x}_F=\max(x_F^*,0)$, where $x_F^*\approx x_F^{\rm mut}$ satisfies the equation $\partial\sigma/\partial x_F=0$ at $x_\alpha=0$. Similarly, in the neutral limit $S\to\infty$, the smaller between x_α and x_F tends to the value 1, i.e.the corresponding stability tends to the neutral threshold, and the larger stability satisfies the equation $\partial\sigma/\partial x_i=0$ at $x_j=1$. For finite S, it can be shown that both stabilities increase with S when S is small, they reach a maximum and then decrease towards the neutral values (see Text S1). This behavior of stability arises from the fact that, under neutral or almost neutral evolution, the advantage in fitness provided by a more stable protein is too small to be fixed in the population against the entropic effect of mutations. This mechanism has been proposed as an explanation of the empirical observation that natural proteins are only marginally stable [3].

Similarly, we can show that the fitness has a minimum as a function of S: It starts from the value $f=1/3$ at $S=0$, then at small S the fitness is reduced because low stability values are penalized, at larger S more stable sequences are attained, and finally in the neutral limit the fitness tends to the maximum possible value $f=1$ while stability decreases (see Text S1). We can therefore distinguish three qualitative behaviors, described in Table 1. We are mainly interested in the parameter range that is far both from the region $SN<1$ at which the minimum stability is close to the lethal threshold $\min(\overline{x}_\alpha,\overline{x}_F)\approx 0$, and from the region of large S at which stabilities are close to the neutral thresholds.

Influence of the mutation bias. The most interesting feature of the evolutionary model presented here is the dependence of stability and fitness on the mutation bias. Unfortunately, this dependence cannot be predicted analytically, since we do not have a detailed model of how the mutation entropy σ depends on GC usage. Numerical results show that, for the folding free energy function that we adopt here, the two stabilities respond differently to the GC usage. This is expected, since small GC usage favors hydrophobic proteins, enhancing unfolding stability (x_F) at the expenses of misfolding stability (x_α). Since fitness depends on both x_α and x_F, it has to trade-off between the two stabilities, and we expect that there is an optimal GC usage at which the fitness is maximal for given S and N, which satisfies the equation $d\varphi/dGC=0$

$$\frac{d\varphi}{dGC}=\frac{S}{1+\overline{x}_\alpha^{-S}+\overline{x}_F^{-S}}\left(\frac{\partial\overline{x}_\alpha}{\partial GC}\overline{x}_\alpha^{-S-1}+\frac{\partial\overline{x}_F}{\partial GC}\overline{x}_F^{-S-1}\right), \quad (7)$$

where $\overline{x}_\alpha$ and $\overline{x}_F$ are determined by the ML equations (5). The maximum fitness is achieved when the quantity

$$\delta=\overline{x}_\alpha^{-S}+\overline{x}_F^{-S}=x_{\min}^{-S}\left[1+(x_{\min}/x_{\max})^S\right] \quad (8)$$

is minimal. Here $x_{\min}$ is the smaller value and $x_{\max}$ the larger value of $\overline{x}_\alpha$ and $\overline{x}_F$. We first discuss the small N regime at which

Table 1. Qualitative behavior of fitness and stability versus neutrality exponent S at fixed GC and population size.

S range	Stability	Fitness
Small	Increasing	Decreasing
Intermediate	Increasing	Increasing
Large	Decreasing	Increasing

At $S=0$ stability is close to the lethal threshold $\min(\overline{x}_\alpha,\overline{x}_F)\approx 0$ without any penalization for the fitness. In the small S regime stability increases with S, but the penalization for low stability decreases even more, with the net effect of a decrease in fitness. At intermediate S both stability and fitness increase with S and stability reaches a maximum that depends on N. Finally, at large S stability decreases with S, since the differences in fitness produced by a given difference in stability become smaller and cannot be fixed against the entropic effect of mutations, while fitness tends to the maximum possible value $f=1$.
doi:10.1371/journal.pcbi.1000767.t001

stabilities are small and they are strongly influenced by the GC usage. In this regime, we expect that there is a value of GC at which $\overline{x}_\alpha$ and $\overline{x}_F$ are equal. Therefore, at small GC usage it holds $x_{\min}=\overline{x}_\alpha$, which increases with GC, whereas at large GC usage it holds $x_{\min}=\overline{x}_F$, which decreases with GC. Consequently, the factor $x_{\min}^{-S}$ has a minimum where $\overline{x}_\alpha=\overline{x}_F$. Conversely, the second factor that appears in δ, $1+(x_{\min}/x_{\max})^S$, has a maximum where $\overline{x}_\alpha=\overline{x}_F$. We expect that the factor $x_{\min}^{-S}$ depends more strongly on GC than the factor $1+(x_{\min}/x_{\max})^S$, in particular if S is large. Therefore, we expect that the minimum δ (i.e. the optimal GC) is reached near the GC usage at which $\overline{x}_\alpha=\overline{x}_F$, and that it approaches this value as S grows. The GC usage at which $\overline{x}_\alpha=\overline{x}_F$ has an interesting interpretation. We can define the selective pressure on the variable x_i as the derivative of φ with respect to x_i, which expresses how fitness responds to a change in stability. If this derivative is large, a large number of attempted mutations will be discarded because of their negative influence on fitness. The ML equations show that the selective pressure is proportional to $-\partial\sigma/\partial x_i$, and it is stronger on the smaller variable $x_{\min}$. Therefore, when the GC usage increases, the selective pressure on unfolding increases, and the selective pressure on misfolding decreases, and they balance when $\overline{x}_\alpha=\overline{x}_F$.

Theoretical considerations and numerical results indicate that there is a second regime at large N. In this limit, the fitness tends to the maximum possible value. Due to the trade-off between unfolding and misfolding stability, it is not possible to maximize $\overline{x}_\alpha$ and $\overline{x}_F$ simultaneously, since they are inversely related. As N increases, $\overline{x}_\alpha$ and $\overline{x}_F$ are expected to converge to the optimal fitness point x_α^{Max},x_F^{Max} and their dependence on GC is expected to become weaker and weaker. We find numerically that x_α^{Max} is smaller than x_F^{Max}, so that for large N, $\overline{x}_\alpha$ is smaller than $\overline{x}_F$ for all GC, and the selective pressure is always stronger on $\overline{x}_\alpha$. In this regime, $x_{\min}^{-S}$ always decreases with GC and its dependence on GC gets weaker. Conversely, the term $1+(x_{\min}/x_{\max})^S$ always increases with GC, and the optimal GC is determined by a balance between these two terms. We now discuss two interesting limiting behaviors of the optimal GC.

1. In the small N regime and for finite S, so that SN is small, $\overline{x}_\alpha$ tends to zero and $\overline{x}_F$ tends to $\max(x_F^{\mathrm{mut}},0)$ independent of S. For small GC usage, x_F^{mut} is positive and $\delta\approx\overline{x}_\alpha^{-S}$ is a decreasing function of GC, since $\overline{x}_\alpha$ increases with GC. For large GC usage, $\overline{x}_\alpha>\overline{x}_F$ and δ increases with GC. Therefore, we expect that the minimum of δ, i.e. the optimal GC, is attained near the GC usage at which $x_F^{\mathrm{mut}}(\mathrm{GC})=0$, which is independent of S and of the neutral thresholds B_α and B_F.
2. In the neutral limit $S\to\infty$, the selective pressure only affects the smallest stability variable, since $\varphi\approx-\log(1+x_{\min}^{-S})$. This tends to $x_{\min}\approx1$ independent of N and GC. Therefore, as discussed above, for large S, the optimal GC is reached when $\overline{x}_\alpha\approx\overline{x}_F\approx1$, i.e. when the two selective pressures balance. The ML equations imply that at this point $\partial\sigma/\partial x_F(1,1,\mathrm{GC})\approx\partial\sigma/\partial x_\alpha(1,1,\mathrm{GC})$, so that the optimal GC does not depend on N. The ML equations also imply that, in the large S limit, $1+\overline{x}_{\min}^S\approx NS/|\partial\sigma/\partial x|_{x=1}$ (see Text S1), which means that the maximum stability and maximum fitness is attained at the GC value at which $|\partial\sigma/\partial x_{\min}|$ is minimum. This prediction is confirmed in Fig. 6 in the Text S1).

Simulations

All simulations presented here are based on the native structure of some natural protein. When not otherwise stated, we exemplify our numerical results using the protein lysozyme, PDB id. `31zt`. In all cases, the starting sequence is the sequence in the PDB. Results are collected after fitness has converged to its stationary value, discarding the first ≈1000 accepted substitutions, which are enough for equilibration, as it can be seen in Fig. 2 in the Text S1.

As an illustration of the stationary states of the evolutionary dynamics, we represent in Fig. 2 the mean stability values $\overline{x}_F$ and $\overline{x}_\alpha$ obtained using the fitness function with $S=1$ for different population sizes from $N=10$ to $N=4000$ and GC usage from $\mathrm{GC}=0.1$ to $\mathrm{GC}=0.9$. The distributions $P(\alpha,F)$, Eq. (3), are narrowly peaked around the plotted points $(\overline{x}_\alpha,\overline{x}_F)$. Sets of points with the same GC usage are joined with solid lines, and sets of points with the same N are joined with dashed line. The data are superimposed to a heat map that shows the value of fitness in colour code. We can see from the figure that both stabilities grow with N. On the other hand, $\overline{x}_\alpha$ grows and $\overline{x}_F$ decreases with GC, so that $\overline{x}_\alpha$ and $\overline{x}_F$ are negatively correlated for fixed population size. For $\mathrm{GC}<0.6$, $\overline{x}_F$ tends to a finite value when $\overline{x}_\alpha$ tends to zero (corresponding to very small N), i.e. the most likely value of x_F in the absence of selection is $x_F^{\mathrm{mut}}>0$ and, for such small GC usage, there is very weak selective pressure on unfolding. One can see from the plot that the GC usage at which $\overline{x}_F$ and $\overline{x}_\alpha$ are equal increases with population size, which implies that the selective pressure on $\overline{x}_\alpha$ increases more than the selective pressure on x_F for increasing population size. In the large population limit both $\overline{x}_\alpha$ and $\overline{x}_F$ tend to finite values independent of GC. We estimated from our numerical results that $x_\alpha^{\max}\approx4$ and $x_F^{\max}\approx10$, so that for large populations it is always $x_\alpha^{\max}<x_F^{\max}$.

Fitness clearly increases with N. The variation of fitness with GC is weaker, but one can nevertheless notice it from the plot. This variation translates into the fact that, for fixed fitness function and population size N, there is an optimal GC usage such that fitness is maximal, as predicted in Eq. (7). The existence of this optimal mutation bias is demonstrated in Fig. 3, where we plot the fitness of populations with constant N and S as a function of their GC usage. For each set of parameters, we obtained the optimal GC usage $\mathrm{GC}_{\mathrm{opt}}(N,S)$ by cubic interpolation, as exemplified in Fig. 3, and plotted it versus N. We found that $\mathrm{GC}_{\mathrm{opt}}$ is small for very small populations, large for intermediate populations, and the bias is almost absent ($\mathrm{GC}\approx0.5$) for very large populations (see Fig. 4). We obtained qualitatively similar results as long as the neutrality exponent S is not too large or too small (in that case, the fitness landscape becomes almost neutral). The population size at which the optimal GC usage is highest increases with decreasing S for small S, while the opposite holds for large S. Our numerical results are consistent with the optimal GC usage becoming less dependent on S in the infinite population limit, see Fig. 3 in the Text S1.

Eq. (4) implies that a trait that confers a selective advantage can only be fixed against the entropic effect of random mutations when the difference in the selection coefficients φ is larger than $1/N$. We therefore verified whether the difference of selective coefficients φ between populations adopting different GC usages is large enough so that the optimal one would be eventually selected. We found that $\Delta\varphi$ decreases with population size, but more slowly than $1/N$, so that $N\Delta\varphi$ increases with N, see Fig. 4 in the Text S1. This implies that two populations evolving with different mutation bias (the optimal one and another one) attain a fitness difference large enough so that the optimal GC usage can be selected.

We tested that our results do not change qualitatively when different protein structures are used in the simulation. To this end, we computed the relationship between the optimal GC usage and population size at neutrality exponent $S=1$ for five proteins of different length and secondary structure (see Methods). All curves, plotted in Fig. 5, have the same shape, although they are shifted in

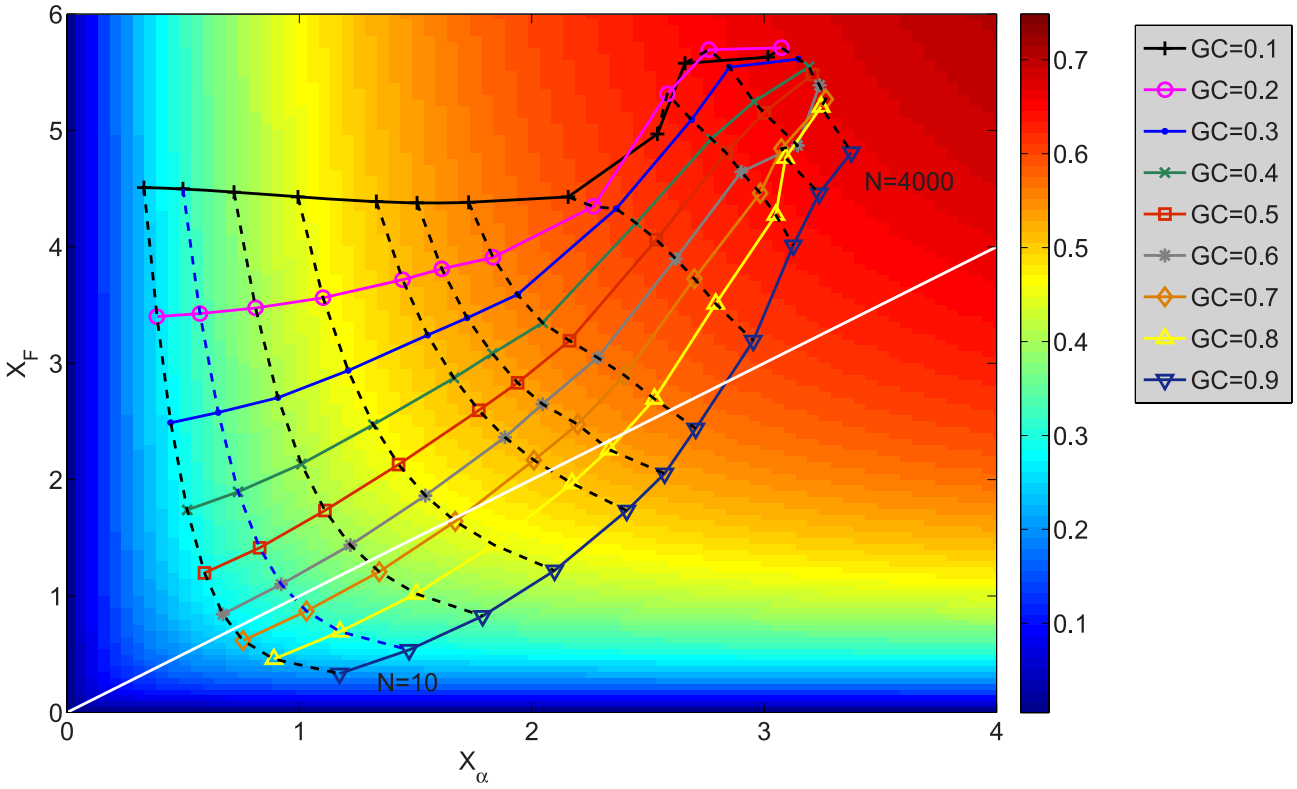

Figure 2. Mean unfolding stability $\bar{x}_F$ versus misfolding stability $\bar{x}_\alpha$ for neutrality exponent $S=1$ (non-neutral regime). The sets of points joined with solid lines correspond to constant GC usage, between 0.1 (largest $\bar{x}_F$) and 0.9 (largest $\bar{x}_\alpha$). $\bar{x}_\alpha$ grows and $\bar{x}_F$ decreases with GC. The sets of points joined with dashed lines correspond to constant population size N, from $N=10$ (smallest stability) to $N=4000$ (largest stability). Both stability variables $\bar{x}_i$ increase with N. Data points are superimposed to a heat map of the fitness function, showing that fitness increases with N. However, constant N lines do not correspond to constant fitness, but there are small variations, from which the optimal GC usage is derived. The solid white line shows $\bar{x}_\alpha=\bar{x}_F$ at which the selective pressures on $\bar{x}_\alpha$ and $\bar{x}_F$ balance. One can see that, at large N, $\bar{x}_\alpha$ is smaller than $\bar{x}_F$ for all GC, so that the selective pressure is stronger on the former.
doi:10.1371/journal.pcbi.1000767.g002

the vertical direction in a way that suggests that shorter proteins are characterized by larger optimal GC usage (but more proteins are needed to confirm this trend). We then combined the five curves. We assumed that a genome composed of these five proteins is evolving with very low mutation rate, so that at most one protein is mutated at each step, consistent with the assumption $\mu N \ll 1$. The global fitness of the organism was obtained through two different ansatz that yielded qualitatively similar results, either as the minimum of the fitness of all proteins P, $f=\min_P(f_P)$ or as the product of the fitnesses, $f=\Pi_P f_P$, assuming absence of epistatic interactions. From these f we then obtained the optimal GC by cubic interpolation. This is represented in Fig. 5, bottom plot for $S=1$. One can see that the qualitative behavior of the individual curves is preserved. We expect therefore that this qualitative behavior would be maintained for a large number of proteins as well.

To further test the robustness of our results we changed the neutral thresholds $\alpha_{\rm thr}$ and $F_{\rm thr}$ up to 20%, examining nine combinations of thresholds for neutrality exponent $S=1$. The results are shown in Fig 6. One can see that the qualitative behavior is unchanged. As expected, when $\alpha_{\rm thr}$ becomes more tolerant the optimal GC usage decreases, and the contrary happens when $\alpha_{\rm thr}$ becomes more strict.

Finally, we verified that the results are robust with respect to the energy parameters used. For such a test, we adopted the contact interaction energies determined by Godzik, Kolinsky and Skolnick (GKS) [41]. These parameters have correlation $r=0.65$ with the BVK parameters adopted in the present study, so that their differences are not small. We determined a new parameter for conformation entropy $s=s_{\rm GKS}=0.20$ by demanding the folding free energies computed with the two sets of energy parameters to coincide on the average. As one can see from the dotted curve in Fig. 7, the qualitative behavior is the same for the two parameter-sets, but the optimal GC usage for GKS parameters is lower than for BVK parameters. This is due to the fact that, for our test protein lysozyme, GKS energy parameters produce a very low normalized energy gap $\alpha=0.024$ instead of $\alpha=0.24$ with BVK parameters, which means that the native conformation is closer in energy to random conformations when GKS parameters are used. Consequently, $\alpha_{\rm thr}$ is very small (we recall that $\alpha_{\rm thr}$ is proportional to the value of α for the native sequence) and the selective pressure on misfolding is very weak. We then increased this selective pressure by setting $B_\alpha=9.5$ instead of 0.95. The resulting curve can be seen in Fig. 7 as a dashed curve. One finds that the maximum GC usage is now much larger, reaching $\mathrm{GC}\approx 0.8$.

Finally, we show in Fig. 8 the optimal GC usage versus the neutrality exponent S for small ($N=40$), intermediate ($N=320$) and large ($N=1280$) populations. For small populations the optimal GC usage increases with the neutrality exponent, from very small values to $\mathrm{GC}\approx 0.5$. For intermediate and large populations the optimal GC usage has a maximum and then it decreases. The maximum value of $\mathrm{GC}_{\rm opt}$ increases with population size, and it is reached at smaller neutrality exponent for intermediate populations ($S=2$ at $N=320$) than for large populations ($S\approx 5$ at $N=1280$).

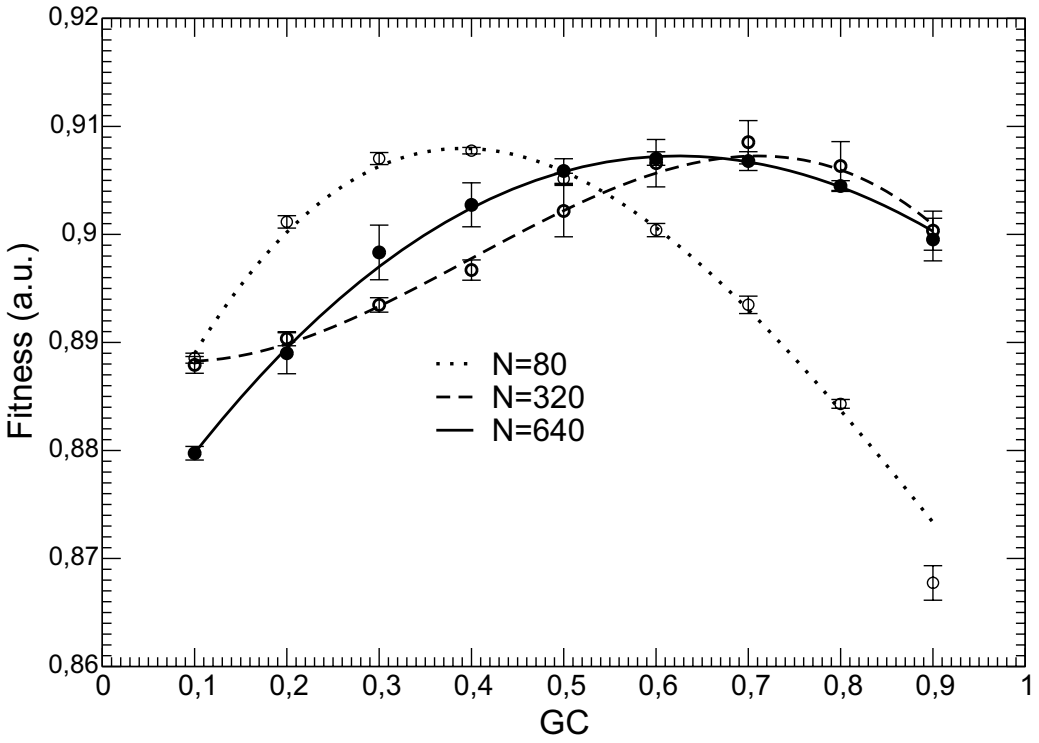

Figure 3. Fitness (in different units for each curve) versus GC usage for neutrality exponent $S=1$ and three different population sizes. The curves have been shifted in the vertical direction so that their maxima coincide. We obtain GC_{opt} by cubic fits, which are plotted as dotted, dashed, and solid lines.
doi:10.1371/journal.pcbi.1000767.g003

We then tested the mean-field prediction that the stability coefficient $x=\min(x_\alpha,x_F)$ has a maximum and the sequence entropy has a minimum as a function of neutrality exponent S. As expected, maximum stability and minimum entropy occur at the same value of S, see Fig. 5 in the Text S1.

Qualitative behavior of the optimal GC. We now discuss the N-dependence of the optimal GC based on the results reported in Fig. 2. As explained above, the existence of the optimal GC usage arises from the trade-off between unfolding stability and misfolding stability in response to changes in the mutation bias. One can observe this trade-off in Fig. 2, from which it appears that $\overline{x}_\alpha$ and $\overline{x}_F$ are negatively correlated for fixed population size. At the optimal GC the derivatives of ${\overline{x}_\alpha}^{-S}$ and ${\overline{x}_F}^{-S}$ with respect to GC, which have opposite sign, become equal in absolute value, as indicated by Eq. (7). One can see from Fig. 2 that at small GC usage $\overline{x}_\alpha$ responds to GC variation more strongly than $\overline{x}_F$, whereas the opposite happens at large GC usage, so that the optimal is reached at intermediate GC. In Fig. 2, the white thick line represents the $\overline{x}_F=\overline{x}_\alpha$ line at which the selective pressures on unfolding and misfolding are equal. One can see from the plot that, for small GC usage and small population sizes, the selective pressure is stronger on x_α (misfolding). Since $\overline{x}_F$ increases faster than $\overline{x}_\alpha$ with population size, the selective pressure on x_α increases with N more than the selective pressure on x_F. Consequently, the GC usage at which $\overline{x}_F=\overline{x}_\alpha$ (white line) increases with population size. As discussed in the section "Influence of the mutation process", this behaviour qualitatively explains why the optimal GC increases with N at small N, since the optimal GC is expected to be near the value at which $\overline{x}_F=\overline{x}_\alpha$. Near $N=320$, the optimal GC attains a maximum as a function of N. For $N>320$, we see that $\overline{x}_\alpha>\overline{x}_F$ for all GC usages, so that the selective pressure is always stronger on misfolding, and we enter what we called the large N regime. In this regime, $\overline{x}_\alpha$ and $\overline{x}_F$ tend to the finite values that yield the maximum absolute fitness (numerical results suggest that they are $x_\alpha^{\max}\approx 4$ and $x_F^{\max}\approx 10$), which are independent of GC, so that the GC dependence of stabilities gets weaker and weaker for large populations. When these limiting values are approached, the $(\overline{x}_\alpha,\overline{x}_F)$ curves that correspond to fixed $N>320$ and varying GC in Fig. 2 change their shape, becoming more convex and centered around $GC=0.5$ (red squares). This behavior corresponds to the fact that the optimal GC decreases towards $GC=0.5$ for very large population size.

According to this reasoning, the maximum value of GC_{opt} versus N is reached at a population size where $x_{\min}=\overline{x}_\alpha$ approaches its limiting value x_α^{Max}. As discussed above and detailed in the Text S1, $x_{\min}$ has a maximum as a function of S for fixed population size. Therefore, the population size at which a given value $x_{\min}$ is reached has a minimum as a function of S, which implies that the population size N at which the optimal GC is largest has a minimum as a function of S. This prediction is in

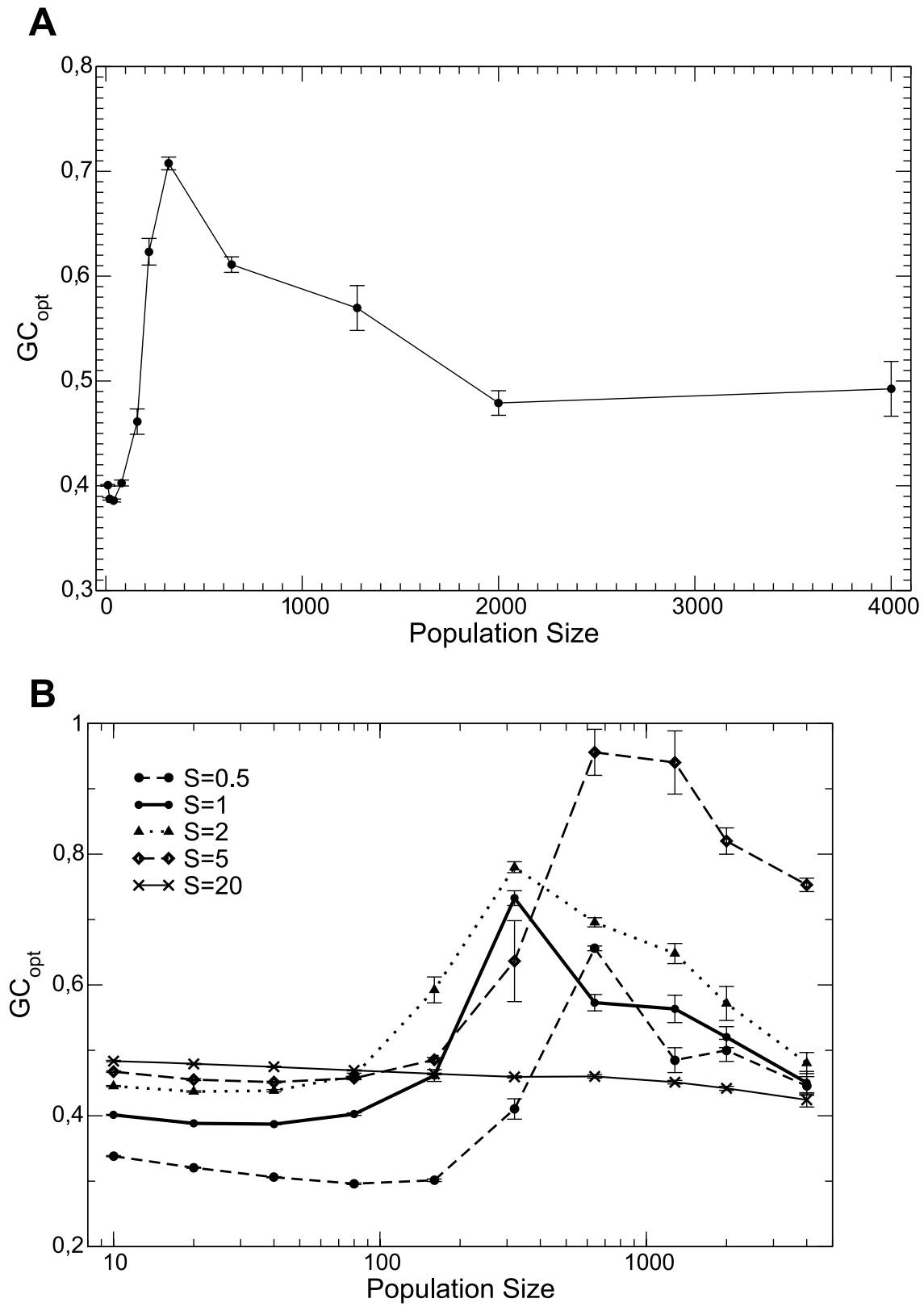

Figure 4. Optimal GC usage $\mathrm{GC_{opt}}$ at which the fitness is maximum versus population size N. The upper plot shows data with neutrality exponent $S=1$ and the bottom plot shows $S=0.5,1,2,5$ and 20. Interpolating lines are drawn as a guide to the eye.
doi:10.1371/journal.pcbi.1000767.g004

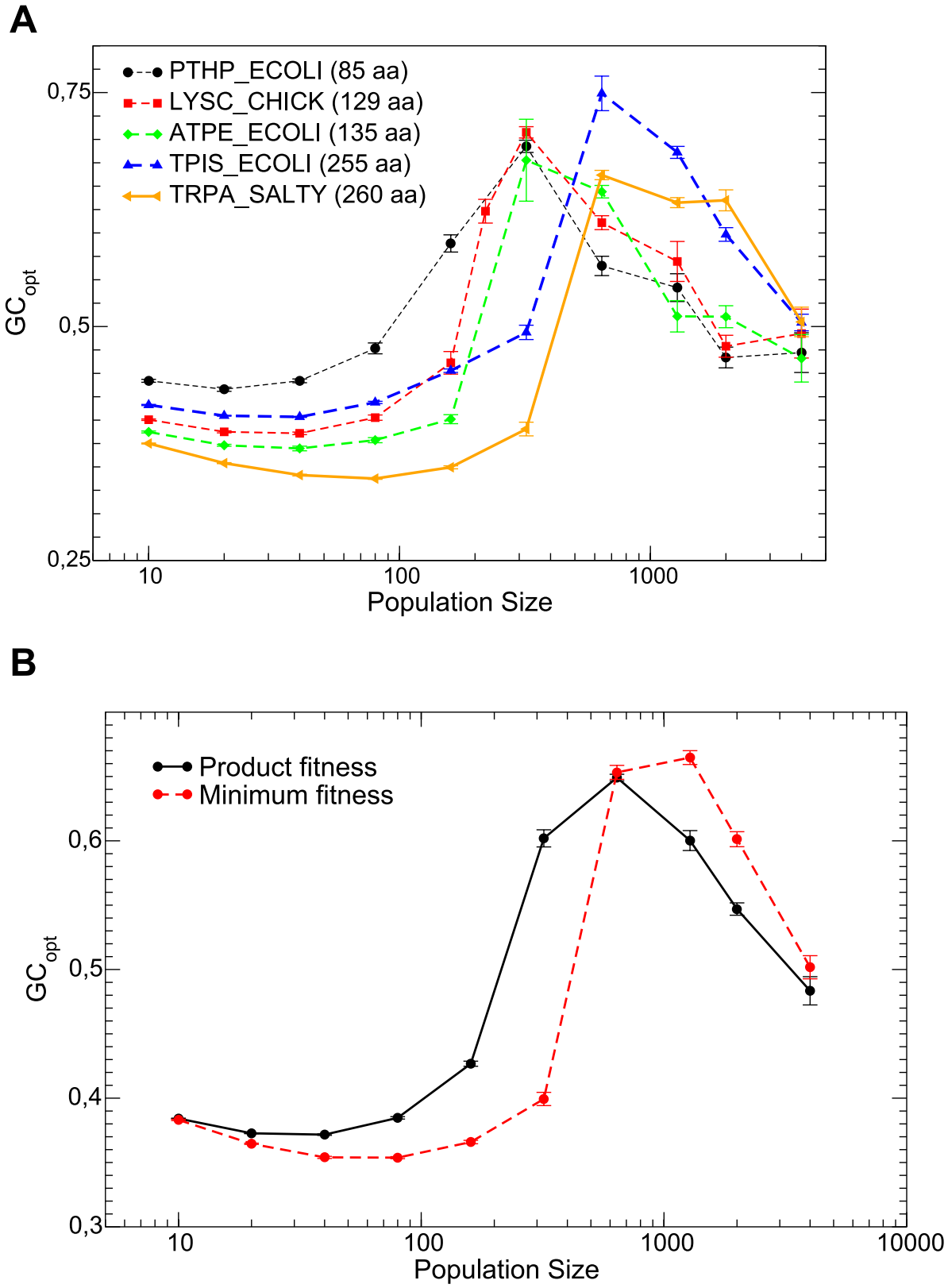

Figure 5. Optimal mutation bias GC_{opt} at which the fitness is maximum versus population size N for different proteins and neutrality exponent $S=1$. Upper plot: Results for individual proteins. Bottom plot: Fitness is obtained for the combination of 5 proteins either as the minimum or as the product over all proteins. Interpolating lines are drawn as a guide to the eye.
doi:10.1371/journal.pcbi.1000767.g005

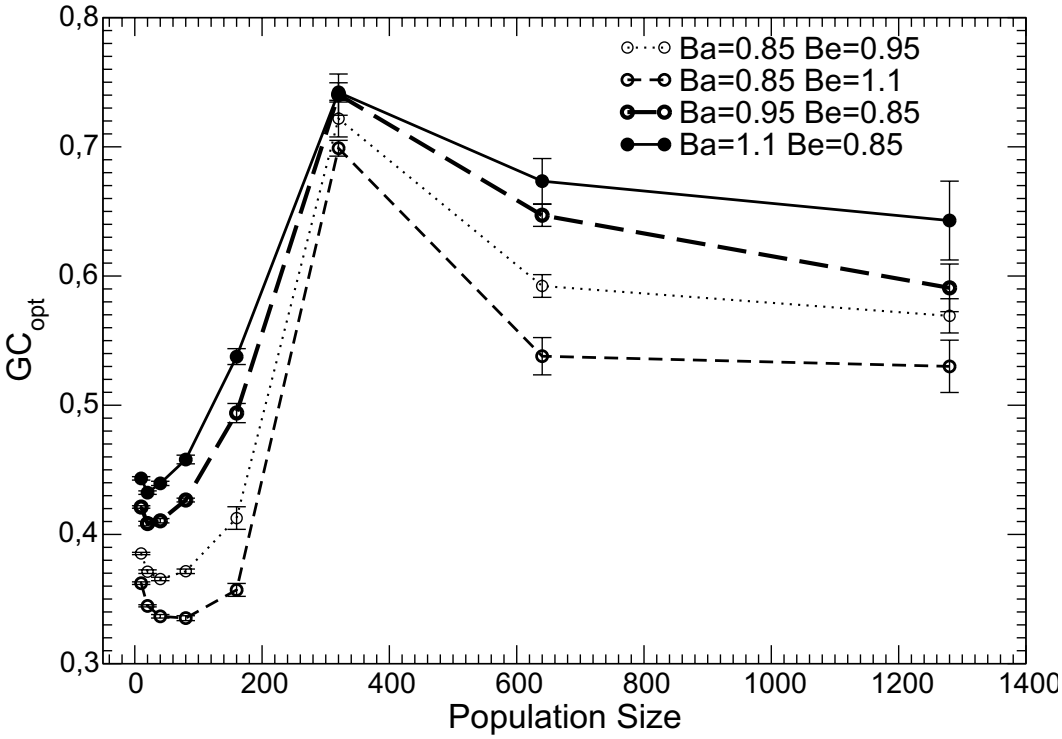

Figure 6. Optimal GC usage GC_{opt} versus population size N for neutrality exponent $S=1$ and different values of the neutral thresholds $\alpha_{thr}=B_\alpha\alpha_0$ and $F_{thr}=B_fF_0$, where the reference energy gap α_0 and unfolding free energy F_0 are those measured for the protein in the PDB. We simulated all nine combinations of the values $\{0.85,1.0,1.1\}$ for either B_α of B_F. We only show four combinations since all other curves are contained between them.
doi:10.1371/journal.pcbi.1000767.g006

qualitative agreement with Fig. 4, bottom plot, which suggests that the minimum of the largest GC_{opt} versus N, $\max_N GC_{opt}(N,S)$, is reached between $S=1$ and $S=2$.

Effective population Size

The results that we have presented suggest that mutation bias towards AT or GC favor protein folding stability for very small and intermediate population sizes, respectively, while very large populations are advantaged in the absence of bias ($GC\approx0.5$). As it will be discussed below, this suggests that species evolving with mutation bias, either towards AT or GC, will have smaller population size than species with no bias. This prediction is consistent with the fact that almost all bacterial species with intracellular lifestyles, implying a reduction of effective population size through bottlenecks, shifted their mutation spectrum to AT, which resulted in small genomic GC content. On the other hand, among bacteria with large GC content some are facultative pathogens, such as *Mycobacterium tuberculosis*, and some live symbiotically in plant nodules, but there is no general tendency allowing for the deduction of their population size from their lifestyles. Therefore, to test our prediction, we tried to directly estimate their effective population size.

The effective population size N_e depends on the breeding structure and the natural history of a population, and in particular it is influenced by the bottlenecks that the population may undergo if a few individuals periodically colonize new environments. Therefore, the effective population size cannot be measured experimentally, but is estimated by fitting some observed population feature to its expected value under evolution in a population with given N_e. Optimal codon usage was used several years ago to estimate the effective population size of *Escherichia coli* [42]. A recent work supports the existence of a correlation between effective population size and synonymous codon usage [43], and the availability of many complete genomes makes it possible to analyze codon usage on a large scale. Codon usage and mutation bias are intimately correlated. It is commonly believed that the mutation bias, rather than selection for optimal codon usage, ultimately influences the global GC content of a genome [18,19]. The definition of the optimal codon usage on which the results that we use here are based considers the excess frequency of preferred codons with respect to the frequency expected under mutation alone, and is therefore not expected to depend on the mutation bias in a trivial way. Dos Reis *el al.* [44] have recently estimated the optimal codon usage in a large number of prokaryotic species. We use their data rather than the analogous data obtained by Sharp *et al.* [45], since Dos Reis *et al.* evaluated the optimal codon usage on the entire genome, whereas Sharp *et al.* concentrated their attention only on ribosomal genes, which can be a biased sample. Fig. 9 shows the average optimal codon usage versus the average GC content at the third codon position, which is not affected by the selection on the amino acid sequence and is expected to be very strongly correlated with the mutation bias. We distinguished species with small (<0.3), intermediate (0.4 to 0.6) and large (>0.7) GC content. Species with intermediate GC content turned out to have significantly larger optimal codon usage, which suggests that they have larger effective population size. The scatter plot and the histogram of the GC content are shown in Fig. 7 and 8) in the Text S1. Error bars in the plot represent the standard error of the mean, and show that the mean values are significantly different. However, data prior to the mean are rather broadly distributed, with standard deviations equal to 0.16 ($GC<0.3$), 0.24 ($GC\in[0.4,0.6]$) and 0.20 ($GC>0.7$).

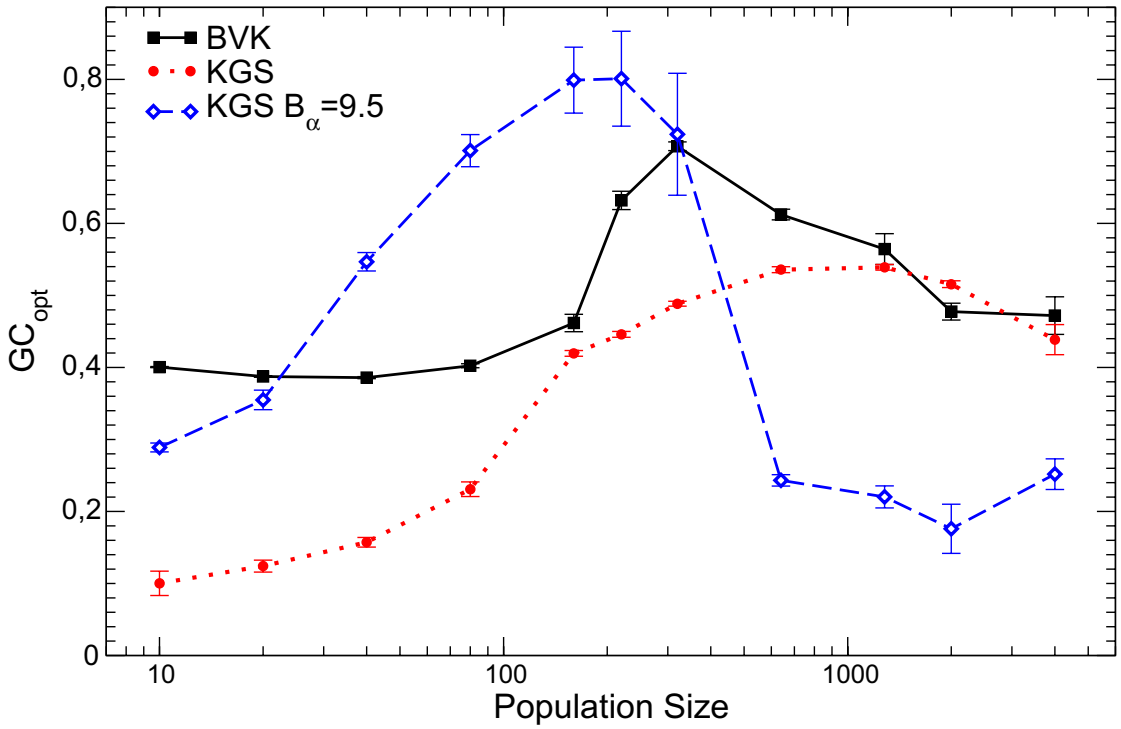

Figure 7. Comparison between the optimal GC usages computed with GKS energy parameters (dotted line and dashed line) and the BVK parameters adopted in the present study (solid line). The conformation entropy is $s_{BVK}=0.074$ for BVK parameters and $s_{GKS}=0.20$ for GKS. The coefficient of the neutral threshold is $B_\alpha=0.95$ for the dotted curve and $B_\alpha=9.5$ for the dashed curve. Other parameters are fixed at $S=1$, $B_F=0.95$.
doi:10.1371/journal.pcbi.1000767.g007

As a second estimate of effective population size, we considered the ratio between non-synonymous and synonymous substitutions K_A/K_S, which is thought to represent the strength of negative selection [8]. We examined values of K_A/K_S computed for pairs of entire genomes, recently published by Daubin and Moran [46]. From their table, we eliminated two pairs of genomes for which the evolutionary divergence, estimated through K_S, was very small (<0.1), corresponding to *Bordetella pertussis/parapertussis* and two strains of *Xylella fastidiosa*, since it is known that the amino acid substitution rate is significantly higher at small time separation [47–49] and in fact these two pairs of genomes showed the two largest values of K_A/K_S. We also eliminated two pairs for which the two compared species had genomic GC content in different bins: two strains of *Prochlorococcus marinus* having GC = 36% and 51%, and the pair *Synechocystis/Synechococcus* having GC = 48% and GC = 65%, respectively. We divided the remaining 19 pairs in 3 bins of low, mean and high GC content and averaged their K_A/K_S. Results, shown in Fig. 9, clearly show that species evolving with no bias are characterized by lower K_A/K_S, hence larger effective population size, in agreement with the analysis of the optimal codon usage and with the prediction of our model.

Finally, we reanalysed our data on protein folding stabilities computationally estimated for orthologous proteins in different prokaryotic genomes [12]. Unfolding and misfolding stabilities are negatively correlated, as predicted by our model (see Fig. 10). We found that most of the organisms evolving with mutation bias have proteins whose misfolding stability is lower than what could be expected based on their unfolding stability, see Fig. 11. This further supports the idea that these species are characterized by reduced effective population sizes.

Discussion

Interplay between mutation bias and population size

We studied here a mathematical model of protein evolution where the genotype to phenotype mapping is determined by the stability of the mutated protein against unfolding and misfolding, predicted using a protein folding model that correlates well with experimental measures. As observed in previous work, the two kinds of stability respond in an opposite way to changes in the GC usage of the mutation process. This fact produces a trade-off between the two kinds of stability, and an interesting phenomenology arises from the impossibility to find a mutation process that optimizes both stabilities at the same time, a concept that in the physical literature has received the name of *frustration*.

We considered three key evolutionary parameters: the effective population size N, the neutrality exponent S, which determines how protein stability influences fitness, and the GC usage that

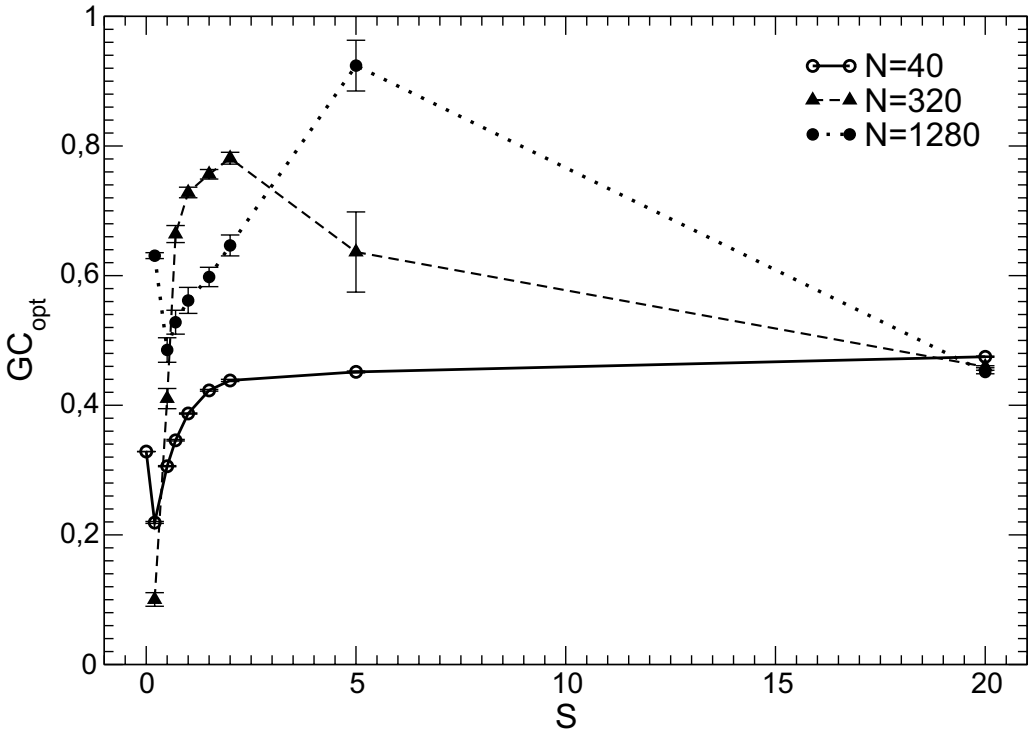

Figure 8. Optimal GC usage GC_{opt} versus neutrality exponent S for three population sizes N.
doi:10.1371/journal.pcbi.1000767.g008

expresses the mutation bias. Despite its importance in shaping the folding properties of proteins, the latter has been rarely considered in evolutionary models. Here we show that, in the non-neutral regime, mutation bias has a very interesting interplay with population size. We suggest that this can explain why some microbial species adopted extreme mutation bias.

At high neutrality exponent, all proteins with stability above the neutral threshold provide the same fitness and evolution is only able to attain the lowest allowed stabilities [3], almost independent of population size. Consistently, our analytic and numerical results indicate that the neutrality exponent S has a non-monotonic influence on protein stability, which reaches a maximum at intermediate S for given population size. The increase of S in our model has its biological counterpart in the increase of the expression level of chaperones, which make proteins more tolerant to stability losses. Therefore, the decrease of stability for increasing S predicted by our model would correspond in the real world to the decrease of protein stability when the chaperone expression is increased. This outcome appears rather plausible. However, given the cost of synthesizing chaperones, in real evolution it is to be expected that the increase of the expression level of chaperones is a consequence of the loss of protein stability, as observed in intracellular bacteria with reduced population size, rather than the other way round.

In the neutral regime the GC usage influences the amino acid composition and consequently the folding properties, favoring proteins more stable with respect to misfolding but less stable with respect to unfolding, without modifying the fitness. In contrast, in the non-neutral regime fitness is a continuous function of stability and the outcome of evolution depends non-trivially on mutation in the sense that for fixed population size there is an optimal mutation bias at which fitness and stability are maximal. This is an unexpected result, which implies that mutation and selection are effectively entangled, and that the mutation spectrum constrains the maximum stability and fitness that an evolving population can attain. The possibility that the mutation rate is optimized as a response to evolutionary forces [50] has received considerable attention in experiments (see Ref. [51] for a recent work) and modelling (see for instance Refs. [52,53]). The main forces influencing mutation rate evolution have been identified as the population size [50], the ruggedness of the fitness landscape [54] and the average negative effect of a mutation [55]. Recently, a theoretical work has established a relation between mutation rate, maximal genome size and thermodynamic response of proteins to point mutations, showing that populations go extinct via lethal mutagenesis when their mutation rate exceeds a few mutations per genome per replication [56]. Simulations of this model confirmed the predicted behaviour, showing that the limiting number of mutations is approximately seven for RNA viruses and about four for DNA-based organisms, with some weak dependence on the number of genes in the organism and the organism's natural death rate [57]. This model predicts that species with high mutation rates tend to have less stable proteins compared to species with low mutation rates. Therefore, the notion that the mutation process

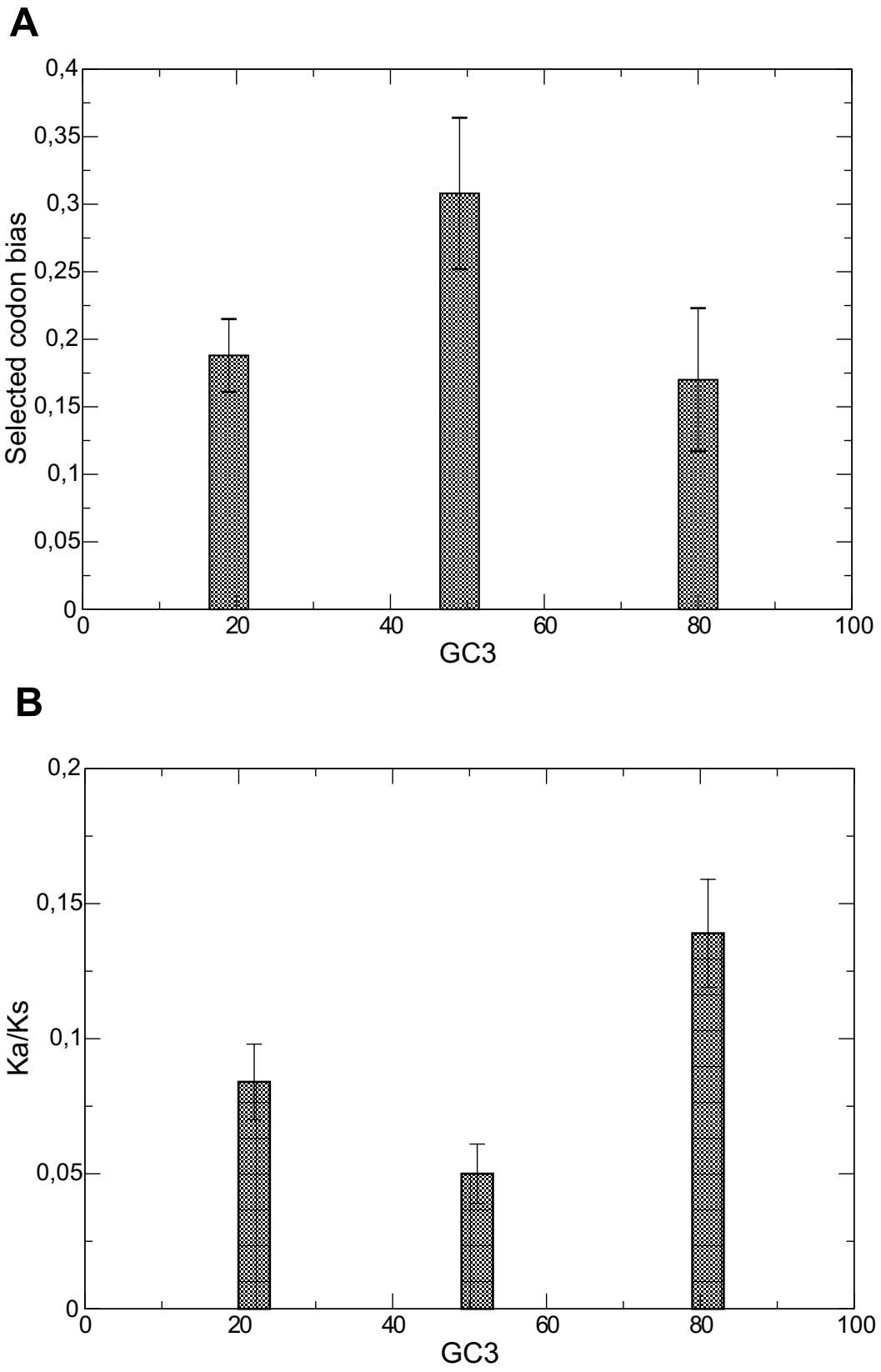

Figure 9. Estimates of quantities correlating with effective population size obtained from genomic data. Upper plot: Optimal codon bias estimated by dos Reis *et al.* [44] versus GC content at synonymous third codon position, shown as mean and standard error of the mean for three bins of GC3 (smaller than 30%, 40 to 60%, larger than 70%). Error bars in the plot represent the standard error of the mean, and show that the mean values are significantly different. However, data prior to the mean are rather broadly distributed, with standard deviations equal to 0.16 ($\mathrm{GC}<0.3$), 0.24 ($\mathrm{GC}\in[0.4,0.6]$) and 0.20 ($\mathrm{GC}>0.7$). Bottom plot: values of K_A/K_S computed by Daubin and Moran [46] are averaged for pairs of bacteria with low, intermediate and high GC content. Both plots support the notion that species with GC content $\approx 50\%$ are characterized by larger effective population size.
doi:10.1371/journal.pcbi.1000767.g009

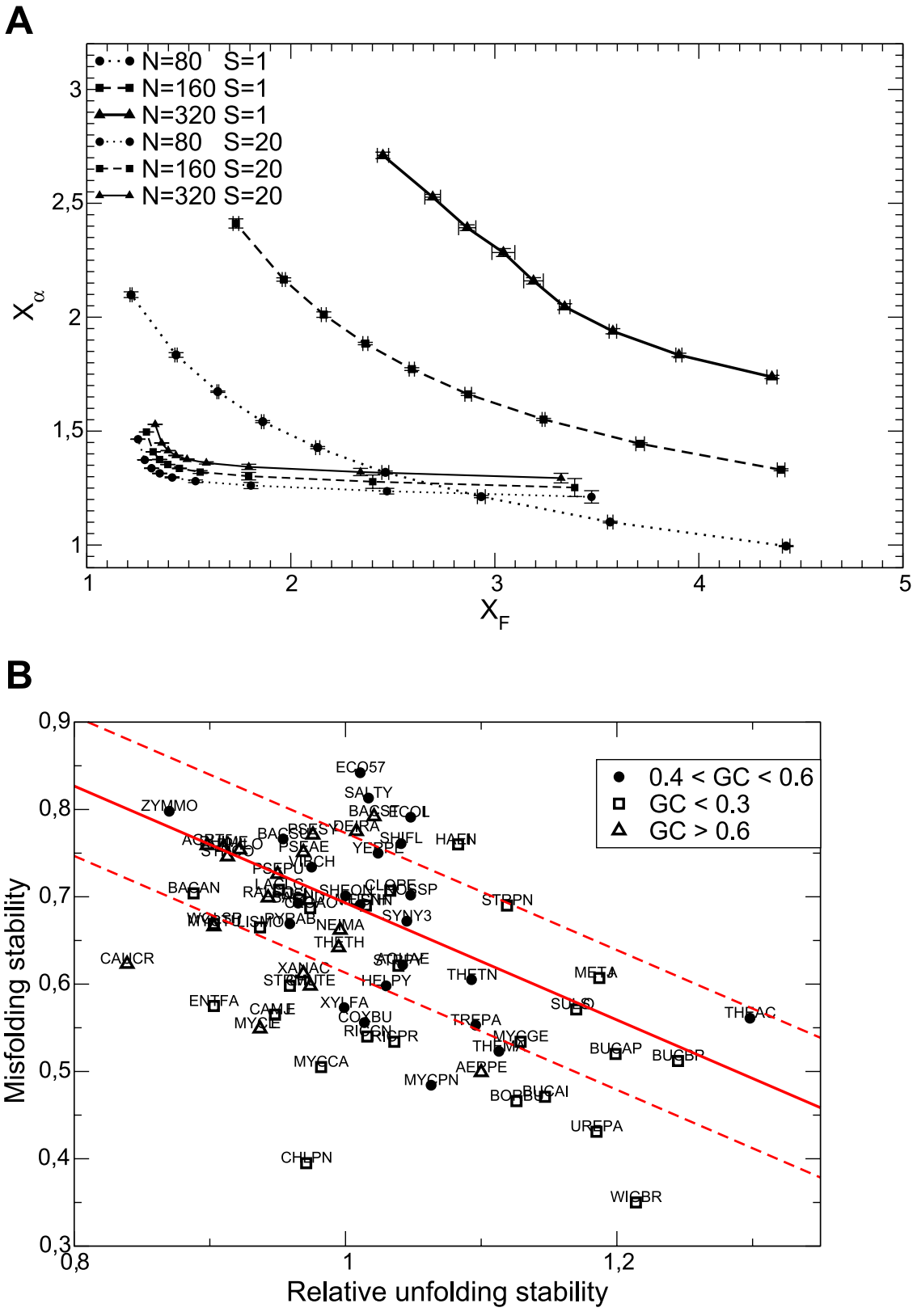

Figure 10. Negative correlation between misfolding and unfolding stability. Upper plot: Simulation results for average misfolding stability x_α versus unfolding stability x_F for various mutation biases, three population sizes and neutrality exponent $S=1$ (non-neutral regime) and $S=20$ (neutral regime). Bottom plot: Estimated misfolding versus unfolding stability for families of homologous proteins in prokaryotic genomes (data from Ref. [12]). We distinguish genomes according to GC content at third codon position. The solid line represents a linear fit of misfolding stability for genomes with moderate or no mutation bias ($0.4 \leq GC \leq 0.6$).
doi:10.1371/journal.pcbi.1000767.g010

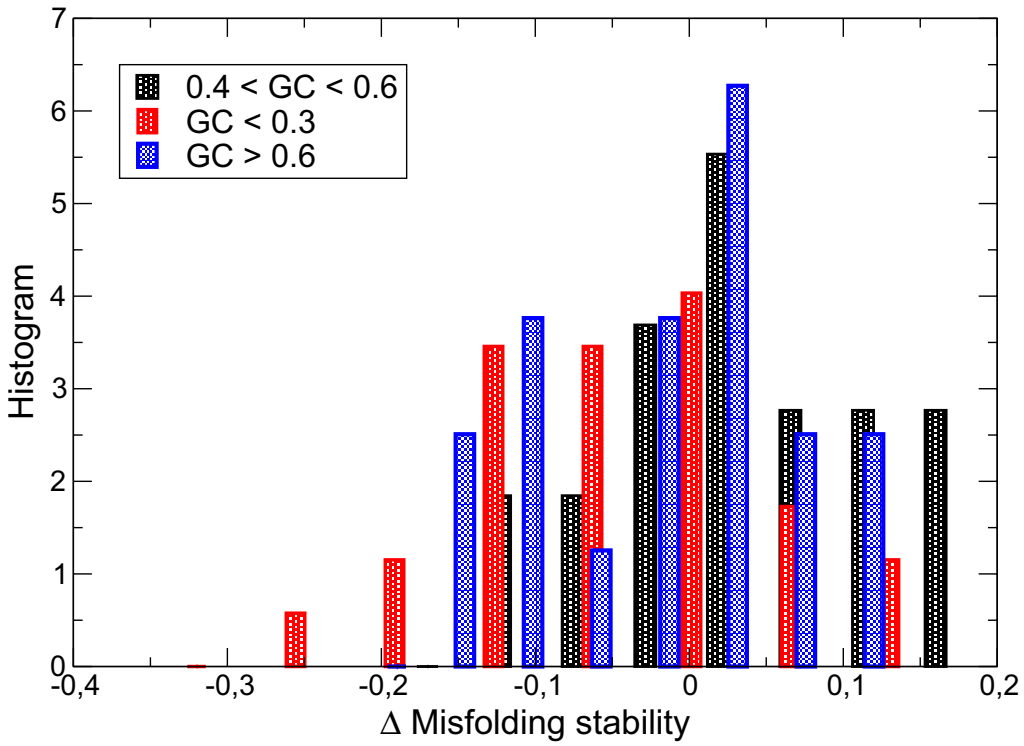

Figure 11. Relationship between GC usage and protein folding stability in orthologous proteins in different prokaryotic genomes (data taken from Ref. [12]). Histogram of the difference between the actual misfolding stability and the misfolding stability expected from the unfolding stability, using the relationship derived from species with moderate bias (continuous line in the previous plot). Notice that species with small and large GC usage have smaller than expected misfolding stability.
doi:10.1371/journal.pcbi.1000767.g011

can influence protein stability, and that the optimal mutation process is influenced by properties of the selection process is not new, but the extension of this concept to the evolution of the mutation bias is novel to our knowledge.

Quite interestingly, small populations attain higher fitness with AT bias, intermediate populations get an advantage with GC usage, and very large populations attain higher fitness with almost absent bias. This result establishes a deep interplay between population size and mutation bias. The ML equations show that the optimal GC usage depends on how the number of stable sequences decreases with the stability values, i.e. it is an effect of probability in sequence space. For very small population size and stabilities the optimal mutation bias is attained at small GC usage, which makes folding easier. At higher stabilities (intermediate population size) the optimal GC usage increases, therewith improving the stability against misfolding at the optimal GC. Approaching the maximal stabilities the optimal GC usage decreases again towards the value 0.5, which means absence of bias in the mutation process.

As a speculative remark, we note that it was not obvious that our model would predict $\mathrm{GC}_{\mathrm{opt}} \approx 0.5$ as the optimal GC usage for very large populations. In this limit the absolute maximum fitness is reached. We have shown numerically (see Text S1) that the optimal GC usage in the infinite population limit is little dependent on the parameters of the fitness function S, α_{thr} and F_{thr}, as long as the selective pressure affects mostly x_{α}, so that in this limit $\mathrm{GC}_{\mathrm{opt}}$ mainly depends on the contact energy parameters and on the genetic code. This conjecture is consistent with our data. Nevertheless, a systematic test requires cumbersome simulations that we did not perform here. We obtained a different result when using the GKS contact energy parameters, which yielded $\mathrm{GC}_{\mathrm{opt}} \approx 0.2$ for $B_{\alpha} = 9.5$ in the very large population limit. However, we notice that these parameters also produced a very small normalized energy gap, which suggests that they might be less suitable for this kind of study.

Influence of the mutation rate

The model that we adopt here is based on the assumption that the population is genetically homogeneous, i.e. the product $N\mu$ of population size times mutation rate is small. This allows us to analytically compute the fixation probability of a new mutation through Eq. (1) instead of explicitly simulating population dynamics. This approximation is considered valid if μ measures the mutation rate of a single protein, in particular if population size is small. However, the high mutation rates of RNA viruses may violate this assumption even for a single protein, and in this case several works [58,59] have shown that the load due to nonviable mutations significantly modifies the evolutionary process

even in the case of a neutral fitness landscape, leading to the evolution of mutational robustness and enhanced folding stability [60–62]. This situation can be studied analytically in the framework of the quasi-species theory [63]. We did not consider this theory here, because it assumes that the population size is infinite and therefore it prevents to study the effect of finite populations that is the main focus of the present work. If we considered a whole evolving genome instead of a single protein, the approximation of very small mutation rate would not be justified, since genomic mutation rates are in a range of 0.003 to 0.004 mutations per genome per generation for DNA-based microbes, including viruses, bacteria, and eukaryotes [55]. In this context, a new interesting effect has to be considered, namely the hitch-hiking effect, which consists in the fixation of mildly disfavoured alleles driven by a positively selected allele present in the same chromosome. However, since treating the hitch-hiking effect would make both the analytic and the numeric study much more complicated, we leave it as a subsequent step.

Robustness of the results

Our model depends on several assumptions and parameters. As evolutionary model, we adopted the Moran process, one of the best studied population genetic models. The theoretical work by Sella and Hirsh [17] shows that other evolutionary processes, such as for instance the Wright-Fisher process, would yield the same qualitative results. The mutation process was modelled using a single parameter, the GC usage. While this parametrization might appear too simplified, it has the merit to focus on a variable whose relevance has been pointed out by a large number of experimental studies, statistical analysis and models.

The ingredients of our model that seem more debatable are the form of the fitness function and its parameters S, α_{thr} and F_{thr}. To test the robustness of our results, we simulated different functional forms of the fitness function, using exponential functions of stability instead of power laws or letting the fitness depend only on the minimum between the two stabilities x_α and x_F. In all cases, we found the same qualitative results: There is an optimal mutation bias at which the fitness is maximal, such that for very small populations the optimal bias is towards AT, and for intermediate populations the optimal bias is towards GC. We then studied in detail the fitness function Eq. (2). Changing the neutrality exponent does not modify the qualitative results as long as the combination of S and N is in the non-neutral regime. Experiments on the evolution of small populations [13,14] and computational studies of protein folding stability [12] suggest that stability does depend on population size for populations subject to repeated bottlenecks, so that for such populations it is justified to assume that the non-neutral regime is the relevant evolutionary regime. We also varied the neutral thresholds α_{thr} and F_{thr} by more than 20%, finding that they do not change the qualitative behavior, although they have a quantitative influence on the optimal GC usage. We observed more important quantitative changes when we changed the contact energy parameters, but even in this case the gross qualitative features of the $\mathrm{GC}_{\mathrm{opt}}$ versus N relationship remain valid.

Meta-population evolution of the optimal bias

The result that the mutation bias directly influences the fitness that a population can attain in its evolution suggests the intriguing possibility that there may be a feedback between mutation and selection such that a particular mutation bias favors optimal protein folding stability, and selection may favor the replication machinery yielding this optimal mutation bias. Nevertheless, the selective advantage of evolving with the optimal GC usage is only apparent after a sufficiently large number of substitutions in protein coding genes. A mutant for GC usage would have a very low selective advantage during the first generations, and therefore its fixation would be a matter of almost neutral genetic drift. After the mutant is fixed, however, our model predicts that the population evolving with optimal bias will accumulate a sufficiently high selective advantage to take over populations with a less favourable GC usage when they, or their hosts in the important case of endosymbiotic bacteria, come to compete. Therefore, we expect this meta-population selection to almost deterministically favour the selection of the strain with optimal GC usage in contrast to the almost neutral fixation of a mutant with optimal GC usage within a single population. Thus the optimal mutation bias can facilitate the selection of more stable proteins and, on a longer time scale, selection at the meta-population level may favor the replication machinery that is most suitable to protein stability.

The population sizes at which we find the maximum of $\mathrm{GC}_{\mathrm{opt}}$ in our model are of the order of a few hundreds individuals for $S=1$. These values appear very small compared with real bacterial populations, even if they tend to grow rapidly for very high or very low neutrality exponent S. We may reconcile our model with biology if we notice that the effective population size is not the same as the total number of individuals of a species. Berg [42] showed that, if a small number of individuals often colonize new habitats with colonization probability almost independent of the founders fitness, the effective population size is given by the number of generations between two colonization events. This is a very small number for obligatory endosymbiotic and parasitic bacteria, and it may also be small for facultative parasites or symbionts, and even for the paradigm of a free living bacterium such as *Escherichia coli* for which Berg [42] estimated an effective population of 10^5 individuals.

The meta-population structure of bacterial species raises the question of whether the molecular evolution properties of a species such as the codon usage bias and the K_a/K_S ratio are primarily determined by the effective size of a local population or by the global size of the meta-population. This is an important question that requires modelling the meta-population dynamics and the different levels of selection that are relevant for it. Our opinion is that both population sizes influence the evolutionary dynamics, and that, despite the losses of stability of small local populations can be in part compensated at the meta-population level, the influence on evolution of the local population size remains important even taking into account these corrections, so that observables such as codon usage bias and K_a/K_S strongly reflect the local structure of the population.

Comparison with observed mutation bias

The distribution of GC content observed in bacterial genomes is remarkably broad. We assume here, as it is widely believed, that these differences in the GC content are mainly determined by different mutation pressures [18,19]. The third codon position, where a shift from A to G and from C to T does not change the coded amino acid in most cases, is thought to strongly reflect the mutation bias. However, the GC content at third codon position is strongly correlated with the GC content at first and second codon position [20,21], and through this correlation, the mutation bias influences the properties of the protein sequence, most notably its hydrophobicity [12,22]. This is surprising, since hydrophobicity is considered the main determinant of folding stability [23], and it is expected to be finely tuned since the protein has to avoid unfolding on one hand, and misfolding and aggregation on the other hand (of course this balance is very different for membrane proteins,

which are not considered here). One possible interpretation is that, due to the trade-off between unfolding and misfolding, the hydrophobicity is to some extent neutral so that it is possible to modify it without significantly affecting the global fitness of the protein. Our results suggest a different interpretation: There may be an optimal range of hydrophobicity, but this range may be different for different values of protein stability. So proteins with low stability, as those found in small populations, may tend to be more hydrophobic than proteins with high stability as those found in large populations, hence leading to a preference for a lower GC usage in their evolution.

Our model predicts that species with large population size will tend to evolve without mutation bias (GC usage equal to 0.5), whereas species with small and intermediate populations will tend to present such a bias, either towards AT or towards GC. This prediction is in qualitative agreement with two independent estimates of effective population size based on optimal codon usage and on the ratio between non-synonymous and synonymous substitutions represented in Fig. 9, and with a computational comparison of unfolding and misfolding stabilities in orthologous bacterial proteins, see Fig. 11. Of course bacterial genomes are rather complex, and we do not expect the mechanism proposed here to explain their GC content as the result of a single factor, population size. Another important factor influencing the GC content has been identified in a previous statistical study, which demonstrated that aerobiosis is an important determinant of GC rich genomes [64]. This interesting result is not in contradiction with our model, since many bacteria with small GC content tend to have an intracellular lifestyle, which in turn can make them anaerobic and at the same time reduce their effective population size.

As mentioned above, the proposed relationship between low GC content and small population size is consistent with the known fact that most bacterial species that adopted an intracellular lifestyle shifted their mutation spectrum towards AT with respect to their free living relatives [26]. This AT bias is, in most cases, the consequence of the loss of repair genes. For instance, three out of the four sequenced species of *Buchnera* lost the gene mutH, which in *Escherichia coli* is responsible of repairing the replication errors produced by methylation of cytosine that causes C to T mutation [65]. Moran proposed that this loss of repair genes and the consequent mutation bias is a selectively nearly neutral event in the evolution of endosymbionts [9]. Nevertheless, the results presented here suggest that this shift has important consequences on the folding properties of the whole proteome. In fact, a strong AT bias, together with reduced population size, is expected to produce severe misfolding problems, as indicated by the low predicted misfolding stability of proteins of intracellular bacteria with respect to orthologous ones in free living bacteria [12], and by the observed positive selection and over-expression of molecular chaperones in endosymbiotic bacteria [66], which is an expensive but effective strategy to reduce misfolding problems. Interestingly, it has been found that the fitness observed in an experimental population subject to frequent bottlenecks can be in part recovered by over-expressing chaperones [15]. Nevertheless, AT bias also enhances stability with respect to unfolding, and the results presented here suggest that its influence on fitness is globally positive for small populations.

The relationship between small population size and GC richness is even less expected. Only a few out of several prokaryotic species having high GC content are obligatory intracellular bacteria, such as for instance *Mycobacterium leprae*, and some are facultative pathogens or plants associated symbionts. Our results suggest the intriguing possibility that they tend to have small population size, although larger than for obligatory endosymbionts. To test this prediction, we estimated the population size using optimal codon usage [44], which has often been used to estimate population sizes. There are several caveats: The selective advantage of optimal codon usage strongly varies from one gene to another, and from one species to another. However, it is expected that the average codon usage bias estimated on the whole genome is correlated with population size. The optimal codon usage is computed subtracting the average mutation background, therefore it should not be trivially influenced by mutation bias. We found significantly reduced selection for optimal codon usage in bacteria evolving with large mutation bias compared to those with moderate or no bias, supporting our prediction that the former are characterized by smaller effective population size. Furthermore, we tested the relationship between GC content and effective population size estimating the latter through the ratio between non-synonymous to synonymous substitutions computed by Daubin and Moran [46] for entire bacterial genomes. This analysis presents important caveats. For instance, the non-synonymous substitution rate has been shown to depend on the time separation between two species [47–49]. We tackled this point by eliminating values of K_A/K_S estimated at short timescales, which are known to be strongly overestimated. Given the above, it is remarkable that the qualitative picture provided by this measure qualitatively coincides with the one obtained analysing optimal codon usage. Both measures strongly support the prediction of our model that species with $GC=0.5$ are characterized by larger effective population size. Nevertheless, among species presenting large mutation bias, those with bias towards GC are estimated through the K_A/K_S measure to have smaller effective population than those with bias towards AT, which is in contrast with our prediction. This point is worth further investigation taking into account more carefully the time dependency of the K_A/K_S estimate [48].

Of course, there exist several exceptions to these predictions, as there are several other factors, some already identified [64,67] and others still unknown, that influence the differences in GC content of prokaryotic species. One remarkable exception to the association between intracellularity and low GC content is the genome of the endosymbiotic bacterium *Hodgkinia cicadicola*, very recently sequenced by Moran's group [68]. This genome is extremely reduced (144 kb), as generally observed for endosymbiotic bacteria, but it shows GC content of 58%, which came as a big surprise since it is probably the most serious exception to the association between genome size and GC content. This genome also challenges the association between endosymbiotic bacteria and AT bias. It has been suggested that *Hodgkinia* belongs to the Rhizobiales division of alpha proteobacteria, characterized by high GC content. Interestingly, the genetic code of *Hodgkinia* underwent a modification such that UGA codes for Tryptophan instead of Stop. This modification is expected to ease the evolution of proteins that are stable with respect to misfolding. Consistently with this expectation, we found that the optimal GC usage for small populations slightly increases when this alternative genetic code is used in simulations, but this effect is too small to reconcile the GC content of *Hodgkinia* with its expected small effective population size (data not shown). Further research is needed to identify the origin of the GC content in this genome that lacks any repair gene [68]. Nevertheless, the association between intracellular lifestyle and AT bias, despite not being deterministic as demonstrated by this counterexample, is still strongly significant.

A second exception is represented by *Prochlorococcus marinus*, a very abundant species of small marine cyanobacteria [69,70]. It is expected that this species has a very large population size, which is

in agreement with a recent estimate of its K_A/K_S ratio [46]. 11 out of 13 fully sequenced strains of this cyanobacterium present low GC content, in the range between 30 to 38 percent, apparently contradicting the association between large population size and lack of mutation bias. However, the two remaining strains have GC content of 50%, as expected according to our model, and one of these was used to estimate the small K_A/K_S ratio that supports the large population size. *Prochlorococcus* has a complex meta-population structure in which the strains with 50% GC content, characterized by large genomes, appear to act as gene reservoirs. These strains are also characterized by a larger cell size than other *Prochlorococcus* strains, which the authors describe as "a feature that may have led to their lower isolation recovery due to the filtration step most often used to separate *Prochlorococcus* from *Synechococcus*. Hence, there are probably more LL-adapted *Prochlorococcus* strains with cell and genome sizes similar to those of *Synechococcus* thriving deep in the euphotic zone. This is apparently confirmed by the dominance of this ecotype at the base of the euphotic zone in the Atlantic Ocean, as revealed by quantitative PCR data" [70]. These strains with large genomes and without mutation bias are found at considerable depth in the ocean and thus at low oxygen pressure. There seems to be a positive association between ocean depth and GC content for *Prochlorococcus* strains, thus a negative association between oxygen pressure and GC content, opposite to the observed general association between oxygen and GC content [64]. Comparative analysis of the sequenced *Prochlorococcus* strains will be necessary to test the hypothesis that there is an association between the GC content and the population size of these strains. Consistent with this possible association, it was found that in the MED4 strain, characterized by the smallest GC content among all *Prochlorococcus* strains, translational selection does not shape the codon usage variation among the genes in this organism [71].

Conclusions

We have shown here that the AT mutation bias can increase the fitness associated with essential proteins if the population size is very small. The same happens with GC mutation bias for intermediate population. These results suggest that the mutation bias is not selectively neutral, but it may be the preferred outcome for the evolution of small populations. We found a deep interplay between the estimated effective population size and the GC content that is consistent with the predictions of our model. Of course this association is not deterministic, since many other factors influence the GC content. However, the influence of population size is an intriguing one that we believe is worth further investigation. Thus, we hope that this proposal will be subject to experimental test in the future.

Materials and Methods

Folding stability

As in our previous work, the unfolding free energy of a protein with sequence $A_\alpha \cdots A_L$ and contact matrix $C_{ij}=1$ if the minimal interatomic distance between residues i and j is below 4.5Å, 0 otherwise, is defined as

$$F(\mathbf{A}) = \sum_{ij} C_{ij} U(A_i, A_j) + sL, \tag{9}$$

where $U(a,b)$ is the contact interaction matrix determined in [72], $s=0.074$ was determined fitting Eq. (9) to a set of experimentally measured unfolding free energy (UB, unpublished) and L is protein length. Although rather simple, this model is accurate enough to allow quantitative predictions of the folding free energy of small proteins that fold with two-state thermodynamics (the correlation coefficient between experimental and predicted free energy is $r=0.92$ over a representative test set of 20 proteins, UB, unpublished result) and of the stability effect of mutations (correlation coefficient $r=0.72$ over a set of 195 mutations, UB, unpublished result). This is comparable to state-of-the-art programs such as Fold-X [73]. However, the computational simplicity of the model makes it affordable to use it for simulating very long evolutionary trajectories with a large number of parameters, which would not be possible using other tools.

The normalized energy gap α measures how alternative compact conformations are higher in energy than the native, and it is defined using the random energy model [74,75] as

$$\alpha(\mathbf{A}) = \frac{\sum_{ij} C_{ij} U(A_i, A_j) - N_c \langle e \rangle_{\mathbf{A}} + \sigma_{e,\mathbf{A}} \sqrt{2 N_c (AL + B)}}{\sum_{ij} C_{ij} U(A_i, A_j)(1 - q_0)} \tag{10}$$

with $A=0.1$, $B=4$, $q_0=0.1$, $N_c=\sum_{ij} C_{ij}$, and $\langle e \rangle_{\mathbf{A}}$ and $\sigma_{e,\mathbf{A}}$ are the mean and standard deviation of the interaction energy of both native and non-native contacts in sequence $\mathbf{A}$.

Protein structures

We studied five proteins with different size and secondary structures: Phosphocarrier protein of *E.Coli* (85 amino acids, PDB id. 1opd), Lysozyme of *G.Gallus* (129 amino acids, PDB id. 3lzt), ATP synthase epsilon chain of *E.Coli* (135 amino acids, PDB id. 1aqt), Triose Phosphate Isomerase of *E.Coli* (255 amino acids, PDB id. 1tre) and Tryptophan Synthase alpha chain of *S. Typhimurium* (260 amino acids, PDB id. 1a50). When not otherwise stated, we exemplify our results with the structure of the protein lysozyme.

Mutation process

Mutations are modelled through the HKY process [28], in which the mutation rate from nucleotide n to n', $T_\mu(n,n')$, is $\mu f(n')$ if $n \to n'$ is a transition, $\mu k f(n')$ if it is a transversion. The transition/transversion ratio is fixed at $k=2$. The microscopic rate μ is assumed to be very small and it does not affect the results. We further assume $\pi(\mathrm{A})=\pi(\mathrm{T})$ and $\pi(\mathrm{C})=\pi(\mathrm{G})$ (Chargaff second parity rule), so that the only parameter of the mutation model is the stationary GC content, $\mathrm{GC}=\pi(\mathrm{C})+\pi(\mathrm{G})$, which we call GC usage.

Simulation of the evolutionary process

Simulations were performed starting from the native sequence, which was changed through random mutations subject to the acceptance probability Eq. (1) computed using the estimated folding stabilities. We checked that simulations converged in all cases after a number of accepted substitutions not larger than a few times the protein length L, and discarded the first $8 \times L$ steps of the trajectory for collecting statistics. The simulations were run until $2000 \times L$ accepted substitutions were collected, which makes it rather cumbersome to simulate large populations for which the acceptance rate is small. For each set of parameters we run 10 independent simulations in order to evaluate the statistical error.

At every step, we randomly draw one mutating DNA site j with probability dependent on the nucleotide n_j that occupies it, $P_j \propto \sum_{n' \neq n_j} T_\mu(n_j, n')$, and we draw a new nucleotide $n' \neq n_j$ with probability proportional to $T_\mu(n_j, n')$. The mutation is then translated to the amino acid sequence, whose stability is computed through Eq. (9) and (10) from which we obtain fitness through Eq.

(2). The fitness is compared to the one of the current wild type sequence and the mutation is accepted with probability given by Eq. (1).

Optimal mutation bias

For fixed N and S the equilibrium fitness f is simulated for 9 GC usages from 0.1 to 0.9 and the results are fitted to a cubic function, from which we obtain the optimal GC at the point where the first derivative vanishes. If f(GC) is monotonically increasing or decreasing the maximum (minimum) GC is chosen. To estimate the error, we estimated GC_{opt} from 10 independent simulations, and we computed mean and standard error of the mean.

Supporting Information

Text S1 Supporting figures and analytic developments
Found at: doi:10.1371/journal.pcbi.1000767.s001 (0.23 MB PDF)

Acknowledgments

We acknowledge contributions of Andreas Buhr in early stages of this work.

Author Contributions

Conceived and designed the experiments: MP UB. Performed the experiments: RM MF UB. Analyzed the data: RM MF MP UB. Wrote the paper: MP UB. Wrote the simulation code: UB.

References

1. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
2. Kimura M (1983) The neutral theory of molecular evolution Cambridge Univ. Press.
3. Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? Proteins 46: 105–109.
4. Muller HJ (1932) Some Genetic Aspects of Sex. American Naturalist 66: 118–138.
5. Wright SG (1938) The distribution of gene frequencies in populations of polyploids. Proc Natl Acad Sci USA 24: 372–377.
6. Fisher RA (1958) The genetical theory of natural selection. DoverNew York.
7. Ohta T (1976) Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor Pop Biol 10: 254–275.
8. Graur D, Li WH (2000) Fundamentals of molecular evolution, Sinauer, Sunderland.
9. Moran NA (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. Proc Natl Acad Sci USA 95: 4458–4462.
10. 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.
11. Lambert DJ, Moran NA (1998) Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proc Natl Acad Sci USA 95: 4458–4462.
12. Bastolla U, Moya A, Viguera E, van Ham RCHJ (2004) Genomic determinants of protein folding thermodynamics. J Mol Biol 343: 1451–1466.
13. Duarte E, Clarke D, Moya A, Domingo E, Holland J (1992) Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc Natl Acad Sci USA 89: 6015–6019.
14. Novella IS, Dutta RN, Wilke CO (2008) A linear relationship between fitness and the logarithm of the critical bottleneck size in vesicular stomatitis virus populations. J Virol 82: 12589–12590.
15. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E (2002) Endosymbiotic bacteria: GroEL buffers against deleterious mutations. Nature 417: 398.
16. Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. BMC Evol Biol 4: 42.
17. Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102: 9541–9546.
18. Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84: 166–169.
19. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams H (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101: 3480–5.
20. 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.
21. Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 24: 1–11.
22. D'Onofrio G, Jabbari K, Musto H, Bernardi G (1999) The correlation of protein hydropathy with the base composition of coding sequences. Gene 1999 238: 3–14.
23. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14: 1–63.
24. Uversky VN (2003) Protein folding revisited. A polypeptide chain at the folding – misfolding – nonfolding cross-roads: Which way to go? Cell Mol Life Sci 60: 1852–1871.
25. Bastolla U, Porto M, Roman HE, Vendruscolo M (2006) A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. BMC Evol Biol 6: 43.
26. Silva FLatorre, Gomez-Valero AL, Moya A (2008) Genomic Changes in Bacteria: From Free-Living to Endosymbiotic Life. Structural Approaches to Sequence Evolution. Bastolla U, Porto M, Roman HE, Vendruscolo M, eds. Springer.
27. Durrett R (2002) Probability models for DNA sequence evolution, Springer.
28. Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
29. Bastolla U, Porto M, Roman HE, Vendruscolo M (2002) Lack of self-averaging in neutral evolution of proteins. Phys Rev Lett 89: 208101.
30. Bastolla U, Porto M, Roman HE, Vendruscolo M (2003) Statistical properties of neutral evolution. J Mol Evol 57: S103–S119.
31. Govindarajan S, Goldstein RA (1998) On the thermodynamic hypothesis of protein folding. Proc Natl Acad Sci USA 95: 5545–5549.
32. 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.
33. Babajide A, Hofacker IL, Sippl MJ, Stadler PF (1997) Neutral networks in protein space. Fol Des 2: 261–269.
34. Bussemaker HJ, Thirumalai D, Bhattacharjee JK (1997) Thermodynamic stability of folded proteins against mutations. Phys Rev Lett 79: 3530–3533.
35. Tiana G, Broglia RA, Roman HE, Vigezzi E, Shakhnovich EI (1998) Folding and misfolding of designed proteinlike chains with mutations. J Chem Phys 108: 757–761.
36. Mirny LA, Abkevich VI, Shakhnovich EI (1998) How evolution makes proteins fold quickly. Proc Natl Acad Sci USA 95: 4976–4981.
37. Dokholyan NV, Shakhnovich EI (2001) Understanding hierarchical protein evolution from first principles. J Mol Biol 312: 289–307.
38. Parisi G, Echave J (2001) Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol 18: 750–756.
39. DePristo MA, Weinreich DM, Hartl DL (2005 Sep) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6(9): 678–87.
40. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH (2005) Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A 102: 606–611.
41. Godzik A, Koli ski A, Skolnick J (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 4: 2107–17.
42. Berg OG (1996) Selection Intensity for Codon Bias and the Effective Population Size of Escherichia coli. Genetics 142: 1379–1382.
43. Petit N, Barbadilla A (2009) Selection efficiency and effective population size in Drosophila species. J Evol Biol 22: 515–26.
44. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: A test for translational selection. Nucl Ac Res 32: 5036–5044.
45. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucl Ac Res 33: 1141–1153.
46. Daubin V, Moran NA (2004) Comment on "The Origins of Genome Complexity". Science 306: 978.
47. Ho SY, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22: 1561–8.
48. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239: 226–35.
49. Peterson GI, Masel J (2009) Quantitative prediction of molecular clock and Ka/Ks at short timescales. Mol Biol Evol. doi 10.1093/molbev/msp175.
50. Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol Microbiol. 60: 820–7.
51. Loh E, Salk JJ, Loeb LA (2010) Optimization of DNA polymerase mutation rates during bacterial evolution. Proc Natl Acad Sci U.S.A. [Epub ahead of print].
52. Nilsson M, Snoad N. Optimal mutation rates in dynamic environments. Bull Math Biol 64: 1033–43.
53. Brumer Y, Shakhnovich EI (2004) Host-parasite coevolution and optimal mutation rates for semiconservative quasispecies. Phys Rev E Stat 69: 061909.
54. Clune J, Misevic D, Ofria C, Lenski RE, Elena SF, Sanjuán R (2008) Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. PLoS Comput Biol 4: e1000187.

55. Drake JW (2009) Avoiding dangerous missense: thermophiles display especially low mutation rates. PLoS Genet 5: e1000520.
56. 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.
57. Chen P, Shakhnovich EI (2009) Lethal mutagenesis in viruses and bacteria. Genetics 183: 639–50.
58. van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96: 9716–9720.
59. Wilke CO (2004) Molecular clock in neutral protein evolution. BMC Genetics 5: 25.
60. Taverna DM, Goldstein RA (2002) Why are proteins so robust to site mutations? J Mol Biol 315: 479–84.
61. Bloom JD, Lu Z, Chen D, Raval A, Venturelli OS, Arnold FA (2007) Evolution favors protein mutational robustness in sufficiently large populations. BMC Biology 5: 29.
62. Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of neutral protein evolution. Genetics 175: 255–66.
63. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58: 465–523.
64. Naya H, Romero H, Zavala A, Alvarez B, Musto H (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 55: 260–264.
65. van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A (2003) Reductive genome evolution in Buchnera aphidicola. Proc Natl Acad Sci USA 100: 581–586.
66. Fares MA, Moya A, Barrio E (2004) GroEL and the maintenance of bacterial endosymbiosis. Trends Genet 20: 413–416.
67. Musto H, Naya H, Zavala A, Romero H, Alvarez-Val n F, Bernardi G (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun 347: 1–3.
68. McCutcheon JP, McDonald BR, Moran NA (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genet 5: e1000565.
69. Kettler, et al. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet 3: e231.
70. Scanlan, et al. Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews 73: 249–299.
71. Banerjee T, Ghosh TC (2006) Gene expression level shapes the amino acid usages in Prochlorococcus marinus MED4. J Biomol Struct Dyn 23: 547–54.
72. Bastolla U, Farwer J, Knapp EW, Vendruscolo M (2001) How to guarantee optimal stability for most representative structures in the protein data bank. Proteins 44: 79–96.
73. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369–87.
74. Derrida B (1981) Random Energy Model: an exactly solvable model of disordered systems. Phys Rev B 24: 2613–2626.
75. Shakhnovich EI, Gutin AM (1989) Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys Chem 34: 187–199.