Protein evolution speed depends on its stability and abundance and on chaperone concentrations.

Permalink
https://escholarship.org/uc/item/99k107r8

Journal
Proceedings of the National Academy of Sciences of the United States of America, 115(37)

ISSN
1091-6490

Authors
Agozzino, Luca
Dill, Ken A

Publication Date
2018-09-01

DOI
10.1073/pnas.1810194115

Peer reviewed
Protein evolution speed depends on its stability and abundance and on chaperone concentrations

Luca Agozzino\textsuperscript{a,b} and Ken A. Dill\textsuperscript{a,b,c,1}

\textsuperscript{a}Laue Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794-5252; \textsuperscript{b}Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794-3800; and \textsuperscript{c}Department of Chemistry, Stony Brook University, Stony Brook, NY 11790-3400

Contributed by Ken A. Dill, July 23, 2018 (sent for review June 13, 2018; reviewed by Irene A. Chen and Claus O. Wilke)

Proteins evolve at different rates. What drives the speed of protein sequence changes? Two main factors are a protein’s folding stability and aggregation propensity. By combining the hydrophobic–polar (HP) model with the Zwanzig–Szabo–Bagchi rate theory, we find that: (i) Adaptation is strongly accelerated by selection pressure, explaining the broad variation from days to thousands of years over which organisms adapt to new environments. (ii) The proteins that adapt fastest are those that are not very stably folded, because their fitness landscapes are steepest. And because heating destabilizes folded proteins, we predict that cells should adapt faster when put into warmer rather than cooler environments. (iii) Increasing protein abundance slows down evolution (the substitution rate of the sequence) because a typical protein is not perfectly fit, so increasing its number of copies reduces the cell’s fitness. (iv) However, chaperones can mitigate this abundance effect and accelerate evolution (also called evolutionary capacitance) by effectively enhancing protein stability. This model explains key observations about protein evolution rates.

Author contributions: K.A.D. designed research; L.A. and K.A.D. performed research; and L.A. and K.A.D. wrote the paper.

Reviewers: I.A.C., University of California, Santa Barbara; and C.O.W., The University of Texas at Austin.

The authors declare no conflict of interest.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

\textsuperscript{1}To whom correspondence should be addressed. Email: dill@laufercenter.org.

This article contains supporting information online at www.pnas.orglookup/suppl/doi/10.1073/pnas.1810194115/dcsupplemental.

Published online August 27, 2018.

SI Appendix

Protein evolution speed depends on its stability and abundance and on chaperone concentrations

What molecular properties determine the rates of cell evolution? Proteins are known to evolve at different rates, partly based on the functions they perform for the cell, but also depending on their physical properties, such as folding stability and propensity for aggregation (1–8), and also depending on their companion chaperoning (9–16). While some evolution takes place over thousands to millions of years, other evolution can be much faster. Cancer cells evolve over a human lifetime. And pathogenic cells can evolve resistance to drugs in just a few years (17–19) or even faster (18). How do the molecular properties of proteins and chaperones determine the speed of evolution? Here, we develop theory for the rates of protein evolution.

Computing the Evolutionary Equilibria and Dynamics of Protein Sequences

The rate that a protein molecule evolves is given by the dependence on time \( t \) of the probability \( P_i(t) \) that a protein sequence \( i \) is fixed in a population by the time \( t \), through mutation and selection. Before considering the dynamics, we note that the equilibrium distribution of such probabilities will be a Boltzmann-like exponential, as shown (20–23) (and given by an alternative derivation using maximum entropy applied to sequence space in SI Appendix, Eq. S4):

\[
P^\ast_i = g_i \frac{e^{-\lambda V_i}}{Q},
\]

where \( V_i \) is the fitness potential, which is related to the fitness landscape \( f_i \) (24) by \( V_i = -\log f_i \); \( g_i \) is the sequence degeneracy—that is, the number of different sequences of a given fitness; \( \lambda \) is the selective pressure, proportional to the effective population size, as shown in ref. 20; and \( Q = \sum_i g_i e^{-\lambda V_i} \) is the sum over the statistical weights (relative populations) of the different sequences of the protein. (The fitness landscape is a mathematical surface, often multidimensional, of the cell’s fitness as a function of the different mutations of a given protein.) Eq. 1 gives the equilibrium population of sequences for a given fitness potential. This equilibrium distribution is useful for considering the dynamics below.

A Zwanzig–Szabo–Bagchi-Like Model of Protein Adaptation Rates

We model a protein’s evolutionary kinetics by adapting Zwanzig–Szabo–Bagchi (ZSB) theory applied to the different problem of protein-folding speeds (25, 26). On the one hand, protein-folding dynamics is quite a different process than protein evolution. In folding, a particular protein explores its conformational degrees of freedom, changing its shape, whereas in evolution, a protein undergoes changes of sequence through mutations and selection. However, the dynamics can be modeled by a similar formalism. We define the transition rate from an ancestor sequence \( i \) to descendant sequence \( j \) as \( W_{ji} \) through a process of mutations and selection steps. Then, the change in population of sequence \( i \) in a small time interval is given by the master equation expressing the “flow” from different sequences into sequence \( j \), minus the flows out from \( j \) to other sequences,

\[
\frac{dP_j(t)}{dt} = \sum_i (W_{ji}P_i(t) - W_{ij}P_j(t)).
\]

To solve the dynamics, we need to know the transition rates \( W_{ij} \); these are dictated by the shape of the fitness potential \( V_i \) since

Significance

Some biological evolution is slow (millions of years), and some is fast (months to years). The speed at which a protein evolves depends on how stable a protein’s folded structure is, how well it avoids aggregation, and how well-chaperoned it is. What are the mechanisms? We compute fitness landscapes by combining a model of protein-folding equilibria with sequence-change dynamics. We find that adapting to a new environment is fastest for proteins that are least stably folded, because those sit on steep downhill parts of fitness potentials. The modeling shows that cells should adapt to warmer environments faster than to colder ones, explains why increasing a protein’s abundance slows cell evolution, and explains how chaperones accelerate evolution by mitigating this effect.
the rates are related to the equilibrium probabilities $P_i^*$, which is given by Eq. 1. Then, we can solve for two key dynamical quantities: (i) the adaptation time, $\tau_A$, or peak time, which is the minimum time required for changes in a sequence $i$, through mutation and selection, to reach the sequence that is optimally adapted to its environment; or (ii) the substitution time, $\tau_S$, also called the exit time, which is the average time required for a sequence $i$ to change and become any other sequence. The inverse of each of these times is a corresponding rate. The substitution rate is also called the evolution rate. The adaptation rate and substitution rate are measured differently and give different insights. SI Appendix gives the details of the dynamical model; here, we just summarize the main points.

A Protein’s Adaptation Rate Depends Strongly on the Selection Pressure

First, we ask how protein adaptation can sometimes be very fast. For this exploration of principle, it is sufficient to adopt the very simplest model of a fitness landscape that has a single peak. We assume the fitness potential is linear in the number of mutations $m$ in a single protein (meaning that the fitness landscape is exponential), with slope $V_0$ and minimum $-V^*$ (which is the fitness point of the optimal sequence) (both $V_0$ and $V^*$ are taken to be positive quantities):

$$ V(m) = -V^* + mV_0. \quad [3] $$

The virtue of the linear landscape here is in allowing for a closed-form expression for the adaptation time (SI Appendix, Eqs. S8–S18),

$$ \tau_A \simeq \frac{(1 + ze^{-\lambda V_0})^L}{\omega_0 L} \quad [4] $$

where $z$ is the number of possible mutations a residue in the protein can have relative to its starting sequence ($z = 19$), $L$ is the total number of residues in the protein, and $\omega_0$ is the average fixation rate for a single point mutation. (If $L$ is large, the adaptation time is independent of the number of mutations; it becomes equally hard to find the peak, no matter what the starting sequence is.)

Fig. 1 shows a key conclusion: A protein’s adaptation speed can vary over nine orders of magnitude as a result of only a twofold change in selection pressure $\lambda$. This huge magnification in Eq. 4 is because the adaptation rate is nearly an exponential function of an exponential $[k_A = 1/\tau_A \sim (e^{\lambda V_0})^{L/z}]$. So, even though evolution “would take forever” if fitness landscapes were flat, even a very slight tilt of a fitness landscape gets amplified into a very fast adaptation for protein sequence evolution. [The treatment is valid in the limit of strong selection and weak mutations, for which populations are monomorphic and mutations do not interact with each other. Other contexts require different methods (27).] This general conclusion holds also if instead we had used other hypothetical functional forms of fitness. Here, we have considered just a single isolated protein. Below, we consider situations where mutations happen in multiple proteins.

Proteins Having the Steepest Fitness Landscapes Adapt the Fastest

Eq. 4 shows another key point, namely, that the adaptation rate $k_A$ increases strongly with the steepness, $V_0$, of the fitness potential. Metaphorically, a ball rolls faster down a steeper hill than down a shallower hill. (In the limit of a small slope, adaptation will follow a random walk in a large space, requiring an exponentially long time.)

The Least-Stable Proteins Adapt the Fastest Because Their Fitness Landscapes Are the Steepest

Above, we asked how external pressures affect adaptation speed. Here, we ask how the properties of the protein itself affects its adaptation speed. So, first, we need a model for how fitness depends on protein properties. Ever since the pioneering work of Drummond et al. (1, 3, 6, 21, 29), a major idea has been the misfolding avoidance hypothesis; namely, that a protein’s fitness is substantially due to its folding-unfolding equilibrium. Here, we give a model of the evolution rates. Consider a protein $i$ having folding stability, $\Delta G_i = G_i^{(0)} - G_i^{(f)}$, where $G_i^{(0)}$ is the native folded state, $G_i^{(f)}$ is the unfolded state, and $G_i^{(c)}$ is the abundance $A_i$. Let the number of different types of proteins in the cell be $m_{\text{tot}}$. A well-known result is how the cell’s fitness potential $V$ is the following nonlinear function of its folding stability (6):

$$ V(T, m_{\text{tot}}; \{\Delta G\}) = -c \sum_{i=1}^{m_{\text{tot}}} A_i \left( \frac{\exp(-\Delta G_i/RT)}{1 + \exp(-\Delta G_i/RT)} \right) $$

Eq. 5 simply states that each protein’s fitness potential is proportional to the product of (its abundance, $A_i$, in the proteome) $\times$ (its fractional degree of folding $[\text{native/(native + unfolded)}]$) $\times$ (the total number of protein types in the cell). Fitness potential is assumed to be linearly proportional to the number of folded copies of the protein, but only up to the point of overexpression. Folding stability and aggregation are not the only physical contributors to evolution rates; conformational flexibility, which we do not study here, can also affect evolvability, particularly in virus proteins (30–32).

Here, we model the evolution rates. We combine the fitness potential in Eq. 5 with the principle given by Eq. 4 that proteins undergo the fastest evolutionary adaptation where protein fitness landscapes are steepest.

First, compare two proteins: One protein is more stable than the other. The logic above says that the less-stable protein will accumulate adaptive mutations faster than the more-stable protein. Second, compare a “fit” protein, which is stably folded and well adapted to its environment, to a mutated version of that same protein, which is less stably folded and less fit. The mutant protein will acquire adaptive mutations faster than the well-adapted protein.

Fig. 2 illustrates that fast adaptation happens where the fitness potential is steep, which is where protein stability is marginal (near $\Delta G_i = 0$, neither stably folded nor substantially unfolded), for a given abundance $A_i$.

The curve in Fig. 2 is general and applicable when both stabilizing and destabilizing directions are accessible to the protein. However, we note that adaptation requires mutations in...
Cells Should Adapt Faster to a Warmer Environment than to a Colder One

How fast can proteins adapt if cells are put in climates of different temperatures? Some unicellular organisms (mesophiles) live in moderate-temperature environments (~40°C for *Escherichia coli*), while others (thermophiles) live in hotter environments. Cells grow the fastest at the temperature of their natural environment (33–35°C), but moved to different environments having different temperatures, they can adapt (36). We compute the speed of adaptation of a cell that is transferred from its normal environment to a new environment having either a higher or lower temperature. We compute rates from SI Appendix, Eq. S16, with a fitness potential given by the thermal folding Eq. 5, and using SI Appendix, Eq. S18 to find its slope along the mutation axis. In this example, we consider a given number of proteins in the cell with either zero or one adaptive mutations to each protein (assuming no epistasis).

Fig. 3 shows the prediction that cells should be able to adapt much faster to a warmer environment than to a cooler environment. (We are unaware of experiments that bear on this.) Fig. 4 illustrates the reason for this, using a fitness landscape. Start with a healthy mesophilic cell in its normal environment, say, at T = 40°C, where it is maximally fit. Its proteins are stably folded. Now, upshift its environment to T = 70°C (path 1) (slowly, in small steps, to avoid killing the cell). Initially, the cell is unfit for its new, warmer environment because its proteins are less stable at this higher temperature, T = 70°C. Now, mutations accumulate rapidly (30 total, in the model example) because the fitness landscape is steep for proteins that are unstable, leading to fast adaptation to the new peak (path 2).

Now, contrast this with cooling. Now, a thermophilic cell starts at T = 70°C, maximally fit, with its proteins stably folded. Cooling causes this cell to be less fit for its new environment at T = 40°C (path 3). However, this is not due to protein stability; cooling proteins that are already stable does not change their native populations. Rather, the reduced fitness upon cooling is because of the Arrhenius temperature factor: Cells naturally grow more slowly in colder temperatures (SI Appendix, Eq. S18). Overall, for this cooling situation, the cell’s fitness landscape has a shallow slope (along path 4), and adaptation to the cold through mutations is slow. In summary, cells should adapt to warm climates faster than to colder ones.

The Substitution Rate vs. Adaptation Rate: They Reflect Different Features of Fitness Terrains

For the rest of this work, we now switch attention from the adaptation rate (how fast an arbitrary sequence evolves to become the sequence that has the maximal fitness) to the substitution rate (also called the evolution rate: how fast an arbitrary sequence changes to become fixed as a different arbitrary sequence). This switch allows us to test predictions against experimental data for the properties studied below. Substitution rates are properties of individual proteins, meaning that the accumulation of multiple mutations can take a long time. In contrast, adaptation involves mutations that can occur in parallel throughout the entire proteome, and therefore those changes can happen much faster. For this reason, for the remainder of the work we will be counting the number of mutations in a single protein, as opposed to what was done in the previous section.

More importantly, these two rate properties reflect different features of fitness landscapes. Whereas our model shows that adaptation rates are proportional to the slope of a fitness landscape (see above), substitution rates, instead, are proportional to the average mutational distance of a given protein to its fitness peak (at equilibrium) (see below and SI Appendix, Eq. S27):

$$\langle W \rangle \approx \mu_0 \sum_{m>0} m e^{-\lambda V(m)/Q} = \mu_0 \langle m \rangle$$  \[6\]

where $\mu_0$ is a rate quantity used as a fitting parameter and $\langle m \rangle$ is the average number of sequence mutations from the optimum in a single protein, and hence, the average mutational distance from an hypothetical optimal sequence at equilibrium. Fig. 5 shows the interpretation of $\langle m \rangle$. It measures the weight under the curve, so substitution rates are highest on fitness landscape contours that are “high-shouldered”: plateaus of high fitness where slopes are shallow. The bluescape Fig. 5 is high-shouldered, with larger $\langle m \rangle$: A mutation in either direction (green arrow) is fit enough, so substitution is fast. The orangescape is not a high plateau or flat. It has smaller $\langle m \rangle$: A mutation downhill (red arrow) is too unfit to be fixed. Because of greater access to allowed directions, Eq. 6 says that substitutions happen faster on the bluescape than on the orangescape. Moving away from the peak on the bluescape still leads to adaptive mutations, and hence, to substitution; moving away on the orangescape leads to nonadaptive mutations. So, the net substitution speed is greater on the bluescape. This high-shouldering principle is valid beyond the simple model used here to illustrate it.

![Fig. 3. Proteins should adapt faster to a warmer than to a cooler climate.](image-url)
Substitution rates of amino acids are measurable and have been the basis for the molecular clock idea (37–39) that substitution rates differ among proteins, but are approximately constant for a given protein. Recent work has shown that the average substitution rate is determined not by functional constraints, but by physical ones. Proteins that are more abundant are observed to evolve more slowly than proteins that are less abundant (40). The expression level-rate (E-R) anticorrelation is the observation that increasing expression levels (protein abundances) lead to reduced rates of their evolution. It has been hypothesized that this is a result of either protein misfolding or protein–protein interaction (21, 29, 41).

Abundant Proteins Evolve Slowly

We model the mechanism of the E-R anticorrelation. In a population of cells, many proteins are not peak-fitness sequences. Increasing the abundance of these imperfect proteins reduces the cell’s fitness relative to a perfectly adapted cell. We consider two mechanisms: (i) misfolding, where fitness, $V_{\text{core}}^\text{peak}(n)$, depends on how perfectly a protein sequence folds in its lowest-energy state to maximize hydrophobic–hydrophobic (HH) contacts in the core of its native structure. The deviation from the fitness peak is a count of the number of defects, $m = 0, 1, \ldots, N_c$. (ii) For aggregation and misinteraction, fitness, $V_{\text{surf}}^\text{peak}(m)$, depends on how perfectly the protein surface is covered with polar (P) residues, to avoid protein–protein sticking through HH contacts. The deviation from perfect fitness is $m = 0, 1, \ldots, N_s$, the number of hydrophobic (H) residues on the surface. Now, to get these fitness landscapes, we use the hydrophobic–polar (HP) lattice model, in which a protein is assumed to have only H or P residues, and different native and mutated protein sequences are enumerated on a 2D square lattice (42). Random mutations over different proteins can reduce either form of “perfectness.” Details are given in SI Appendix, Eqs. S30 and S31. The main distinction between these mechanisms is their dependence on abundance $A$: $V_{\text{core}}^\text{peak}(n) \propto A$ and $V_{\text{surf}}^\text{peak}(n) \propto A^2$. We calculate the substitution rates for these two different mechanisms using Eq. 6.

The E-R Anticorrelation Is Explained by Either Misfolding or Aggregation or Both

Fig. 6 compares the misfolding and aggregation models to experiments. Both models predict a general E-R anticorrelation. And, both are consistent with the (not very precise) data (29). So, we have no basis for favoring one mechanism over the other. Previous modeling has also observed the E-R anticorrelation, but based on assuming an anticorrelation between $\Delta G$ and mutational $\Delta \Delta G$’s taken from the Protein Data Bank (6, 7). Our more microscopic mechanism here of the full evolutionary landscape allows us also to study aggregation and chaperone effects at a single-protein level.

The model explains the E-R anticorrelation as follows. Fig. 7 (yellow surface) shows the substitution rate as functions of both protein stability and abundance. In a very stable protein, a mutation that removes a hydrophobe from the core is usually acceptable (a high-shouldered terrain), so it has a high substitution rate. In contrast, in a weakly stable protein, removing a hydrophobe from the core can unfold the protein, so it is not adaptive (not a high-shouldered terrain), so fewer possible substitutions are acceptable, and the substitution rate is slower. The abundance effect is as follows. This is an integration over all of the proteins in the cell, and most of them are imperfect and not maximally stable. So, increasing the abundance manifests as increasing the concentration of imperfect proteins, which are not high-shouldered, leading to slower average substitution.

Chaperones Are Evolutionary Accelerators

The speed of cell evolution is modulated by chaperones in the cell. Chaperones are biomolecular complexes that help other proteins (their clients) to fold. Experiments show that chaperones are generally evolution accelerators (they have been called evolutionary capacitors). That is, increasing a cell’s chaperone concentrations can speed up the cell’s evolution (9, 10, 12, 14, 43).

What is the mechanism of evolutionary acceleration by chaperones? The blue surface in Fig. 7 shows the effect of adding
chaperones within the present model (details of calculations are in SI Appendix, section 1). Chaperones are active ATP-driven devices that shift the balance from misfolded and unfolded states to native states of client proteins, resulting in stabilizing the client proteins. And, according to theory above, more-stable proteins have terrains that are more high-shouldered, leading to faster amino acid substitution. Hence, chaperones accelerate evolution. (A subtle point is that while chaperones accelerate substitution, they can slow down adaptation, since chaperones can make “near-perfect” proteins appear perfect to the cell. Such near-perfect proteins have no selective disadvantage in cells with chaperones.)

Fig. 8 also shows the prediction that evolutionary acceleration can be different for different types of chaperones. To show this, we use the model of Santra et al. (44) that shows that GroEl binds misfolded protein, sending it either to unfolded (U) or native (N) states, whereas DnaK binds misfolded protein and only sends it to U; SI Appendix, Fig. S2. The model predicts that GroEl is effective at lower concentrations, and on a different set of client proteins, than DnaK. Why the difference in chaperone concentrations? In short, clients of GroEl see only an unstable U and stable N state, so those clients mostly fold. In contrast, clients of DnaK mostly see an M state that is almost as stable as N, so more chaperone is needed to produce more N. That is why GroEl and DnaK should have different values of “evolutionary capacitance” (see SI Appendix, Fig. S5 for a graphical explanation).

Conclusions
We model here how rates of protein evolution depend on folding and aggregation properties. The present model gives a single framework for understanding disparate observations. (i) Adaptation speed depends on the slope of a fitness landscape. This depends strongly on selection pressures. It is fastest for the least-stably folded proteins. Cells that are shifted to warm climates become unstable, so they adapt rapidly. (ii) Substitution speed depends on how high-shouldered the terrain is of a fitness potential. Abundant proteins evolve slowly. This is because most proteins are not perfectly stable, so increasing their abundance shifts the cell to non-high-shouldered terrains of fitness, leading to slow substitution. (iii) This effect is mitigated by chaperones, which increase protein stabilities, increasing their substitution rates. This modeling describes how
protein evolution rates depend on their folding and aggregation properties.

1. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102:14338–14343.

2. 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.

3. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

4. Yang J-R, Zhuang S-M, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol 6:421.

5. Serohijos AWR, Ryan Lee SY, Shakhnovich EI (2013) Highly abundant proteins favor more stable 3D structures in yeast. Biophys J 104:L1–L3.

6. Serohijos AWR, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep 2:249–256.

7. Yang JR, Liao BY, Zhuang SM, Zhang JZ (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 109:E831–E840.

8. Levy ED, De S, Teichmann SA (2012) Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci USA 109:20461–20466.

9. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336–342.

10. Santra M, Farrell DW, Dill KA (2017) Bacterial proteostasis balances energy and proteomic robustness. Genome Biol Evol 9:1766.

11. Nathan ML, Baquer F (2000) Mutation frequencies and antibiotic resistance. Antimicrob Agents Chemother 44:1771–1777.

12. Baym M, et al. (2016) Spatiotemporal microbial evolution on antibiotic landscapes. Science 353:1147–1151.

13. Greulich P, Wacław B, Allen R (2012) Mutational pathway determines whether drug gradients accelerate evolution of drug-resistant cells. Phys Rev Lett 109:1–5.

14. Selga L, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102:14475–14475.

15. Echave J, Wilke CO (2017) Biophysical models of protein evolution: Understanding the patterns of evolutionary sequence divergence. Annu Rev Biophys 46:85–103.

16. Barton NH, Coe JB (2009) On the application of statistical physics to evolutionary biology. J Theor Biol 259:317–324.

ACKNOWLEDGMENTS. This work was supported by Laufer Center and National Science Foundation Grant MCB1344230.