The effective temperature of mutations

Imre Derényi* and Gergely J. Szöllősi†
ELTE-MTA "Lendulet" Biophysics Research Group, Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary

Biological macromolecules experience two seemingly very different types of noise acting on different time scales: i) point mutations corresponding to changes in molecular sequence and ii) thermal fluctuations. Examining the secondary structures of a large number of microRNA precursor sequences and model lattice proteins, we show that the effects of single point mutations are statistically indistinguishable from those of an increase in temperature by a few tens of Kelvins. The existence of such an effective mutational temperature establishes a quantitative connection between robustness to genetic (mutational) and environmental (thermal) perturbations.

Biological systems are robust [1], and robustness is considered to be a fundamental feature of complex evolvable systems [2]. Molecular phenotypes, such as stable protein folds and functional RNA structures, have provided fundamental insight into the origin and principles of robustness in biological systems [3–10]. In these systems perturbations on different time scales, i.e., rare changes in sequence and omnipresent thermal fluctuations, seem very different. Surprisingly, several computational [11–14] and experimental [15–19] studies suggest a qualitative similarity between the effects of mutations and temperature. Stable proteins are in general more tolerant to point mutations [12, 15], and in the case of RNA, the set of structures explored by thermal fluctuations are highly correlated with the minimum free energy structures of single point mutants[11, 14].

The correlation between the effects of point mutations and temperature is less surprising if we recognize that each degree of freedom of the molecule has an average thermal energy of $k_{\mathrm{B}}T/2 \approx 2.5/2$ kJ/mol (where k_{B} is the Boltzmann constant and $T \approx 300$ K is the absolute temperature) and the typical free energy change associated with a point mutation is also of the order of a $k_{\mathrm{B}}T$ (approximately 4 kJ/mol for a protein [20], and about twice this large for the breaking of a hydrogen bond in a nucleotide base pair). Despite this similarity in energies, for single instances of the system, i.e., individual copies of protein or RNA molecules, permanent changes in sequence are clearly different from ephemeral thermal kicks. However, the distinction between mutational effects and thermal fluctuations becomes less manifest in large populations and over longer time scales where many possible point mutations are explored as molecules are copied (transcribed and translated) repeatedly via mechanisms prone to errors. From this perspective, perturbations of the phenotype (e.g., the protein fold or RNA secondary structure) resulting from mutations can be expected to have similar effects to thermal perturbations: both jostle the system between structural states with energies that differ by only a few times the thermal energy scale.

Here we demonstrate that this qualitative analogy between mutational and thermal perturbations can be taken to a quantitative level, and the effect of point mutations is well described as an effective increase in temperature. For the original, so called wild type (WT), sequence let the probability of the system being in any of its possible structural states (e.g., protein folds or RNA secondary structures) be denoted by

$$P_{\mathrm{WT}}(i,T) = \frac{1}{Z_{\mathrm{WT}}(T)} \exp\left(-\frac{G_{\mathrm{WT}}(i,T)}{k_{\mathrm{B}}T}\right) , \qquad (1)$$

where $G_{\mathrm{WT}}(i,T)$ is the free energy of structure i and $Z_{\mathrm{WT}}(T)$ is the partition function, and let the probability of the same structure transformed to an effective temperature T_{eff} be defined as

$$P_{\mathrm{eff}}(i,T,T_{\mathrm{eff}}) = \frac{1}{Z_{\mathrm{eff}}(T,T_{\mathrm{eff}})} \exp\left(-\frac{G_{\mathrm{WT}}(i,T)}{k_{\mathrm{B}}T_{\mathrm{eff}}}\right) , \qquad (2)$$

with normalization constant $Z_{\mathrm{eff}}(T,T_{\mathrm{eff}})$. Note that this transformation preserves the relative contributions of the enthalpic and entropic components of the free energy at the original temperature T. The mutation averaged probability of the system assuming structure i can be defined as

$$P_{\mathrm{mut}}(i,T) = \frac{1}{N_{\mathrm{mut}}} \sum_{m=1}^{N_{\mathrm{mut}}} p_m(i,T) , \qquad (3)$$

where the summation index m runs over all possible point mutations (of total number N_{mut}) and $p_m(i,T)$ denotes the probability of structure i for mutation m.

Examining the secondary structures of a large number of microRNA (miRNA) precursor sequences and model lattice proteins, we demonstrate that in most cases there exists a well defined effective temperature T^*_{eff}, for which the transformed probability distribution of the wild type approximates the average probability distribution of single point mutants with unanticipated precision:

$$P_{\mathrm{mut}}(i,T) \approx P_{\mathrm{eff}}(i,T,T^*_{\mathrm{eff}}) . \qquad (4)$$

In other words, we show that the effect of mutations are fully described by considering only the free energies $G_{\mathrm{WT}}(i,T)$ of the secondary structures i of the single *wild type* sequence and a single effective temperature T^*_{eff}.

We used the 23766 miRNA precursor sequences of miRBase version 9.0 [21] upto the length of 250 nucleotides. For each of these WT sequences we (i) generated all of its single point mutants, and then (ii) determined the equilibrium probability distributions (at $T = 300$ K) of the secondary

* derenyi@elte.hu
† ssolo@elte.hu

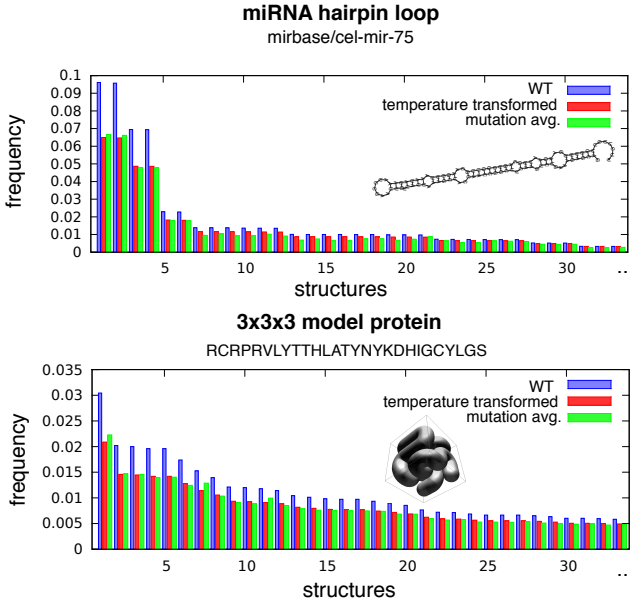

FIG. 1. **The effects of single point mutations and an increased effective temperature on the frequency of the 33 most frequent structures.** For a typical miRNA and a typical lattice protein we plot the equilibrium frequencies of structures (enumerated in decreasing order of their frequencies) in the WT distribution (blue bars); in the distribution transformed to the the optimal effective temperature $T^*_{\rm eff}$ defined in the main text (red bars); and in the mutation averaged distribution (green bars). Due to the large number of structures, only the first 33 structures are shown. The complete distributions (after binning) are displayed in Fig. 2. The insets illustrate the minimum free energy structures.

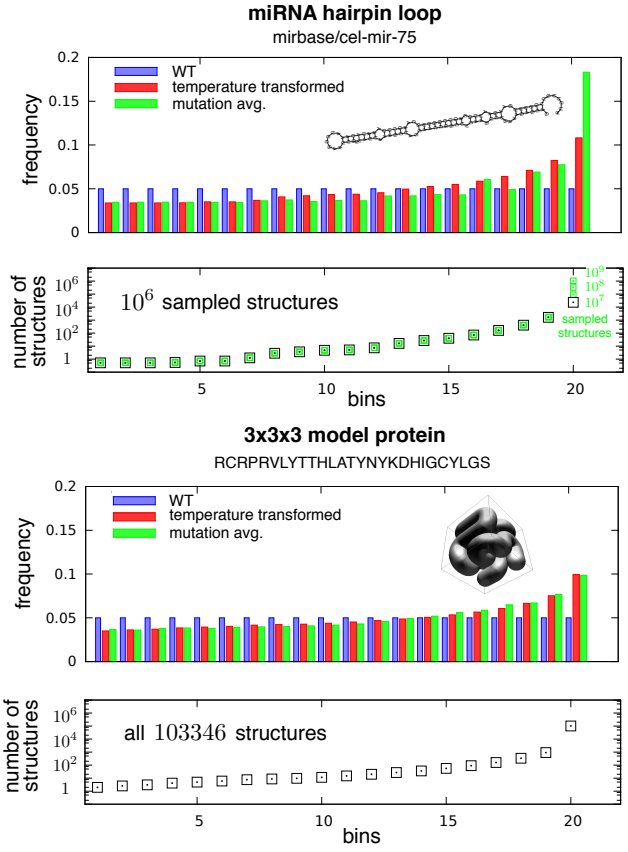

FIG. 2. **The effects of single point mutations and an increased effective temperature on the binned distribution of structures.** For a typical miRNA and a typical lattice protein (which are the same as in Fig. 1) the structures were grouped into 20 bins in such a way that each bin accommodated exactly $1/20$th of the WT equilibrium probability distribution (blue bars). The bottom panels share bins with the top panels, and show the number of structures in each bin for the particular miRNA and a lattice protein considered. For the miRNA sequence smaller green squares show the result of sampling additional structures up to a total of 10^7, 10^8 and 10^9 samples. The bottom panels show that a few of the most stable structures account for most of the probability. The histogram of the mutation averaged probability distribution (green bars) is tilted to the right, indicating that the mutations tend to destabilize the most stable structures. Raising the temperature has a similar effect, as demonstrated by the histogram of the WT distribution transformed to the optimal effective temperature $T^*_{\rm eff}$ defined in the main text (red bars). The insets illustrate the minimum free energy structures.

structures of the WT and the mutant sequences, by using the computational tools developed by Hofacker et al. [22] and sampling 10^6 structures for each sequence. To establish the generality of our results we also analyzed a markedly different system, the model of $3\times3\times3$ compact lattice proteins [5, 23]. In this model a protein is represented by a self-avoiding chain of beads placed on a $3\times3\times3$ square lattice and the free energy of a given structural state of a given sequence is defined as the sum of the interaction energies between nearest neighbors [24]. We first randomly selected 12000 sequences, and designated them as the WT. Subsequently, similarly to the miRNA precursor sequences, for each of these WT sequences we (i) generated all of its single point mutants, and (ii) determined the free energies of all 103346 possible structural states of the WT and the mutant sequences, in order to compute the equilibrium probabilities defined in Eqs. (1) to (3).

Two typical examples, a miRNA and a lattice protein, are shown in Fig. 1, displaying the above defined three distributions truncated to the first 33 most probable structures (according to the WT distributions). In order to visualize the entire distributions, in Fig. 2 we grouped the structures into 20 bins in such a way that each bin accommodated exactly $1/20$th of the WT equilibrium probability distribution ($P_{\rm WT}$). The bins were filled from left to right with structures in decreasing order of their WT probability (starting with the most probable one, i.e., with the structure having the lowest free energy $G_{\rm WT}$). To ensure that each bin contained equal probability, structures at bin boundaries were split among neighboring bins. The (generally non-integer) number of structures belonging to each bin is shown in the panels below the histograms illustrating that a few of the most stable structures account for most of the probability. Examining Fig. 2 we can see that the histogram of the WT equilibrium distribution ($P_{\rm WT}$, blue bars) is uniform by construction. The histogram of the mutation averaged probability distribution ($P_{\rm mut}$, green bars), however, deviates from uniformity and is tilted to the right, in-

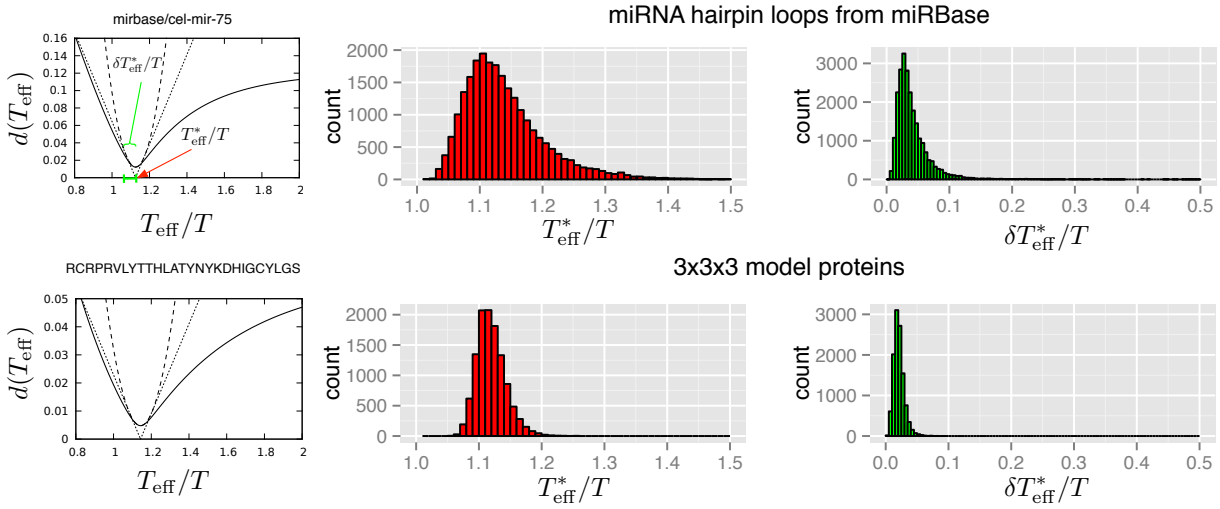

FIG. 3. **The distribution of the optimal effective temperature for miRNAs and lattice proteins.** In the leftmost panels the Euclidean distance $d(T_{\rm eff})$ (solid lines) is plotted as a function of the effective temperature $T_{\rm eff}$ for the two example sequences from Fig. 2. The other panels summarize results for all sequence considered. The middle panels show the distributions of the optimal effective temperatures $T^*_{\rm eff}$ (illustrated by the red arrow in the top left panel), while the rightmost panels show the distribution of uncertainties $\delta T^*_{\rm eff}$ (illustrated by the green interval in the top left panel).

dicating that the mutations tend to destabilize the most stable structures. Raising the temperature has a similar effect.

To quantify the similarity between the temperature transformed and mutation averaged distributions we need to use a distance measure. If there were an effective temperature where the temperature transformed distribution would coincide with the mutation averaged distribution, then any distance measure would give a zero distance at that particular temperature. The main finding of our work is that although an exact coincidence never occurs, for almost all of the investigated systems there exists a well defined effective temperature where the temperature transformed distribution is very similar to the mutation averaged distribution. Under these circumstances the effective temperature at which the distance is minimal is highly insensitive to the choice of the distance measure (which we checked for most widely used measures). Thus, for simplicity, we use the Euclidean distance

$$d(T_{\rm eff}) = \sqrt{\sum_i \left[P_{\rm eff}(i,T,T_{\rm eff}) - P_{\rm mut}(i,T)\right]^2}\,, \qquad (5)$$

as the distance measure, and define the optimal effective temperature $T^*_{\rm eff}$, where this distance is minimal. Figs. 1 and 2 show that the transformed distributions at the optimal temperature ($P_{\rm eff}$, red bars) approximate the mutation averaged distributions remarkably well. The only noticeable discrepancy (mostly observed for miRNAs) occurs at the rightmost bin, which contains rarely visited, high energy, non-native structures.

In the leftmost panels of Fig. 3 the Euclidean distance $d(T_{\rm eff})$ (solid lines) is plotted as a function of the effective temperature $T_{\rm eff}$ for both examples from Fig. 2. If $d(T^*_{\rm eff})$ were exactly zero the $d(T_{\rm eff})$ curve would have a sharp, symmetric, "V"-shaped bottom. This is because the square of the Euclidean distance ($d^2(T_{\rm eff})$) is an analytic function, which to leading order is expected to have a quadratic minimum. In reality, however, $d(T^*_{\rm eff})$ is never exactly zero, and the bottom of the $d(T_{\rm eff})$ becomes rounded, exhibiting a dull, parabolic minimum at $T^*_{\rm eff}$ (red arrow in the top left panel of Fig. 3). We quantify this "dullness" using the half width (green interval in the top left panel) of the fitting parabola (dashed lines) between the points where the parabola touches the two tangents (dotted lines) emanating from the horizontal axis at $T^*_{\rm eff}$:

$$\delta T^*_{\rm eff} = \sqrt{\frac{2d(T^*_{\rm eff})}{d''(T^*_{\rm eff})}}\,, \qquad (6)$$

where $d''(T^*_{\rm eff})$ denotes the second derivative of $d(T_{\rm eff})$ at the minimum. This "dullness", having the dimension of temperature, provides an estimation for the uncertainty in determining the effective temperature. Note that the dimensionless value of the minimal distance $d(T^*_{\rm eff})$ has no obvious physical meaning and, unlike the optimal temperature $T^*_{\rm eff}$ and the uncertainty $\delta T^*_{\rm eff}$, it is sensitive to the choice of the distance measure.

The middle and right panels of Fig. 3 show the distributions of the optimal effective temperatures $T^*_{\rm eff}$ and their uncertainties $\delta T^*_{\rm eff}$, respectively, for all studied miRNAs (top) and model proteins (bottom). The typical uncertainty ($\sim 0.02T$ for the proteins and $\sim 0.03T$ for the miRNAs) is considerably smaller than both the width of the corresponding $T^*_{\rm eff}$ distribution and the typical value of $T^*_{\rm eff}$, indicating the existence of a well defined optimal effective temperature for nearly all miRNAs and proteins.

For the miRNAs the length dependence of $T^*_{\rm eff}$ can also be examined, where the length L is defined as the number of nucleotides). The data show that $T^*_{\rm eff} \approx 12T/L + T$ (Pearson's $\rho = 0.582$, with $p < 2.2 \times 10^{-16}$). This is not surprising if

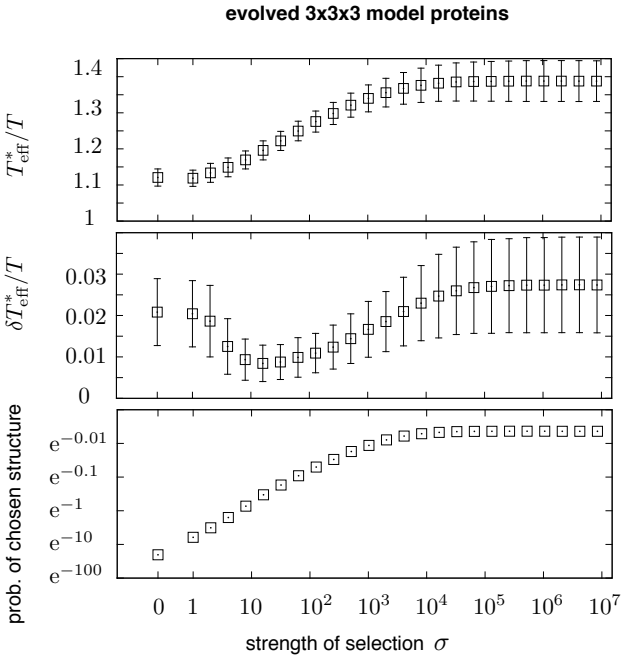

FIG. 4. **The effect of selection on the effective temperature.** Starting from 1000 random amino acid sequences and randomly choosing a structural state, we performed selection for the increased equilibrium frequency of the chosen structure. In each of the 1000 runs 10000 subsequent attempts of introducing random point mutations were executed. Each attempt was accepted (i) with probability 1 if it increased the equilibrium frequency of the chosen state; and (ii) with probability $(p'/p)^\sigma$, if it decreased the equilibrium frequency, where p and p' denote the equilibrium frequencies before and after the point mutation, respectively, and σ is the strength of the selection. The upper panel shows the optimal effective temperature, the middle panel the uncertainty, and the bottom panel the average equilibrium probability of the chosen structure as a function of the selection strength.

we realize that the most probable effect of a point mutation on the most stable structures is the breaking of a few hydrogen bonds and the concomitant increase of the free energy by an amount consistent with $12k_{\rm B}T$. Distributing this increase uniformly among the L nucleotides as a result of averaging over point mutations is analogous to raising the temperature by about $12T/L$.

A similar length dependence of $T^*_{\rm eff}$ is expected for the protein model (and any other systems), as well. Our data for $L = 27$ amino acids suggest a temperature increase of about $3T/L$, which is consistent with an energy cost of about $3k_{\rm B}T$ of changing an amino acid in the most stable structures. The wider distribution of $T^*_{\rm eff}$ for miRNAs compared to model proteins results, at least in part, from the variation in length of the miRNAs.

The above results demonstrate a strong analogy between the effects of mutational and thermal fluctuations on the structures of biological macromolecules, and justify the introduction of the concept of effective mutational temperature. The generality of our results are supported by the analysis of two markedly different systems: (i) real microRNA precursor sequences, shaped by millions of years of evolution; and (ii) randomly selected artificial lattice proteins, lacking any evolutionary history. Thus, our results also indicate that the existence of the effective mutational temperature does not require the systems to have undergone natural selection and, therefore, it must have shaped the evolution of macromolecules from the beginning of life.

An advantage of the model protein system is that it also allows the study of the effect of selection on the value of the effective mutational temperature. To address this issue we performed Monte Carlo simulations of selection for the increased equilibrium frequency of a prespecified structure (see Fig. 4 for the results and the details of the simulations). The results show that the effective mutational temperature increases with the strength of selection (which is not surprising, since the effect of a point mutation on a better adapted and, therefore, more fragile system is expected to be larger); and the uncertainty of the effective mutational temperature remains always small (interestingly, with a minimum near the selection strengths, where the prespecified structure starts to dominate the equilibrium distribution).

The diversity of proteins and nucleic acids we find in the living world are the result of evolution; their properties are determined by the laws of physics and chemistry [25]. To decipher the interplay of these two kinds of causality we have to understand the relationship between molecular sequence and function. The existence of a well defined mutational temperature demonstrates a general property of this relationship: the biophysics of RNA and protein structures imposes a statistical equivalence between the effects of mutational and thermal perturbations. In other words, the biological noise introduced by point mutations is quantitatively analogous to the physical noise generated by thermal fluctuations.

In an evolutionary context this implies that selection for robustness against either of these will produce, as a correlated by-product, robustness against the other. Our result suggests an explicit model of how maintaining stability, a major constraint in the evolution of biological macromolecules, leads to stability against point mutations (both at the short time scales of molecular replication and the long time scales of organism reproduction), and also facilitates opportunities for molecular innovation by allowing increased neutral variation [15, 26].

ACKNOWLEDGMENTS

This work was supported by the Hungarian Science Foundation (grant K101436). GJSz was supported by the Marie Curie CIG 618438 “Genestory” and the Albert Szent-Györgyi Call-Home Researcher Scholarship A1-SZGYA-FOK-13-0005 supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 “NationalExcellence Program”.

[1] A. Wagner, *Robustness and evolvability in living systems* (Princeton University Press, Princeton, N.J., 2005).
[2] H. Kitano, Nat Rev Genet **5**, 826 (2004).
[3] J. M. Smith, Nature **225**, 563 (1970).
[4] M. A. Huynen, P. F. Stadler, and W. Fontana, Proc Natl Acad Sci U S A **93**, 397 (1996).
[5] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[6] E. van Nimwegen, J. P. Crutchfield, and M. Huynen, Proc Natl Acad Sci U S A **96**, 9716 (1999).
[7] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual, Science **314**, 1898 (2006).
[8] G. J. Szöllősi and I. Derényi, Math Biosci **214**, 58 (2008).
[9] E. J. Hayden, E. Ferrada, and A. Wagner, Nature **474**, 92 (2011).
[10] A. Wagner, Trends Genet **27**, 397 (2011).
[11] L. W. Ancel and W. Fontana, J Exp Zool **288**, 242 (2000).
[12] J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold, Proc Natl Acad Sci U S A **102**, 606 (2005).
[13] A. Sakata, K. Hukushima, and K. Kaneko, Phys Rev Lett **102**, 148101 (2009).
[14] G. J. Szöllősi and I. Derényi, Mol Biol Evol **26**, 867 (2009).
[15] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, Proc Natl Acad Sci U S A **103**, 5869 (2006).
[16] P. Domingo-Calap, M. Pereira-Gómez, and R. Sanjuán, J Evol Biol **23**, 2453 (2010).
[17] A. S. Lauring, J. Frydman, and R. Andino, Nat Rev Microbiol **11**, 327 (2013).
[18] J. L. Payne and A. Wagner, Science **343**, 875 (2014).
[19] K. Kümpornsin, C. Modchang, A. Heinberg, E. H. Ekland, P. Jirawatcharadech, P. Chobson, N. Suwanakitti, S. Chaotheing, P. Wilairat, K. W. Deitsch, S. Kamchonwongpaisan, D. A. Fidock, L. A. Kirkman, Y. Yuthavong, and T. Chookajorn, Mol Biol Evol (2014), 10.1093/molbev/msu140.
[20] C. S. Wylie and E. I. Shakhnovich, Proc Natl Acad Sci U S A **108**, 9916 (2011).
[21] S. Griffiths-Jones, R. J. Grocock, S. Van Dongen, A. Bateman, and A. J. Enright, Nucleic acids research **34**, D140 (2006).
[22] I. Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster, Monatsh. Chem. **125**, 167 (1994).
[23] D. J. Lipman and W. J. Wilbur, Proc Biol Sci **245**, 7 (1991).
[24] S. Miyazawa and R. L. Jernigan, J Mol Biol **256**, 623 (1996).
[25] M. J. Harms and J. W. Thornton, Nat Rev Genet **14**, 559 (2013).
[26] A. Wagner, Proc Biol Sci **275**, 91 (2008).