Large phenotype jumps in biomolecular evolution

F. Bardou
IPCMS, CNRS and Université Louis Pasteur, 23 rue du Loess, BP 43, F-67034 Strasbourg Cedex 2, France

L. Jaeger
Chemistry and Biochemistry Department, University of California—Santa Barbara, Santa Barbara, California 93106-9510, USA
(Received 30 October 2003; published 24 March 2004)

By defining the phenotype of a biopolymer by its active three-dimensional shape, and its genotype by its primary sequence, we propose a model that predicts and characterizes the statistical distribution of a population of biopolymers with a specific phenotype that originated from a given genotypic sequence by a single mutational event. Depending on the ratio g_0 that characterizes the spread of potential energies of the mutated population with respect to temperature, three different statistical regimes have been identified. We suggest that biopolymers found in nature are in a critical regime with $g_0 \simeq 1-6$, corresponding to a broad, but not too broad, phenotypic distribution resembling a truncated Lévy flight. Thus the biopolymer phenotype can be considerably modified in just a few mutations. The proposed model is in good agreement with the experimental distribution of activities determined for a population of single mutants of a group-I ribozyme.

DOI: 10.1103/PhysRevE.69.031908 PACS number(s): 87.15.He, 87.15.Cc, 05.40.Fb, 87.23.Kg

I. INTRODUCTION

The biological function (or phenotype) of a biopolymer, such as a ribonucleic acid (RNA) or a protein, is mostly determined by the three-dimensional structure resulting from the folding of linear sequence of nucleotides (RNA) or aminoacids (proteins) that specifies a genotype. Generally, a natural biopolymer sequence (or genotype) codes for a specific two-dimensional or three-dimensional structure that defines the biopolymer activity. But one sequence can simultaneously fold in several metastable structures that can lead to different phenotypes. Thus, random mutations of a sequence induce random changes of the metastable structure populations, which generates a random walk of the biopolymer function. Understanding this phenotype random walk is a basic goal for ''quantitative'' biomolecular evolution.

The statistical properties of RNA secondary structures considered as a model for genotypes have been investigated in depth in the recent years [1]. The neutral network concept [2,3], i.e., the notion of a set of sequences, connected through point mutations, having roughly the same phenotype, has been shown to apply to RNA secondary structures. Thus, by drifting rapidly along the neutral network of its phenotype, a sequence may come close to another sequence with a qualitatively different phenotype, which facilitates the acquisition of new phenotypes through random evolution. Moreover, in the close vicinity of any sequence with a given structure, there exist sequences with nearly all other possible structures [4], as originally proposed in immunology [5]. Thus, even if the sequence space is much too vast to be explored through random mutations in a reasonable time (an RNA with 100 bases only has 10^{60} possible sequences), the phenotype space itself may be explored in a few mutations only, which is what matters biologically. These ideas have been brought into operation in a recent experiment [6] showing that a particular RNA sequence, catalyzing a given reaction, can be transformed into a sequence having a qualitatively different activity, using a small number of mutations and without ever going through inactive steps.

This paper investigates the phenotype space exploration at an elementary level by studying the statistical distribution of a population of biopolymers in a specific three-dimensional shape, that originated from a given genotypic sequence by a single mutational event. It complements studies of the evolution from one structure to another structure [7], that consider only the most stable structure for each sequence and neglect the thermodynamical coexistence of different structures for the same sequence. It also provides more grounds to the recent work that suggests that RNA molecules with novel phenotypes evolved from plastic populations, i.e., populations folding in several structures, of known RNA molecules [8]. It is experimentally evident, for instance, in Ref. [6], that some mutations change the biopolymer chemical activity by a few percent while other mutations change it by orders of magnitude. This is not unexpected since, depending on their positions in the sequence, some residues have a dramatic influence on the 3D conformation while others hardly matter. Thus, the function random walk statistically resembles a Lévy flight [9–11] presenting jumps at very different scales. The respective parts of gradual changes and of sudden jumps in biological evolution is a highly debated issue. While the gradualist point of view has historically dominated, evidences for the presence of jumps have accumulated at various hierarchical levels from paleontology [12], to trophic systems, chemical reaction networks and neutral networks, and molecular structure [7]. The jump issue will be treated here by studying the statistical distribution describing the phenotype effects of random mutations of a biopolymer genotype.

To address the question of the statistical effects of random mutations of functionally active biopolymers, we propose a model inspired from disordered systems physics that naturally predicts the possibility of broad distributions of activities of randomly mutated biopolymers. With two energy parameters describing the polymer energy landscape, this models is shown to exhibit a variety of behaviors and to fit

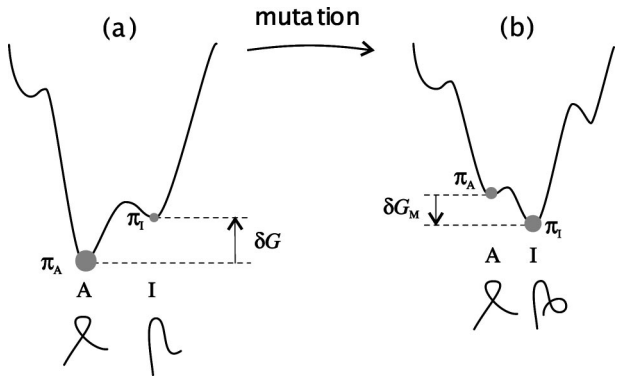

FIG. 1. Schematic representations of the molecular energy landscapes. (a) For the nonmutated molecule. (b) For the mutated molecule. Only the two lowest energy conformations A (active) and I (inactive) are taken into account. Their 3D conformations are indicated symbolically. The shaded dots indicate the populations π_A and π_I at thermal equilibrium.

experimental data. Natural biopolymers are in a critical regime, related to the activity distribution broadness, in which a single mutation may have a large, but not too large, effect.

II. PHYSICAL MODEL OF SHAPE POPULATION DISTRIBUTION

The most favorable conformational state of a biopolymer sequence with a given biological activity is generally considered to be the most stable one within the sequence energy landscape. The ruggedness of the energy landscape might vary depending on the number of other metastable conformational states accessible by the sequence. The typical energy spacing between these states can be small enough so that several states of low energy can be thermally populated. For simplicity, we will consider a sequence that is able to fold into its two lowest energy conformational states, an active state A of specific biological function, and an inactive state I of unknown function [24,25], but whose energy is the closest to A's (higher or lower) (see Fig. 1). The differences between the free energies of the unfolded and folded states for A and I are denoted ΔG_A and ΔG_I, respectively.

A mutation, i.e., a random change in the biopolymer sequence, modifies the biopolymer energy landscape so that ΔG_A and ΔG_I are transformed into $(\Delta G_A)_M$ and $(\Delta G_I)_M$. Note that the conformer state A of the mutant, its three-dimensional shape, is the same as before whereas the conformer state I does not have to be the same as before. To take into account the randomness of the mutational process, the mutant free energy difference $\delta G_M \equiv (\Delta G_I)_M - (\Delta G_A)_M$ is taken either with a Gaussian distribution

$$P_G(\delta G_M) \equiv \frac{1}{\sqrt{2\pi}\delta G_0} e^{-(\delta G_M - \overline{\delta G})^2/2\delta G_0^2} \tag{1}$$

or with a two-sided exponential (Laplace) distribution

$$P_e(\delta G_M) \equiv \frac{1}{2\,\delta G_0} e^{-|\delta G_M - \overline{\delta G}|/\delta G_0}, \tag{2}$$

where $\overline{\delta G}$ is the mean of δG_M and where δG_0 characterizes the width of the distribution. These two energy distributions are commonly used for disordered systems [13] and enable us to cover a range of situations from narrow (Gaussian) to relatively broad (exponential) distributions. Assuming thermodynamic rather than kinetic control, the populations π_A and $\pi_I = 1 - \pi_A$ of conformers A and I, respectively, are given by Boltzmann statistics

$$\pi_A = \frac{1}{1 + e^{-\delta G_M/RT}}, \tag{3}$$

where R is the gas constant and T is the temperature.

From the distributions of free energy differences and Eq. (3), one infers the probability distributions $P_{e \text{ or } G}(\pi_A)$ of the population of conformer state A after a mutation using $P_{e \text{ or } G}(\pi_A) = P_{e \text{ or } G}(\delta G_M)|d\,\delta G_M / d\pi_A|$. For the Gaussian model, one obtains

$$P_G(\pi_A) = \frac{\exp\left[\dfrac{-[\ln \pi_A - \ln(1-\pi_A) - \overline{g}]^2}{2g_0^2}\right]}{\sqrt{2\pi} g_0 \pi_A (1-\pi_A)}, \tag{4}$$

where $\overline{g} \equiv \overline{\delta G}/(RT)$, $g_0 \equiv \delta G_0/(RT)$. The ratio g_0 of the scale of energy fluctuations and of the thermal energy appears frequently in the study of the anomalous kinetics of disordered systems. For the exponential model, one obtains

$$P_e(\pi_A) = \frac{e^{-\overline{g}/g_0}}{2g_0 \pi_A^{1-1/g_0} (1-\pi_A)^{1+1/g_0}} \quad \text{for} \quad \pi_A \leqslant \pi_m, \tag{5a}$$

$$P_e(\pi_A) = \frac{e^{+\overline{g}/g_0}}{2g_0 \pi_A^{1+1/g_0} (1-\pi_A)^{1-1/g_0}} \quad \text{for} \quad \pi_A \geqslant \pi_m, \tag{5b}$$

with the same definitions for $\overline{g}$ and g_0, and $\pi_m \equiv (1 + e^{-\overline{g}})^{-1}$ (median population of A). Note that changing $\overline{g}$ into $-\overline{g}$ is equivalent to performing a symmetry on $P_{e \text{ or } G}(\pi_A)$ by replacing π_A by $1 - \pi_A$.

III. TYPES OF DISTRIBUTIONS

To analyze the different types of population distributions, we focus for definiteness on the Gaussian model. A qualitatively similar behavior is obtained for the exponential model. Figure 2 represents examples of $P_G(\pi_A)$ for the Gaussian model with $\overline{g} = -1$ and various g_0's. The negative value of $\overline{g}$ implies that A is on average less stable than I, and hence that π_A is predominantly less than 50%. For small g_0, the distribution $P_G(\pi_A)$ is narrow since the width δG_0 of the free energy distribution is small compared to RT so that there are only small fluctuations of population around the most probable value. When the energy broadness g_0 increases, the single narrow peak first broadens till, when $g_0 \gtrsim 1.976$ it splits into two peaks, close, respectively, to $\pi_A = 0$ and to $\pi_A = 1$. The broad character of $P_G(\pi_A)$ can be intuitively

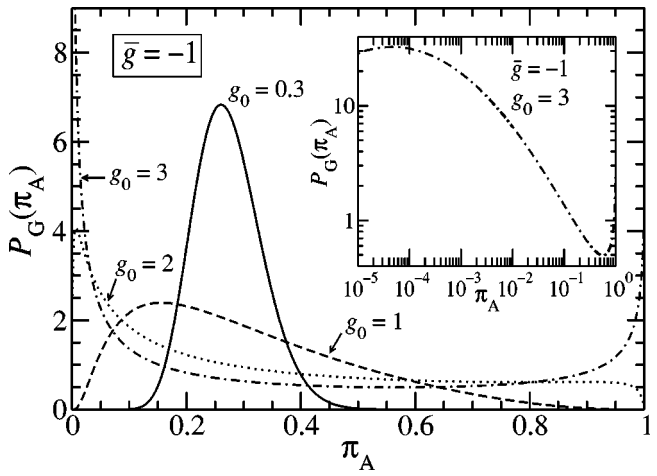

FIG. 2. Distributions $P_G(\pi_A)$ of shape populations of mutated molecules for $\bar{g}=-1$. They are narrow and single peaked for small enough g_0 and broad and double peaked for large enough g_0. The transition from one to two peaks occurs at $g_0 \simeq 1.976$ in agreement with Eq. (6). Inset: logarithmic plot of $P_G(\pi_A)$ for $g_0=3$ showing the broad character of the small π_A peak.

understood as a consequence of the nonlinear dependence of π_A on δG_M. Thus, when the fluctuations of δG_M are larger than RT, i.e., when $g_0 \gtrsim 1$, the quasiexponential dependence of π_A on δG_M [Eq. (3)] nonlinearly magnifies δG_M fluctuations to yield a broad π_A distribution, even if δG_M fluctuations are relatively small compared to the mean δG. A similar mechanism is at work for tunneling in disordered systems [14,15].

A global view of the possible shapes of $P_G(\pi_A)$ is given in Fig. 3. For any given $\bar{g}$, when increasing g_0 starting from 0, the single narrow peak of $P_G(\pi_A)$ first broadens then it splits into two peaks when $\bar{g}=\bar{g}_{\mathrm{sgn}(\bar{g})}(g_0)$ with

$$\bar{g}_{\pm}(g_0) \equiv \pm \left[g_0 \sqrt{g_0^2-2} + \ln\left(\frac{g_0 - \sqrt{g_0^2-2}}{g_0 + \sqrt{g_0^2-2}} \right) \right]. \qquad (6)$$

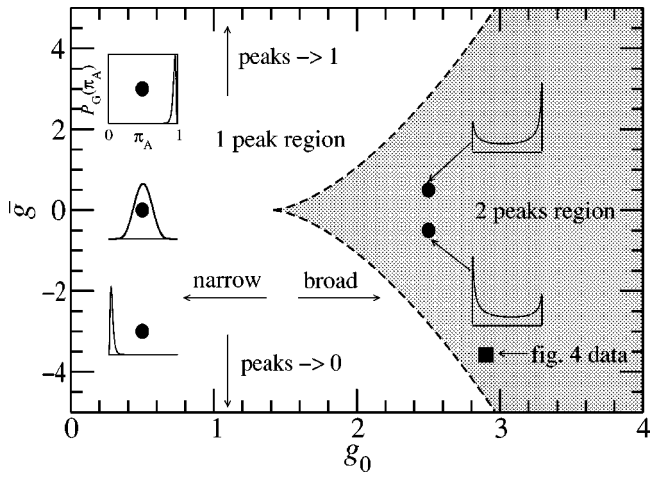

FIG. 3. Possible shapes of $P_G(\pi_A)$. The shaded area indicates the two-peaks region. The dashed line gives the transition from one to two peaks [cf. Eq. (6)]. Insets show examples of $P_G(\pi_A)$ corresponding to the g_0 and $\bar{g}$ indicated by the black dots (P_G's not to scale). The black square corresponds to the fit of Fig. 4 data.

[This expression results from a lengthy but straightforward study of $P_G(\pi_A)$.] When g_0 increases further, these two peaks get closer to $\pi_A=0$ and to $\pi_A=1$ while acquiring significant tails (see Sec. VII). For any given g_0, increasing $\bar{g}$ roughly amounts to moving the populations π_A towards larger values as expected since larger $\bar{g}$'s correspond to stabler states A. However, distinct behaviors arise depending on g_0. If $g_0 < \sqrt{2}$, whatever the value of $\bar{g}$, the distribution $P_G(\pi_A)$ is always sufficiently narrow to present a single peak. If $g_0 > \sqrt{2}$, the distribution $P_G(\pi_A)$ is sufficiently broad to have two peaks when, furthermore, the distribution is not too asymmetric, which occurs for $\bar{g} \in [\bar{g}_-(g_0), \bar{g}_+(g_0)]$. In short, depending on $\bar{g}=\delta\bar{G}/RT$, which characterizes mainly the peak(s) position, and on $g_0 = G_0/RT$, which characterizes mainly the distribution broadness, the distributions $P_G(\pi_A)$ are either unimodal or bimodal, either broad or narrow. This variety of behaviors is reminiscent of beta distributions.

IV. FROM SHAPE POPULATIONS TO CATALYTIC ACTIVITIES

Up to now, we have discussed the distribution $P(\pi_A)$ of the population of a shape A that is functionally active. However, as far as it concerns biopolymers with enzymatic functions, what is usually measured is a chemical activity a, i.e., the product of a reaction rate k for the conformer A by the population π_A of this conformer. The reaction rates are given by the Arrhenius law $k=k_0 e^{-E_a/RT}$ where k_0 is a constant and E_a is the activation energy. Thus, the chemical activity writes, using Eq. (3):

$$a = k_0 e^{-E_a/RT} \frac{1}{1+e^{-\delta G_M/RT}}. \qquad (7)$$

Random mutations may induce random modifications of E_a, δG_M, or both. Fluctuations of δG_M have been treated above. One can introduce fluctuations of E_a in the same way. We do not do it here in detail but present only the general trends.

The effects of adding an activation energy distribution in addition to the free energy difference distribution are twofold. For small activities, the distribution $P(a)$ of chemical activities is similar to the small $P(\pi_A)$ peak at small π_A. Indeed, the reaction rate k depends exponentially on E_a, just as the population π_A depends exponentially on δG_M when $\pi_A \ll 1$. Moreover, the product of two broadly distributed random variables is also broadly distributed [26] with a shape similar to the one of $P(\pi_A)$. For large activities, on the other hand, π_A and k behave differently because π_A is bounded by 1 while k is unbounded. Thus, if the k distribution is broad enough, the distribution of a at large a may exhibit a broadened structure compared to the $\pi_A \simeq 1$ peak of $P(\pi_A)$.

In summary, the distribution of chemical activities $P(a)$ is similar to the distribution of shape populations $P(\pi_A)$ when $P(\pi_A)$ presents a large $\pi_A \simeq 0$ peak (conditions for this to occur are explicited in Sec. VI). Thus, by observing

the shape of the $a \simeq 0$ peak in the activity distribution $P(a)$, one does not easily distinguish between activation energy dispersion, which affects k, and free energy difference dispersion, which affects π_A. On the other hand, at large a, $P(a)$ is differently influenced by activation energy dispersion and by free energy difference dispersion. The available experimental data (see Sec. V) enables us to precisely analyze $P(a)$ at small activities but not at large activities. Thus, for practical purposes, it is not meaningful in this paper to consider a distribution of activation energies on top of a distribution of free energy differences. In the sequel, we will thus do as if only the distribution of free energies was involved, stressing that similar effects can be obtained from a distribution of activation energies.

V. ANALYSIS OF EXPERIMENTAL DATA

Comparison of the theoretical distributions of Eqs. (5) and (4) with experimental data enables us to test the relevance of the proposed model. We have analyzed the measurements of the catalytic activities of a set of 157 mutants derived from a self-splicing group I ribozyme, a catalytic RNA molecule [16] (out of the 345 mutants generated in Ref. [16], we only considered the 157 ones with single point mutations). The original ''wild-type'' molecule is formed of a conserved catalytic core that catalyzes the cleavage of another part of the molecule considered as the substrate. The set of mutants is derived from the original ribozyme by systematically performing all single point mutations of the catalytic core, i.e., of the part of the molecule that most influences the catalytic activity. Nucleotides out of the core which, in general, influence the catalytic activity less, are left unmutated. Thus, in our framework, this set of mutants can be seen as biased towards deleterious mutations. Indeed, mutations of the quasioptimized core are likely to lead to much less active mutants, while mutations of remote parts are likely to leave the activity essentially unchanged. If all parts of the molecule had been mutated, more neutral or quasineutral mutations would have been obtained. Another point of view, which we adopt here is to consider the catalytic core as a molecule in itself, on which all possible single point mutations have been performed.

The 157 measured activities are used to calculate a population distribution with inhomogenous binning (cf. broad distribution). Two bins required special treatment: the smallest bin, centered in 0.5%, contains 40 mutants with nonmeasurably small activities ($<1\%$ of the original activity); the largest bin, centered in 95%, contains the six mutants with activities larger than 90% of the original ''wild'' RNA activity (the largest measured mutant activity is 140%). These two points, whose abscissae are arbitrary within an interval, are not essential for the obtained results. At last, as very few mutants have activities larger than the wild-type ribozyme, the proportionality constant between activity and population is set by matching a population $\pi_A = 1$ to the activity of the wild-type ribozyme.

The obtained distribution (see Fig. 4) has a large peak in $\pi_A \simeq 0$, indicating that most mutations are deleterious, with a long tail at larger activities and a possible smaller peak in

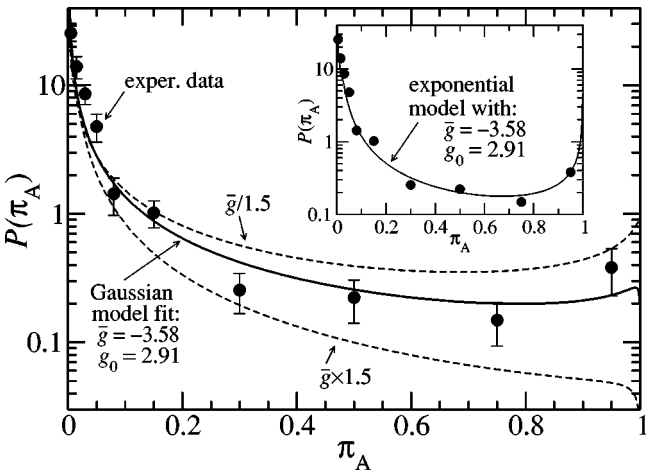

FIG. 4. Analysis of an experimental distribution of activities. Experimental data are derived from Ref. [16]. Error bars give the one standard deviation statistical uncertainty. The solid line is a two parameter fit $(\bar{g}, g_0)$ to the model of Gaussian energy distribution. The dashed lines correspond to the same g_0 and modified $\bar{g}$'s, which enables us to estimate the uncertainty on $\bar{g}$. Inset: comparison of the data to the model of exponential energy distribution (g_0 and $\bar{g}$ are not fitted again but taken from the Gaussian model fit).

$\pi_A \simeq 1$. This nontrivial shape is well fitted by the Gaussian model of Eq. (4) with $\bar{g} = -3.6$ and $g_0 = 2.9$ (the uncertainty on these parameters is about 50%, see dashed lines in Fig. 4). One infers $\overline{\delta G} \simeq -2.1$ kcal/mol and $\delta G_0 \simeq 1.7$ kcal/mol ($T = 300$ K). The order of magnitude of these values is compatible with thermodynamic measurements performed on similar systems [17–20]. This confirms the plausibility of the proposed approach. The inset of Fig. 4 shows the population distribution in the *exponential* model with $\bar{g}$ and g_0 values taken from the *Gaussian* fit. The agreement with the experimental data is also quite good. Thus, the proposed approach soundly does not strongly depend on the yet unknown shape details of the energy distribution. Finally, one can estimate the broad character of the activity distribution from the statistical analysis of the experimental data. Indeed, according, e.g., to the Gaussian model fit, the typical, most probable, population π_A is found to be $\simeq 6 \times 10^{-6}$ while the mean population is $\simeq 0.15$. Thus, the activity distribution spans more than four orders of magnitude.

VI. COARSE GRAINING DESCRIPTION: ALL OR NONE FEATURES

The variation of activity of a biopolymer upon mutation is often described as an ''all or none'' process: mutations are considered either as neutral (the mutant retains fully its activity and $\pi_A \simeq 100\%$) or as lethal (the mutant loses completely its activity and $\pi_A \simeq 0\%$). Satisfactorily, a coarse graining description of the proposed statistical models exhibits such all or none regimes for appropriate $(\bar{g}, g_0)$ values, as well as other regimes.

To obtain a quantitative coarse graining description, we define the mutants with ''no'' activity as those with population that has less than 12% $[\simeq \pi_A(\delta G_M = -2RT)]$ in the A shape. Their weight is

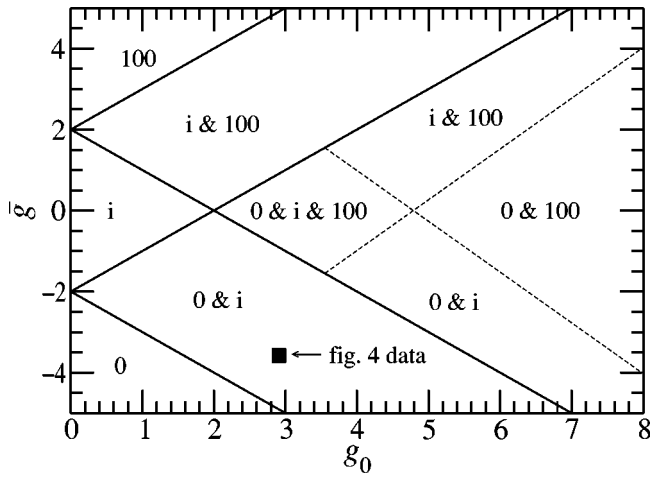

FIG. 5. Coarse graining features of the population distribution $P_G(\pi_A)$ in the Gaussian model. In each region, the population ranges dominating the distribution have been indicated (0 for $\pi_A \leqslant 12\%$, *i* for $12\% \leqslant \pi_A \leqslant 88\%$, and 100 for $\pi_A \geqslant 88\%$).

$$w_0 = \int_0^{12\%} P_{e \text{ or } G}(\pi_A) d\pi_A = \int_{-\infty}^{-2RT} P_{e \text{ or } G}(\delta G) d\delta G. \tag{8}$$

Similarly, the mutants with ''full,'' respectively, ''intermediate,'' activity are defined as those with $\pi_A \geqslant 88\%$, respectively $12\% \leqslant \pi_A \leqslant 88\%$, and their weight is $w_{100} = \int_{2RT}^{\infty} P_{e \text{ or } G}(\delta G) d\delta G$, respectively, $w_i = \int_{-2RT}^{+2RT} P_{e \text{ or } G}(\delta G) d\delta G$. Taking for definiteness the Gaussian model leads to

$$w_0 = \Phi\left(-\frac{2}{g_0} - \frac{\overline{g}}{g_0} \right), \tag{9}$$

where $\Phi(u) = \int_{-\infty}^{u} e^{-t^2/2} dt/\sqrt{2\pi}$ is the distribution function of the normal distribution. Similarly, one has $w_i = \Phi[(2 - \overline{g})/g_0] - \Phi[-(2+\overline{g})/g_0]$ and $w_{100} = 1 - \Phi[(2-\overline{g})/g_0]$. Approximate expressions for $\Phi(u)$ [$\Phi(u) \simeq -e^{-u^2/2}/(\sqrt{2\pi}u)$ for $u \ll -1$, $\Phi(u) \simeq 1/2 + u/\sqrt{2\pi}$ for $|u| \ll 1$ and $\Phi(u) \simeq 1 - e^{-u^2/2}/(\sqrt{2\pi}u)$ for $u \gg 1$] give the regimes in which each weight w is negligible ($w \ll 1$), dominant ($1 - w \ll 1$), or in between. For instance, w_0 is negligible for $\overline{g} > g_0 - 2$, dominant for $\overline{g} < -g_0 - 2$, and intermediate for $-g_0 - 2 < \overline{g} < g_0 - 2$. These inequalities indicate the transition from one regime to another. To be strictly in one regime typically requires that $\overline{g}/g_0$ is larger or greater than 1 from the corresponding criterion, e.g., w_0 is strictly negligible when $\overline{g}/g_0 > 1 + (g_0 - 2)/g_0$. The transitions from one regime to another one are in general exponentially fast (solid lines in Fig. 5). However, in the region ($g_0 > 2$, $|\overline{g}| < g_0 - 2$), the transitions from one regime to another one are smooth (dashed lines in Fig. 5) since, in this region, the weights vary slowly, e.g., $w_i \simeq 4/(g_0\sqrt{2\pi})$.

The resulting coarse graining classification of $P_G(\pi_A)$ is represented in Fig. 5. The ''all or none'' behavior, denoted ''0 & 100,'' appears in the region $g_0 \gtrsim 6/\sqrt{\pi/2}$ and $|\overline{g}| \lesssim \sqrt{\pi/2} g_0 - 6$ as the result of a large dispersion of energy differences associated to a moderate average energy difference. We note that all possible types of distributions are actually present in this model: probabilities concentrated at small, intermediate or large values (0, *i*, or 100); probabilities spread over both small and intermediate (0 & *i*), both small and large (0 & 100, all or none) or both intermediate and large (*i* & 100) values; probabilities spread over small, intermediate and large values at the same time (0 & *i* & 100). The coarse graining classification of Fig. 5 complements the number of peaks classification of Fig. 3 without overlapping it. Indeed, there exist parameters g_0 and $\overline{g}$ for which, e.g., two peaks coexist but one of these peaks has a negligible weight. Thus the presence of a peak is not automatically associated to a large weight in the region of this peak.

VII. ZOOMING IN THE $\pi_A \simeq 0$ PEAK: LONG TAILS

To go beyond the coarse graining description, we zoom in the $\pi_A \simeq 0$ peak. As shown in the inset of Fig. 2, the small activities, labeled as ''no activity'' in a coarse graining description, actually consist of nonzero activities with values scanning several orders of magnitude. This can be analyzed quantitatively, e.g., in the Gaussian model. For $\pi_A \simeq 0$, the activity distribution given by Eq. (4) is quasi-lognormal:

$$P_G(\pi_A) \simeq \frac{1}{\sqrt{2\pi} g_0 \pi_A} \exp\left[\frac{-(\ln \pi_A - \overline{g})^2}{2 g_0^2} \right]. \tag{10}$$

Thus, $P_G(\pi_A)$ has as a power-law-like behavior [15,21]

$$P_G(\pi_A) \simeq \frac{1}{\sqrt{2\pi} g_0 \pi_A} \quad \text{for} \quad e^{\overline{g} - \sqrt{2} g_0} \lesssim \pi_A \lesssim e^{\overline{g} + \sqrt{2} g_0}, \tag{11}$$

in the vicinity of the lognormal median $e^{\overline{g}}$. This corresponds to an extremely long tailed distribution, since $1/\pi_A$ is not even normalizable. It presents the peculiarity that, for *a* and $a+1$ belonging to $[\overline{g} - \sqrt{2} g_0, \overline{g} + \sqrt{2} g_0]$, the probability to obtain a population π_A of a given order of magnitude *a*, i.e., $\pi_A \in [e^a, e^{a+1}]$, does not depend on the considered ordered of magnitude *a*, since

$$\int_{e^a}^{e^{a+1}} P_G(\pi_A) d\pi_A \simeq \text{const.} \tag{12}$$

Thus, if a living organism has to adapt the chemical activity of one of its biopolymer constituents, it can explore several order of magnitude of activity by only few mutations within the biopolymer. The activity changes mimic a Lévy flight [22] as revealed, e.g., by the experimental data in Ref. [6]. The large activity changes will raise self-averaging issues [15] that will add up to those generated by correlations along evolutionary paths [23].

Three broadness regimes corresponding to three evolutionary regimes can be distinguished. If g_0 is very large, the mutant activities span a very large range. This regime might be globally lethal because, in most cases, the mutant activity will be either too low or too large to be biologically useful.

However, under conditions of intense stress, the large variability might allow the system to evolve radically. With $g_0 = 10$, for instance, the activity range covers typically 12 orders of magnitude from $10^{-6}e^{\overline{g}}$ to $10^{6}e^{\overline{g}}$ [see Eq. (11)]. If g_0 is moderately large, the mutant activities span just a few orders of magnitude. This regime is broad enough to permit significant changes, but not too broad to avoid producing too many lethal changes. With $g_0 = 3$, for instance, the activity range covers typically 3–4 orders of magnitude from $10^{-1.8}e^{\overline{g}}$ to $10^{1.8}e^{\overline{g}}$. If g_0 is small, the lognormal distribution peak can be approximated by a Gaussian [15]

$$P_G(\pi_A) \simeq \frac{1}{\sqrt{2\pi} g_0 e^{\overline{g}}} \exp\left[\frac{-(\pi_A - e^{\overline{g}})^2}{2(g_0 e^{\overline{g}})^2}\right]. \qquad (13)$$

The distribution is now narrow and the ranges of values is typically $[e^{\overline{g}}(1-2g_0), e^{\overline{g}}(1+2g_0)]$. This type of distribution is not adapted for producing large changes, but rather for performing fine tuning optimization. With $g_0 = 0.1$, for instance, the activity range covers only $\pm 20\%$ around $e^{\overline{g}}$.

We remark that the group-I ribozyme which we have analyzed corresponds to $g_0 \simeq 2.9$, right in the critical regime of moderately large g_0. One can guess from experimental studies of other biopolymers or from chemical considerations that most biopolymers will fall in this range since δG_0 is typically on the order of a few kilocalories while RT is $\simeq 0.6$ kcal. (Note that δG_0 corresponds to the free energy change between the biopolymer native 3D state and an unfolded state, in which the biopolymer has lost its three-dimensional shape but not its full secondary structure.) It would be interesting to perform further statistical data analysis to see how, e.g., the available protein mutagenesis studies fit with our present model.

The energy statistics associated mutations is likely to be determined at gross scale by the basic biophysics of the molecules involved. This fixes a range for δG_0. It is nonetheless plausible and suggested by our discussion that there is an evolutionary preferred type of activity distribution, and hence of sequences, that may imply a fine tuning of $g_0 = \delta G_0 / RT$ within the constraints on δG_0 coming from biophysics (see Fig. 6) so that each mutation typically generates a significant, but not systematically lethal, activity change. If one considers that the activity changes must cover between, say, one and seven orders of magnitude, then the allowed g_0 range is 1–6 [see Eq. (11)].

To answer the question whether the energy statistics is solely dictated by molecular biophysics or whether it is also influenced by evolutionary requirements, one may compare the energy statistics of molecules from different thermal environments. The conservation of the δG_0 range across psychrophilic and thermophilic molecules would stress the domination of biophysics factors. Note that our model would then imply different stochastic evolutionary dynamics, through the width of the activity distribution, for psychrophilic and thermophilic environments. Conversely, the conservation of g_0 would reveal the importance of evolutionary requirements.

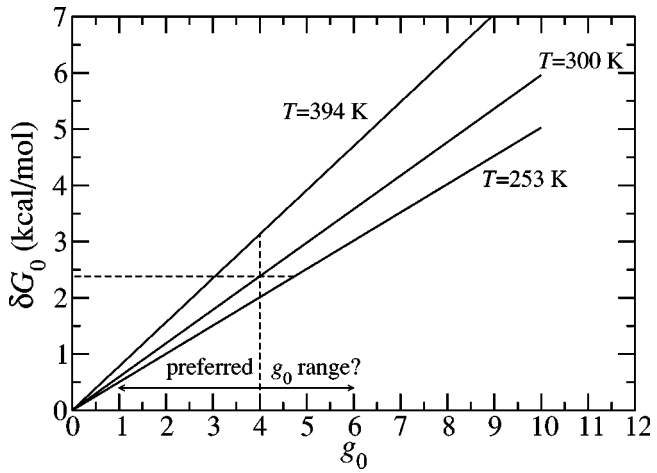

FIG. 6. Free energy dispersion δG_0 as a function of g_0 for different temperatures T. The upper and lower temperature limits for life are, respectively, $\simeq 121$ °C (394 K) and $\simeq -20$ °C (253 K). Depending on whether the energy statistics is determined by the biophysics or by evolutionary requirements, the range of either δG_0 or of g_0 is fixed (see, for example, the dashed lines). Evolutionary requirements suggest $1 \lesssim g_0 \lesssim 6$.

VIII. CONCLUSIONS

In this paper, we have presented a model for the distribution of biopolymer activities resulting from mutations of a given sequence. The model is characterized by the statistics of the energy differences between active conformations and inactive conformations. A similar model would be obtained by considering the statistics of activation energies. The model fits the measured activity distribution of a ribozyme with energy parameters in the physically appropriate range. It is also able to reproduce commonly observed behaviors such as all or none.

Importantly, the peak of small activities exhibits three distinct types depending on the broadness of the distribution of energy differences. Real biopolymers are in a critical regime allowing the exploration of different ranges of activities in a few mutations without being too often lethal. This critical regime seems the most favorable evolutionary regime and could be the statistical engine allowing molecular evolution. Thus the present work supports the idea that, for evolution to take place, the temperature and the physicochemistry dictating the free energy scales of biopolymers must obey a certain ratio. At last, it suggests that, by looking at small variations of this ratio, one might be able to classify biopolymers. One expects, for instance, that biopolymer sequences that are locked in a shape with a specific function, will have smaller g_0 than rapidly evolving biopolymers sequences that could acquire new functions by undergoing major structural changes. Thus, at the origin of life or during rapidly evolving punctuations, biopolymers with larger g_0 than those characterizing highly optimized, modern RNA and protein molecules, could have contributed to the emergence of novel phenotypes, leading thus to an increase of complexity.

[1] W. Fontana, BioEssays **24**, 1164 (2002).
[2] M. Kimura, Nature (London) **217**, 624 (1968).
[3] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1993).
[4] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker, Proc. R. Soc. London, Ser. B **255**, 279 (1994).
[5] A.S. Perelson and G.F. Oster, J. Theor. Biol. **81**, 645 (1979).
[6] E.A. Schultes and D.P. Bartel, Science **289**, 448 (2000).
[7] W. Fontana and P. Schuster, Science **280**, 1451 (1998).
[8] L.W. Ancel and W. Fontana, J. Exp. Zool. **288**, 242 (2000).
[9] *Lévy Flights and Related Topics in Physics*, edited by M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, Vol. 450 of *Lecture Notes in Physics* (Springer-Verlag, Berlin, 1995).
[10] *Anomalous Diffusion*: *From Basics to Applications*, edited by R. Kutner, A. Pękalski, and K. Sznajd-Weron, Proceedings of the XIth Max Born Symposium Held at Lądek Zdrój, Poland, 1998 (Springer-Verlag, Berlin, 1999).
[11] F. Bardou, J.-P. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling* (Cambridge University Press, Cambridge, 2002).
[12] N. Elredge and S. J. Gould, in *Models in Paleobiology*, edited by T. J. M. Schopf (Freeman Cooper & Co, San Francisco, 1972), pp. 82–115.
[13] B. Doliwa and A. Heuer, Phys. Rev. E **67**, 031506 (2003).
[14] V. da Costa, M. Romeo, and F. Bardou, J. Magn. Magn. Mater. **258–259**, 90 (2002).
[15] M. Romeo, V. da Costa, and F. Bardou, Eur. Phys. J. B **32**, 513 (2003).
[16] S. Couture, A.D. Ellington, A.S. Gerber, J.M. Cherry, J.A. Doudna, R. Green, M. Hanna, U. Pace, J. Rajagopal, and J.W. Szostak, J. Mol. Biol. **215**, 345 (1990).
[17] L. Jaeger, E. Westhof, and F. Michel, J. Mol. Biol. **234**, 331 (1993).
[18] L. Jaeger, F. Michel, and E. Westhof, J. Mol. Biol. **236**, 1271 (1994).
[19] P. Brion and E. Westhof, Annu. Rev. Biophys. Biomol. Struct. **26**, 113 (1997).
[20] P. Brion, F. Michel, R. Schroeder, and E. Westhof, Nucleic Acids Res. **27**, 2494 (1999).
[21] E.W. Montroll and M.F. Shlesinger, J. Stat. Phys. **32**, 209 (1983).
[22] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).
[23] U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, Phys. Rev. Lett. **89**, 208101 (2002).
[24] J.N. Onuchic, Z. LutheySchulten, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[25] In our model, the inactive state may be replaced by an ensemble of inactive states with a given energy [24].
[26] With Gaussian distributions of δG_M and E_a, one can be more specific. Both π_A and k are then lognormally distributed at small values. Thus, the product $a = k\pi_A$ is also lognormally distributed [15].