Simple stochastic model for the evolution of protein lengths

C. Destri and C. Miccio
Dipartimento di Fisica G. Occhialini, Università di Milano-Bicocca and INFN, Sezione di Milano, Piazza della Scienza 3, I-20126 Milano, Italy
(Received 18 April 2007; published 30 July 2007)

We analyze a simple discrete-time stochastic process for the theoretical modeling of the evolution of protein lengths. At every step of the process, a new protein is produced as a modification of one of the proteins already existing, and its length is assumed to be a random variable that depends only on the length of the originating protein. Thus a random recursive tree is produced over the natural numbers. If (quasi) scale invariance is assumed, the length distribution in a single history tends to a log-normal form with a specific signature of the deviations from exact Gaussianity. Comparison with the very large Similarity Matrix of Proteins database shows good agreement.

DOI: 10.1103/PhysRevE.76.011924 PACS number(s): 87.10.+e, 02.50.−r

I. INTRODUCTION

Nowadays, it is well established that the great variety of proteins in biological systems have been produced during the course of evolution by means of gene mutations that take effect at the coding level [1–6]. The main mechanisms are duplication of genome segments that contain sequences coding for one or more protein domains [7–10]; divergence of the duplicated sequences by insertion, deletion, and substitution of one or more base pairs [11–14]; domain rearrangements, such as gene fusions and gene fission [15,16], domain recombination [17,18], gene shuffling (recombination between dissimilar genes) [19], and domain insertions and deletions [20,21]. By means of these microscopic mechanisms, iterated a huge number of times throughout the ages of evolution, an initial protein population, most likely very small and poorly assorted, has been enormously increased to the present very large number and complex variety.

A valuable framework for the effective modeling of the evolution of genes and proteins could be provided by stochastic processes. In the most general formulation, one should take into account the complex organization of biological systems into independent organisms grouped in turn into species, genera, and kingdoms, as well as the complicated effects of natural selection. However, since all evolution mechanisms generate new biological material by means of modifications of the biological material already existing, in the case of proteins we may imagine a simpler, more abstract discrete-time stochastic process over the space of all amino acid sequences, such that at each time step $t=1,2,3,\ldots$ a new protein is generated with some prescribed random mechanism from the set of proteins already existing. Clearly, the discrete time of the model has nothing to do with the time of the true biological evolution process, except that it is (almost) a monotonically increasing function of the latter, at least on time scales large enough (great mass extinctions correspond most likely to periods when this monotonicity is lost). Moreover, a single time step in the process would correspond to some averaging over a multitude of different effects, both at the microscopic or biochemical level, and at the macroscopic level of the selection-based evolution mechanisms.

This abstract stochastic process can be specified by $\Pr(p_{t+1}|\{p\}_t)$, that is, the conditional probability that the $t+1$ protein has the amino acid sequence p_{t+1}, given that there is already a set $\{p\}_t \equiv \{p_1, p_2, \ldots, p_t\}$ of distinct proteins at time t. In principle, $\Pr(p_{t+1}|\{p\}_t)$ might embody the effects of many, if not all, of the complicated biochemical and evolutionary mechanisms alluded to above, and should depend explicitly on time. Moreover, the initial configuration of proteins can be assumed to coincide with the actual set of distinct amino acid sequences present in nature at some moment in the distant past, when their total number was much smaller than the present one. In any case, a huge amount of information is required for a complete specification of a model endowed with detailed predictive power. On the other hand, all that we can hope to reproduce in reasonably simple terms, to a large extent independent of the details of the model, are some broad characteristics of the distribution of proteins currently observed. This is indeed our main working hypothesis, based on the fact that the very large number of proteins in the Similarity Matrix of Proteins (SIMAP) database do show simple universal properties [24,25]. Then we may rely on the basic universality property, typical of a wide class of stochastic processes, which is the ability to forget the details of the initial transient regime and to relax toward a statistical equilibrium or quasiequilibrium state that depends only on very general features of the conditional probability $\Pr(p_{t+1}|\{p\}_t)$, and is characterized by few, weakly time-dependent "macroscopic" parameters.

II. A STOCHASTIC PROCESS FOR PROTEIN LENGTHS

In the present work, we concentrate our attention on the distribution of protein lengths, that is, the observed frequency of proteins with a specific number of amino acids over the set of all known proteins. Thus we can consider only the protein length as the random variable of the stochastic process. By definition, this random length takes values in the natural numbers, and we denote it with the symbol ℓ. We also observe that by construction all the proteins produced in the process can be ordered according to the time of production, starting from $t=N_0$, with N_0 the number of distinct proteins in the initial configuration, and arriving at $t=N$, with N

of the order of the number of distinct proteins that exist now in nature, which is of order 10^7 or more. The statistical dynamics of the process is fully determined by $\Pr(\ell_{t+1}|\ell_t,\ell_{t-1},\ldots,\ell_1)$, that is, the conditional probability that the $t+1$ protein has length ℓ_{t+1}, given that the preceding proteins have the indicated lengths. This conditional probability might depend explicitly on the formal time t.

As already discussed above in a more general context, the detailed biological mechanisms that constrain $\Pr(\ell_{t+1}|\ell_t,\ell_{t-1},\ldots,\ell_1)$ are far too complex to be explicitly incorporated in the model. Therefore, we shall make simple and workable assumptions about the conditional probability, relying, in practice, on some sort of central limit theorem for the probability that a protein taken at random from the state of a very long random process has a certain length.

As a first simplifying assumption on the conditional probability $\Pr(\ell_{t+1}|\ell_t,\ell_{t-1},\ldots,\ell_0)$, we make that of *locality*. That is, we assume that a given protein length can be produced from a preceding length independently of all the other lengths already produced. Hence we can write

$$\Pr(\ell_{t+1}|\ell_t,\ell_{t-1},\ldots,\ell_1)=\sum_{s=1}^{t} q_s W_s(\ell_{t+1}|\ell_s), \tag{1}$$

where the nonnegative weights q_s are properly normalized, $\Sigma_{s=1}^{t} q_s=1$, and $W_s(\ell|\ell')$ can be interpreted as Markovian transition probabilities. In the absence of any other information, one would assume equal *a priori* probabilities among the different proteins, that is, $q_s=1/t$ and $W_s(\ell|\ell')=W(\ell|\ell')$, with no explicit s dependence. This might appear in conflict, however, with the global changes of ecosystems, as well as with the complex organization of biological systems in kingdoms and species (which suggests that all proteins existing nowadays can be roughly divided into subsets of similar proteins having almost independent evolutionary histories, as least not too far in the past). We may take this into account by restricting the predictions of our stochastic process to suitably chosen subsets of the proteins of the SIMAP database, according, for instance, to given kingdoms. Moreover, we can neglect, on average, the global changes of ecosystems by placing the start of the process not too deep in the past. All together, let us assume that

$$\Pr(\ell_{t+1}|\ell_t,\ell_{t-1},\ldots,\ell_1)=\frac{1}{t}\sum_{s=1}^{t} W(\ell_{t+1}|\ell_s). \tag{2}$$

Our stochastic process now differs from a random walk on the natural integers only because at each step any one of the already existing lengths, rather than only the last generated one, may serve as the starting point for a jump to a new length. We are therefore dealing with the so-called random recursive tree (RRT) [2,3] (more precisely, a random recursive forest) embedded by $W(\ell|\ell')$ into the natural integers. It follows that the probability $P(\ell,t)$ that the tth protein has exactly length ℓ satisfies the non-Markovian evolution equation

$$P(\ell,t+1)=\sum_{\ell'=1}^{\infty} W(\ell|\ell')Q(\ell',t), \tag{3}$$

where $Q(\ell,t)$ is the average length distribution (that is, the mean fraction of proteins of length ℓ) at time t, and therefore evolves as

$$(t+1)Q(\ell,t+1)=tQ(\ell,t)+P(\ell,t+1). \tag{4}$$

Together, Eqs. (3) and (4) define a stochastic process with memory, and should be compared to the Markov chain recursion for a simple random walk (RW) on all possible lengths, which would read instead

$$P^{\mathrm{RW}}(\ell,t+1)=\sum_{\ell'=1}^{\infty} W(\ell|\ell')P^{\mathrm{RW}}(\ell',t) \tag{5}$$

without any memory of the past. We still do have to make some choice of the explicit form of $W(\ell|\ell')$, in which case our stochastic process could be quite easily simulated on a computer. We expect, however, that the large-time asymptotic regime of the process depends only on very general features of the functional form of $W(\ell|\ell')$, again thanks to the universality hypothesis, which has its roots in the law of large numbers. In any case, to investigate the statistics of the lengths produced, we need beforehand some useful properties and formal manipulations valid for any $W(\ell|\ell')$.

We recall first of all that, by definition, the nonnegative numbers $W(\ell|\ell')$ satisfy the normalization condition

$$\sum_{\ell=1}^{\infty} W(\ell|\ell')=1.$$

These transition probabilities are the elements of a matrix W, the so-called stochastic matrix in the case of Markov processes. Without loss of generality, we may take W to be ergodic, that is, such that any finite length can be produced after a suitable number of steps starting from any other finite length.

Next we can exploit the linearity of Eq. (3) to simplify the choice of initial conditions for $P(\ell,t)$. As already stated above, the process is assumed to start with N_0 distinct proteins, which we may take to have n distinct lengths ℓ_j, $j=1,2,\ldots,n$, each repeated n_j times, so that $\Sigma_j n_j=N_0$. This defines the initial length distribution

$$Q(\ell,N_0)=\frac{1}{N_0}\sum_{j=1}^{n} n_j\delta_{\ell\ell_j}$$

when N_0 was the total number of distinct proteins. As the process may start from any one of these initial proteins with equal probability $1/N_0$, we may regard n_j/N_0 as the probability that the process starts exactly from the length ℓ_j. Therefore the solution of Eq. (3) can be written

$$P(\ell,t) = \frac{1}{N_0}\sum_{j=1}^{n} n_j P(\ell,t-N_0+1|\ell_j)$$

$$= \sum_{\ell'=1}^{\infty} P(\ell,t-N_0+1|\ell')Q(\ell',N_0), \tag{6}$$

where $P(\ell,t|\ell')$ is the special solution that is concentrated at the arbitrary value ℓ' at $t=1$, that is, $P(\ell,1|\ell')=\delta_{\ell\ell'}$. Similarly, we have

$$Q(\ell,t) = \sum_{\ell'=1}^{\infty} Q(\ell,t-N_0+1|\ell')Q(\ell',N_0), \tag{7}$$

where $Q(\ell,t|\ell')$ is the solution of Eq. (4) specialized to $P(\ell,t|\ell')$, that is,

$$Q(\ell,t|\ell') = \frac{1}{t}\sum_{s=1}^{t} P(\ell,s|\ell'). \tag{8}$$

Clearly, ℓ' is the length of a specific protein which plays the role of the root for the RRT, while the complete process is a forest of RRTs each having a root in one of the N_0 initial proteins and growing in parallel. That is, the unique protein labeled by t has a fixed probability n_j/N_0 of belonging to the tree rooted in a protein of length ℓ_j.

We may now introduce the matrix notation

$$W(\ell|\ell') \equiv [W]_{\ell\ell'},$$

$$P(\ell,t|\ell') \equiv [P(t)]_{\ell\ell'},$$

$$Q(\ell,t|\ell') \equiv [Q(t)]_{\ell\ell'},$$

which allows us to write the evolution equation for $P(t)$ more compactly as

$$P(t+1) = WQ(t) = \frac{1}{t}\sum_{s=1}^{t} WP(s); \tag{9}$$

or equivalently as

$$tP(t+1) = \sum_{s=1}^{t-1} WP(s) + WP(t) = (t+W-1)P(t), \tag{10}$$

which has the formal solution

$$P(t) = \prod_{s=1}^{t-1}\left(1+\frac{W-1}{s}\right) = \frac{W^{\overline{t-1}}}{(t-1)!}, \tag{11}$$

where $z^{\bar{n}}$ stands for the so-called *raising factorial product* $z(z+1)\cdots(z+n-1)$ [22]. The raising factorial generates the (unsigned) Stirling numbers of the first kind as coefficients of its expansion in simple powers of z:

$$z^{\bar{n}} = \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} z^k, \quad n>0,$$

where we adopted the square bracket notation of Ref. [22] for the Stirling numbers. Hence, from Eq. (11), we can write

$$P(t) = \frac{1}{(t-1)!}\sum_{s=1}^{t-1} \begin{bmatrix} t-1 \\ s \end{bmatrix} P^{\mathrm{RW}}(s), \tag{12}$$

where $P^{\mathrm{RW}}(t)=W^t$ is the formal solution of the standard random walk. Notice that, by Eq. (12), $P(t)$ is indeed properly normalized, that is,

$$\sum_{\ell=1}^{\infty} P(\ell,t|\ell') = 1,$$

since W^s is also a stochastic matrix and Stirling numbers satisfy the normalization

$$\sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} = n!, \quad n>0.$$

In fact, Eq. (12) shows that the quantity

$$p_s(t) = \frac{1}{(t-1)!}\begin{bmatrix} t-1 \\ s \end{bmatrix}$$

has the interpretation, for the abstract RRT, of the probability that the node added at time t is at a chemical distance s from the root of the tree, that is, the original node present at $t=1$. In terms of proteins, $p_s(t-N_0+1)n_j/N_0$ is therefore the probability that the tth protein is obtained through s changes from one of the N_0 initial proteins of length ℓ_j.

Notice also that the evolution equation (9) allows one write an alternative expression for $Q(t)$ that is local in time (but generally nonlocal in "space") with respect to $P(t)$,

$$Q(t) = W^{-1}P(t+1) = \frac{1}{t!}\sum_{s=1}^{t} \begin{bmatrix} t \\ s \end{bmatrix} W^{s-1}. \tag{13}$$

One can see that $p_{s+1}(t+1)$, which by construction satisfies

$$p_{s+1}(t+1) = \frac{1}{t}\sum_{k=1}^{t} p_s(k),$$

represents the average fraction of nodes at a distance s from the root [2].

For very large n we can use the approximation

$$z^{\bar{n}} \simeq \frac{\Gamma(n)}{\Gamma(z)} n^z \left[1+O\left(\frac{1}{n}\right)\right], \tag{14}$$

which follows from Euler's infinite product representation of the Γ function [23]. From Eqs. (11), (13), and (14) we then find

$$Q(t) \simeq \frac{\exp[(W-1)\log t]}{\Gamma(W+1)}, \tag{15}$$

where we neglected all inverse powers of t in the exponent, relying for uniformity on the boundedness of its spectrum of W. The crucial point of Eq. (15) is the very slow logarithmic dependence on time, which appears evident upon comparison with the formal solution $W^t=\exp(t\log W)$ of the Markovian case.

In order to provide more explicit expressions for $P(t)$ and $Q(t)$, we need some special assumption on the stochastic matrix W. We do that in the next section.

III. AVERAGE PROPERTIES OF SCALE-INVARIANT MODELS

We describe here a class of examples that can be treated in detail at the analytic level. These are characterized by the assumption that our stochastic process is (almost) scale invariant. Intuitively, one expects that longer proteins can be changed throughout evolution more easily than shorter ones. Exact scale invariance would mean that changes are proportional to length.

To implement this picture, we first extend the lengths ℓ from the positive integers to all real positive numbers. This is a common extension in studying stochastic dynamical evolutions of biopolymer lengths (e.g., microtubules [4] and α helices in proteins [5]). It will become apparent in the following that, in fact, it has very little impact on our conclusions.

Next we change variables, from ℓ to its logarithm $x = \log \ell$ and assume homogeneity in x, namely, that

$$W(\ell|\ell')d\ell = W(e^x|e^{x'})d(e^x) = \mathcal{W}(x-x')dx$$

is translation invariant, i.e., a function only of x-space differences. The process is very simple: at each time step the random walker may pick any one of the previously visited points as starting point for the next step, whose value x is extracted with the one-step probability distribution function (PDF) $\mathcal{W}(x)$. In terms of protein lengths, at each step the length is rescaled by a factor e^x. Since the true variables are discrete, we may take $\mathcal{W}(x)$ to be very smooth for all x. Likewise, since $\ell \geqslant 1$, we may take $\mathcal{W}(x)$ to vanish very quickly (let us say "faster than any power") for $x \to -\infty$. For $x \to +\infty$, we assume instead quite reasonably that $\mathcal{W}(x)$ vanishes fast enough to have finite moments at least up to order 4. We then introduce the following notation for the first two cumulants:

$$\mu = \int dx x \mathcal{W}(x), \quad \sigma^2 = \int dx (x-\mu)^2 \mathcal{W}(x),$$

that is, the mean value and the squared fluctuations.

We can now define the process probability in x space as

$$\mathcal{P}(x-x',t) \equiv e^x P(e^x, t|e^{x'}),$$

and in the same way we can introduce the average distribution $\mathcal{Q}(x,t)$, which by Eq. (8) satisfies

$$\mathcal{Q}(x,t) = \frac{1}{t}\sum_{s=1}^{t} \mathcal{P}(x,s).$$

Since the stochastic matrix that corresponds to $\mathcal{W}(x-x')$ is diagonal in Fourier space, we can now write the formal expression Eq. (11) as

$$\mathcal{P}(x,t) = \int \frac{dk}{2\pi} e^{ikx} \hat{\mathcal{P}}(k,t), \tag{16}$$

where

$$\hat{\mathcal{P}}(k,t) = \prod_{s=1}^{t-1} \left(1 + \frac{\hat{\mathcal{W}}(k)-1}{s}\right)$$

and $\hat{\mathcal{W}}(k)$ is the Fourier transform of $\mathcal{W}(x)$. Clearly, by Eq. (13), the Fourier transform of $\mathcal{Q}(x,t)$ reads

$$\hat{\mathcal{Q}}(k,t) = \frac{\hat{\mathcal{P}}(k,t+1)}{\hat{\mathcal{W}}(k)}.$$

The correct normalization of either $\mathcal{P}(x,t)$ or $\mathcal{Q}(x,t)$ follows from that of $W(x)$, which implies $\hat{\mathcal{W}}(0)=1$. Other consequences of the probabilistic nature of $W(x)$ are the symmetry $\hat{\mathcal{W}}(k)^* = \hat{\mathcal{W}}(-k)$ and the bound $|\hat{\mathcal{W}}(k)| \leqslant 1$. In addition, with the natural requirements made above on the one-step PDF $\mathcal{W}(x)$, the function $\hat{\mathcal{W}}(k)$ has the expansion near $k=0$

$$\hat{\mathcal{W}}(k) \simeq 1 - i\mu k - \frac{1}{2}(\mu^2+\sigma^2)k^2 + \cdots, \tag{17}$$

and vanishes for large $|k|$.

In this context of a continuous logarithmicspace, the extension to a generic initial distribution is very simple: it amounts to multipling both Fourier transforms $\hat{\mathcal{P}}(k,t)$ and $\hat{\mathcal{Q}}(k,t)$ by the Fourier transform of the initial distribution.

Using Eq. (15), we have for large t

$$\mathcal{P}(x,t) = \int \frac{dk}{2\pi} e^{ikx} \frac{\exp[(\hat{\mathcal{W}}(k)-1)\log t]}{\Gamma(\hat{\mathcal{W}}(k))} \tag{18}$$

and

$$\mathcal{Q}(x,t) = \int \frac{dk}{2\pi} e^{ikx} \frac{\exp[(\hat{\mathcal{W}}(k)-1)\log t]}{\Gamma(\hat{\mathcal{W}}(k)+1)} \tag{19}$$

up to fully negligible inverse power corrections in t. For any given $\hat{\mathcal{W}}(k)$, the Fourier integral in Eqs. (18) and (19) can be computed numerically to high accuracy through fast Fourier transform. Moreover, for large t we can derive very similar asymptotic expansions in inverse powers of $\log t$ valid for any $\hat{\mathcal{W}}(k)$ in the class described above. Since our main interest is in the average distribution profile $\mathcal{Q}(x,t)$, we concentrate our attention on this.

The leading term for large t is determined only by the first two terms of the $\hat{\mathcal{W}}(k)$ expansion (17) near $k=0$, with the quadratic term providing the Gaussian dominance in Eqs. (18) and (19) according to the central limit theorem. From the first and second derivatives in $k=0$ of the Fourier transform $\hat{\mathcal{Q}}(k,t)$, we first compute the mean value and standard deviation of the process for large t as

$$\bar{\mu}_t \equiv \langle x \rangle_t = \mu(\log t + \gamma - 1),$$

$$\bar{\sigma}_t^2 \equiv \langle (x - \mu_t)^2 \rangle_t = (\mu^2 + \sigma^2)(\log t + \gamma - 1) - \left(\frac{\pi^2}{6} - 1\right)\mu^2, \tag{20}$$

where $\gamma = 0.577\,215\,6\ldots$ is the Euler-Mascheroni constant. Then, in terms of the standard centered scaled variable

$$\xi = \xi(x,t) = \frac{x - \bar{\mu}_t}{\bar{\sigma}_t},$$

we have to leading order

$$\mathcal{Q}(x,t) \simeq \frac{e^{-\xi^2/2}}{\sqrt{2\pi\bar{\sigma}_t^2}}. \tag{21}$$

Going back to the length variable ℓ through the definition $x = \log(\ell/\ell')$, we find the log-normal distribution

$$Q(\ell,t|\ell') \simeq \frac{e^{-[\log \ell - \log(\ell' e^{\bar{\mu}_t})]^2/(2\bar{\sigma}_t^2)}}{\ell\sqrt{2\pi\bar{\sigma}_t^2}}, \tag{22}$$

peaked around the rescaled initial length $\ell' e^{\bar{\mu}_t - \bar{\sigma}_t^2}$.

Subleading contributions to the above results, of relative order $1/\sqrt{\log t}$ and smaller, *at fixed* values of ξ, can be computed by the standard perturbation technique around Gaussian integrals: one includes also terms of order higher than k^2, say of order $n \geqslant 4$, in the power expansion of $\hat{\mathcal{W}}(k)$ around $k=0$, and then expands to order k^n also the exponentials of such terms; for completeness, also the expansion of the inverse Γ function must be properly extended; finally, one integrates explicitly each term of the complete expansion in terms of multiple derivatives of the leading Gaussian. One obtains in this way an n-degree polynomial in ξ times the Gaussian $e^{-\xi^2/2}$. The $n+1$ coefficients of the polynomial are fixed by the first $n+1$ moments of the distribution, which in turn can be computed directly from the Taylor series in $k=0$ of the Fourier transform $\hat{\mathcal{Q}}(k,t)$ or $\hat{\mathcal{P}}(k,t)$ (by construction, we must impose $\langle \xi \rangle_t = 0$ and $\langle \xi^2 \rangle_t = 1$ for the first two moments). Taking into account the specific form of these Fourier transforms, it is more convenient to calculate the cumulants of $\mathcal{Q}(x,t)$ or $\mathcal{P}(x,t)$, since their n-order cumulant is given by $\log t$ times the n-order *moment* of the one-step PDF $\mathcal{W}(x)$, plus the n-order derivative with respec to k of $\log[\Gamma(\hat{\mathcal{W}}(k)+1)]$ or $\log[\Gamma(\hat{\mathcal{W}}(k))]$ evaluated at $k=0$. Moreover, the latter contributions are systematically subleading as compared to the moments of $\mathcal{W}(x)$, so that we have, for the third-order and the fourth-order cumulants of $\mathcal{Q}(x,t)$ (that is, the average skewness $\bar{s}_t$ and kurtosis $\bar{\kappa}_t$ of the process, up to normalization conventions)

$$\bar{s}_t \equiv \langle \xi^3 \rangle_t = \frac{\mu_3}{\mu_2^{3/2}} \frac{1}{\sqrt{\log t}} \left[1 + O\left(\frac{1}{\log t}\right)\right],$$

$$\bar{\kappa}_t \equiv \langle \xi^4 \rangle_t - 3 = \frac{\mu_4}{\mu_2^2} \frac{1}{\log t} \left[1 + O\left(\frac{1}{\log t}\right)\right], \tag{23}$$

where μ_3 and μ_4 are the third- and fourth-order moments of $\mathcal{W}(x)$, while $\mu_2 = \mu^2 + \sigma^2$ is the analogous notation for the second moment. In this expression, one may regard the expectation values as evaluated with $\mathcal{P}(x,t)$ rather than with $\mathcal{Q}(x,t)$, since the differences are due solely to the change $\Gamma(\hat{\mathcal{W}}(k)) \to \Gamma(\hat{\mathcal{W}}(k)+1)$ from Eq. (18) to Eq. (19), and are subleading. We can now recognize a distinctive mark of the RRT over the real line: for sufficiently large time, the kurtosis of the average distribution profile is certainly positive, since the fourth moment of any $\mathcal{W}(x)$ is positive definite. Another important characteristic, which will be further discussed later on, is the positivity of the ratio between the skewness $\langle \xi^3 \rangle_t$ and the third moment μ_3 of $\mathcal{W}(x)$.

The extension of the main results Eqs. (21)–(23) to the case of a generic initial distribution is straightforward. In particular, to the cumulants of $\mathcal{Q}(x,t)$ one would have to add the cumulants of the initial distribution, which are constant in time and therefore subleading. Thus Eqs. (20) would get additive constants and Eqs. (23) would stay unchanged. This is the standard way to see how the process forgets abouts the initial conditions (in a logarithmically slow way).

IV. PROFILE FLUCTUATIONS

Let us assume that, for a given stochastic matrix W and initial distribution $Q(\ell, N)$, we can explicitly compute $P(\ell,t)$ and $Q(\ell,t)$, at least for large t, as in the preceding section. To compare the result to the length distribution in a single evolution history, or very few of them, which is indeed our case, we need to gather information also on the fluctuations of the profile of the length distribution from one history to another.

Typically, one would like to rely on the law of large numbers. For ergodic Markov chains (with finitely many possible events), this law states that the probability that the frequency of a certain event in a given history differs from its equilibrium probability by any nonzero amount vanishes when the history becomes infinitely long. In our case, the elementary events are the observed protein lengths, and the frequency in a given history is just the profile of the length, distribution in a given evolution history. The quantity $Q(\ell,t)$ discussed above is just the expectation value of the profile, that is, its average over all possible histories. In a Markovian setup with finitely many possible events, there would be no difference between the profile of a specific history and its expectation value in the $t \to \infty$ limit, which means vanishing profile fluctuations in the limit and negligible ones for sufficiently large t. The stochastic process at hand, however, is not Markovian, having the (very specific, simple, and itself random) RRT form of memory, and it has a number of possible events that is in principle arbitrarily large. In this case, we expect, thanks to stronger forms of the law of large numbers like the central limit theorem (and have indeed verified in the example class of the preceding section), that the average length distribution $Q(\ell,t)$ assumes, for t large enough, a universal nonconstant form, which depends only on very general properties of W.

What we need then is also that the fluctuations of the frequency for large t do not completely spoil the profile of its expectation value $Q(\ell,t)$. Notice, for instance, that this is not true for random walks, not even when they are recurrent (as is generally true in one dimension, which is our case). In other words, in the standard RW the frequency of times the walker visits any given small region keeps fluctuating strongly from one very long history to the other, never resembling the mean frequency profile. This is due to the characteristic dispersion of order $\sqrt{t}$ of the RW, which implies that each elementary event occurs an insufficient number of times of order $1/\sqrt{t}$ to guarantee a good convergence of the frequency along a single history (it would be even worse in $d>1$ dimensions).

On the contrary, the random memory of the RRT dramatically helps the application of the law of large numbers, since the logarithmic time dependence leads to a much slower drift and diffusion, strongly reducing the impact of fluctuations on the length distribution. One could say that the length distribution is an almost "self-averaging" array of random variables, which for sufficiently long time does not differ too much from its expected value. Indeed, at least in the case of the abstract RRT, there exist mathematically rigorous theorems about the convergence of the chemical distance profile of any RRT toward a normal form [1]. In this section, we provide some quantitative numerical evidence of the same property for length distribution profiles using a specific model for $W(\ell|\ell')$.

We first revert to the realistic situation of lengths as positive integers not smaller than some lower cutoff $\ell_{\min}\geqslant 1$; next, we consider the following RRT process (written as computer pseudocode)

$$\ell = \text{integer part of } e^{x}\ell(n_t);$$

$$\text{if } \ell \geqslant \ell_{\min} \text{ then } \ell(t+1)=\ell, \tag{24}$$

where n_t is an integer chosen at random from 1 to t and x is extracted with the one-step PDF $\mathcal{W}(x)$ over the continuous logarithmicspace; finally, we pick for $\mathcal{W}(x)$ the maximum entropy form compatible with our general setup, namely, a Gaussian with mean μ and standard deviation σ. This minimum bias choice could even be regarded as natural in view of the many different "microscopic" and "macroscopic" mechanisms on which the stochastic process should depend, as discussed in the Introduction. However, we make it here mainly for numerical definiteness. In any case the analysis of the preceding section and the discussion below, at the end of this section, should make it clear that other choices of $\mathcal{W}(x)$ in the same class would lead to relative changes that vanish as $1/\log t$, while preserving important characteristic properties like the positivity of the kurtosis.

It is quite easy on modern personal computers to accurately simulate the process (24) by running many very long random histories. In our simulations, we produced 10^5 length distributions with the discrete time t running from $N_0\lesssim 50$ to $N=5\times 10^6$. For the sake of definiteness, we started from 25 initial lengths chosen at random from 30 to 50 and set $\mu=0.16$ and $\sigma=0.19$. This setup was determined in such a way as to fit the overall scales of the length distribution in the SIMAP database, as will be discussed in the next section. In particular, it turns out that the effects on the profiles of the lower cutoff $\ell_{\min}$ are fully negligible, so that the scale-invariant framework adopted in the preceding section should apply. Indeed, one can also check that the discreteness of the lengths $\ell(t)$ does not play any significant role at all with respect to the continuous case.

For each distribution we computed the mean and standard deviation in the variable $x=\log \ell$:

$$\mu_t = \frac{1}{t}\sum_{j=1}^{t} x(j), \quad \sigma_t^2 = \frac{1}{t}\sum_{j=1}^{t}[x(j)-\mu_t]^2, \tag{25}$$

at prescribed intermediate values of t. Likewise, we computed the skewness and kurtosis

$$s_t = \frac{1}{t}\sum_{j=1}^{t}\xi^3(j), \quad \kappa_t = -3+\frac{1}{t}\sum_{j=1}^{t}\xi^4(j), \tag{26}$$

where as usual $\xi(j)=[x(j)-\mu_t]/\sigma_t$.

These four parameters are still random variables which fluctuate from one RRT to the other. Moreover, except for the mean μ_t, their average values over all possible RRT realizations of t steps do not coincide with the corresponding parameters of the average distribution $\mathcal{Q}(x,t)$, since such average values receive contributions also from the profile fluctuations. Only the average $\langle\mu_t\rangle$ is given by the quantity $\bar{\mu}_t$ in the first equation (20). The differences between $\langle\sigma_t\rangle$, $\langle s_t\rangle$, $\langle\kappa_t\rangle$ and $\bar{\sigma}_t$, $\bar{s}_t$, $\bar{\kappa}_t$ in the second equation (20) and Eqs. (23), respectively, cannot even be estimated with the help of $\mathcal{Q}(x,t)$ alone. This is true *a fortiori* for the fluctuations. Therefore it is important to provide some (numerical) evidence of their behavior for large times. In particular, s_t and κ_t, provide a measure of the deviation from Gaussianity of a given profile (we refer to the above-mentioned mathematical literature for some rigorous bounds in the case of abstract RRTs).

We also kept track of all the logarithmicspace profiles, after a suitable coarse graining: we fixed beforehand a uniform binning grid of K intervals of width $h\ll 1$ over a portion of the real line large enough to contain almost all ξ points produced [e.g., the interval $(-5,5)$ to comprise all points within 5σ]; then we computed the fraction q_k of ξ points in a given RRT that fall in the kth interval of the grid. At this stage, using continuous or discrete lengths does make a difference, since a binning grid too fine over the logarithms of integer lengths will induce spurious fluctuations. Hence, in the discrete case, for each integer j repeated n_j times in a given length distribution, we filled the real interval $(j-1/2, j+1/2)$ with n_j double precision lengths chosen at random; only after this smoothing did we compute the distribution over the regularly space grid in logarithmicspace.

By construction, the average of the discretized density $q_k(t)/h$ over all possible histories will reproduce the integral of the average profile $\mathcal{Q}(x,t)$ as a function of ξ over the kth interval of the grid. Then an estimate of the profile fluctuations is the standard deviation of $q_k(t)/h$ for each k.

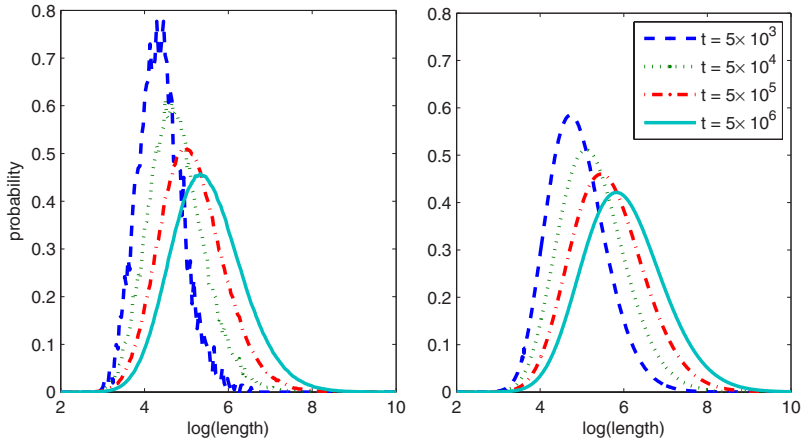

FIG. 1. (Color online) Evolution of the logarithmic space length distribution in a specific history (left panel) and on average (right panel), starting from the same initial conditions.

With $q_k(t)$ we computed another important (and more robust) measure of deviation from Gaussianity, that is, the entropy

$$S_t = \log h - \sum_{k=1}^{N} q_k(t) \log q_k(t).$$

In fact, in the $t \to \infty$ limit of an infinite RRT and then $h \to 0$ of vanishing grid width, a Gaussian profile for ξ would have maximal entropy equal to $(\log 2\pi + 1)/2 = 1.418\,938\,53\ldots$.

In Fig. 1, we show the evolution of the logarithmic length distribution along a single history, and, for comparison, the evolution of the average profile $\mathcal{Q}(x,t)$ obtained by numerically integrating through fast Fourier transform Eq. (19) and superposing the results as in Eq. (7).

In Fig. 2, we show the distributions of the statistical estimators defined above for few values of t equally spaced in logarithmicspace. We see that the parameters that measure deviation from Gaussianity, that is, s_t, κ_t, and S_t, have mean values that slowly tend to the Gaussian values with smaller and smaller fluctuations as $t \to \infty$. The convergence behavior is roughly the ubiquitous one, $1/\log t$, with variances that vanish faster than the peak movement. Also, the variance of the standard deviation seems to slowly converge. On the other hand, the fluctuations of the mean do not appear to converge at all; this is reflected in the reduction more slowly than $1/\log t$ of the standard deviation of the ξ profile fluctuations. In Table I we provide further numerical evidence through the standard deviations over the 10^5 sample histories of μ_t, σ_t, s_t, κ_t, and S_t.

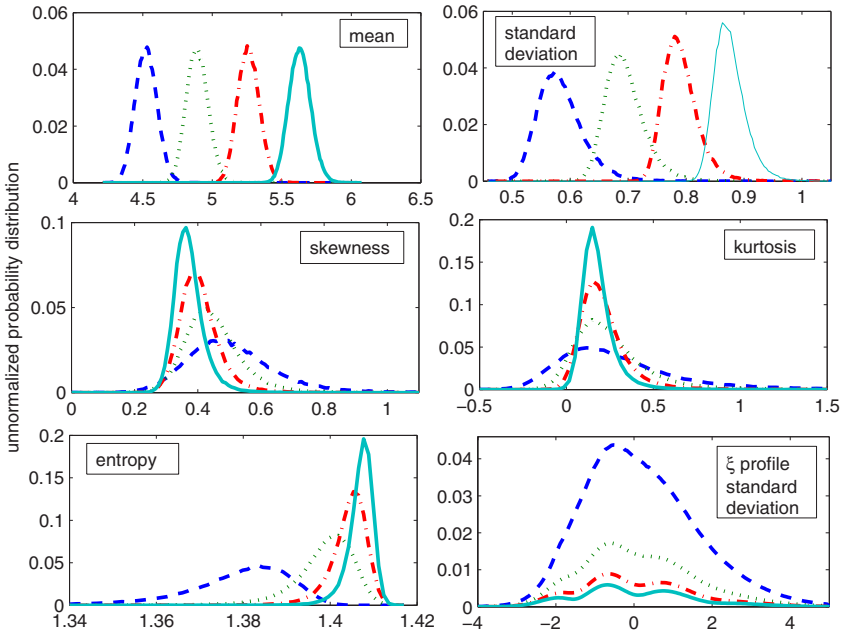

FIG. 2. (Color online) Distributions, as fraction of the total, over 10^5 sample histories of the indicated statistical estimators of the (natural) logarithmic space length profile for the same times as in Fig. 1.

TABLE I. Standard deviations over 10^5 samples of the statistical estimators of mean, standard deviation, skewness, kurtosis, and entropy of the logarithmic space length distribution.

t	$\Delta\mu_t$	$\Delta\sigma_t$	Δs_t	$\Delta\kappa_t$	ΔS_t
5×10^3	0.0781	0.0401	0.1425	0.3840	0.0131
5×10^4	0.0784	0.0343	0.0927	0.2282	0.0072
5×10^5	0.0787	0.0304	0.0647	0.1440	0.0045
5×10^6	0.0785	0.0277	0.0482	0.0980	0.0031

These results remain qualitatively unchanged under generalization from the Gaussian one-step PDF chosen above to a generic $\mathcal{W}(x)$ of the class discussed in the previous section. For fixed μ and σ, only t-independent numerical variations appear due to the change in the higher moments of $\mathcal{W}(x)$. In particular, the skewness can be made to assume prevalently positive or negative values by choosing a $\mathcal{W}(x)$ with positive or negative third moment (while keeping the first moment $\mu>0$), while the kurtosis distribution remain always peaked around positive values, with a variance that appears to vanish faster than the mean. This is in agreement with the properties of the average length distribution as given by Eq. (23).

In summary, very large RRTs over the space of possible protein lengths are indeed almost auto averaging objects, and it is sensible to compare the average properties of the random process to a few, or even a single, realizations of it.

V. COMPARISON WITH THE OBSERVED LENGTH DISTRIBUTIONS

To test our simple model we compare here the predicted length distributions with the real length distributions of proteins observed in nature. In this last decade the number of known protein sequences has been rapidly growing and is still growing now at a steady pace. A huge number of protein sequences coming from very many different species are now stored in various databases.

In particular, the SIMAP [24] database collects almost all amino acid sequences from public databases and completely sequenced genomes. On September 2006 it was storing more than 5.5 million nonredundant proteins coming from more than 100 000 different species.

We report in Table II a coarse subdivision of all SIMAP proteins and their corresponding species in five (nonstandard) main kingdoms: bacteria, viruses, plants, invertebrates (animalia), and vertebrates (animalia). In Fig. 3, we provide plots of the corresponding length distributions.

One can see that all SIMAP distribution profiles have a globally similar shape, with a well-defined overall position and scale. There are, however, also large fluctuations on smaller scales. In particular, the curves of viruses, invertebrates, and vertebrates show very high and narrow peaks in correspondence to certain specific values of length. Of course, on general grounds, our model is too simple and generic to make predictions on other than global properties of the profiles, so we should restrict ourselves to the lowest moments or cumulants of the distribution, and perform robust coarse graining on the data for more refined analysis. We believe, in any case, that these peaks are to a large extent spurious, being due to an over-representation in the SIMAP database of those particular protein lengths. SIMAP, in fact, contains a lot of proteins that do not necessarily come from completely sequenced genomes: this fact makes the collection of proteins nonhomogeneous over the species present in the database, and so it is possible that certain peculiar lengths are more represented, since they correspond to proteins of many more different species than other lengths. If the collection were homogeneous over the species, we would expect length distributions without high narrow peaks, and also less fluctuating in general. At any rate, we verified that the global analysis reported below is almost insensitive to the removal by hand of the high and narrow peaks.

TABLE II. Number of species and proteins for each kingdom in SIMAP in September 2006.

Kingdom	Number of species	Number of proteins
Bacteria	111.30	2217301
Viruses	14631	319885
Plants	31232	1156929
Animalia		
Invertebrates	25951	383760
Vertebrates	19341	772605
Environmental samples	1453	32591
Synthetic	822	14660

The SIMAP database provides a very large sample of real proteins, which can be assumed to be statistically significant. We believe, therefore, that it is sensible as a testbed for our model, and we make the basic assumption that the SIMAP length distributions for different kingdoms as (almost) independent realizations of our stochastic process. The motivation is that different kingdoms have been going through almost independent evolutionary histories for a long time, and, even if one cannot forget that far enough in the past there was no distinction at all, the main characteristic of the stochastic process of forgetting the initial conditions suggests that at most a negligible trace remains of the common remote past in each kingdom distribution.

In Table III we list the measured values of the mean, standard deviation (SD), skewness, kurtosis, and entropy of the logarithmic length distributions for the five kingdoms separately, and for the cumulative all-kingdom distribution. Except for the entropy, these parameters can be computed directly from the statistical estimators as in Eqs. (25) and (26) without any coarse graining. To compute the entropy, we performed a coarse graining in the logarithmic space with the same procedure as in the preceding section. One can see that the kurtosis is always positive, in accordance with the average property of the model [Eq. (23)] and with the results of the previous section on the fluctuations. We also notice that the cumulative kurtosis is definitely higher than the individual ones, due to the fluctuations in the lower cumulants. Again, this is consistent with the interpretation of the king-

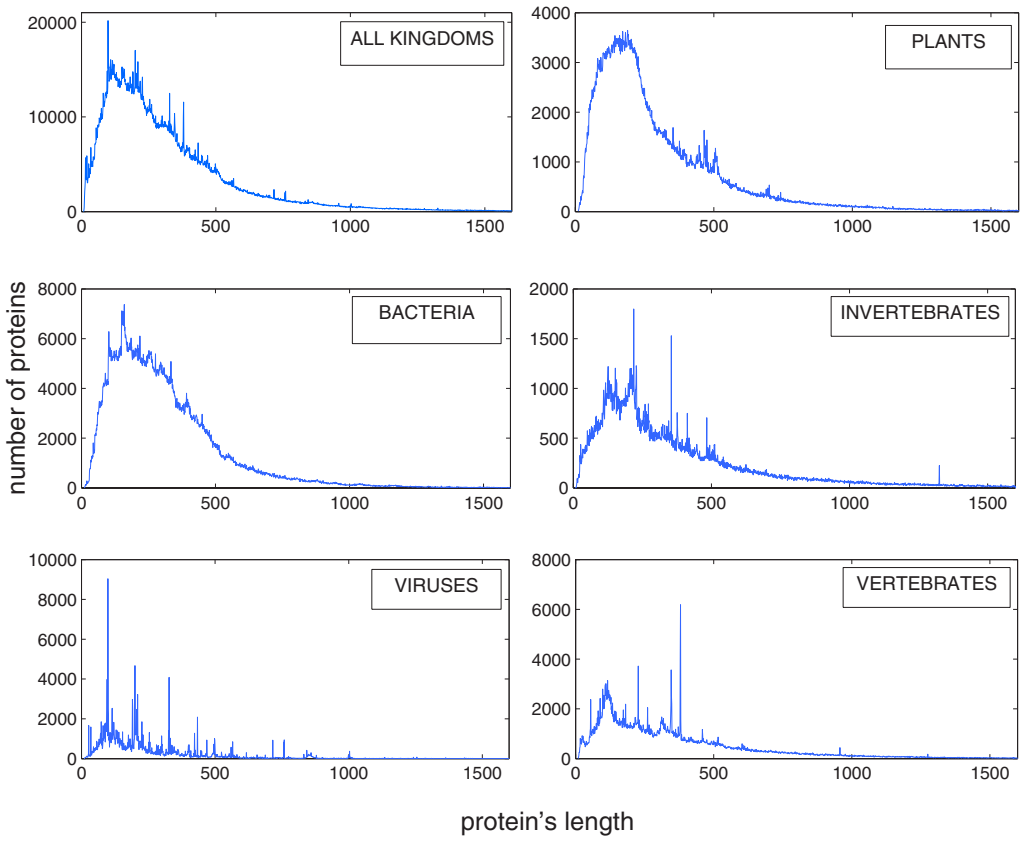

FIG. 3. (Color online) Length distributions of SIMAP proteins. Each box shows an enlargement of the (unnormalized) length distributions of proteins coming from all kingdoms ($\langle l\rangle$=335, l_{max}=38 031), bacteria ($\langle l\rangle$=316.9, l_{max}=36 805), viruses ($\langle l\rangle$=273.9, l_{max}=7312), plants ($\langle l\rangle$=314.5, l_{max}=20 925), invertebrates ($\langle l\rangle$=416.1, l_{max}=23 015), and vertebrates ($\langle l\rangle$=397.1, l_{max}=38 031).

dom distributions as independent realizations of the process. Except for the viruses, the entropy is always close to the upper Gaussian limit, with the plant distribution the closest to a normal form.

In Fig. 4 we show the Gaussian fits of the length distributions. As expected from the data in Table III, these fits appear rather good, apart from fluctuations, which are more important when the entropy is lower, that is, for viruses and vertebrates. Explaining these fluctuations is beyond the scope of our model. Moreover, one must remember that the SIMAP database is incomplete, and, as discussed above, probably biased toward particular species; these features contribute to local irregularities.

With fine-tuned choices of the two main parameters of $\mathcal{W}(x)$, μ and σ, a specific large value of t, and values of the initial lengths in the range 30–50, one can produce simulations with Eq. (24), like those reported in Fig. 2, whose distribution profiles fit well the peak positions and sizes of the SIMAP length distributions.

Our choice for the initial distribution is based on the quite natural assumption that today's proteins are evolved from shorter peptide ancestors [6,27]. In any case, according to the model, t-independent changes in the initial distribution might affect the final distribution only by terms of relative order $1/\log t$.

We remark also that, in our purely probabilistic framework, no fine quantitative check can be performed, for some good reasons.

First, even assuming that for each kingdom the proteins in the database constitute a statistically significant fraction of

TABLE III. Global statistical indicators of the SIMAP length distributions in logarithmic space.

Kingdom	Mean	SD	Skewness	Kurtosis	Entropy
Bacteria	5.53	0.68	−0.20	0.32	1.408
Viruses	5.26	0.79	0.30	0.53	1.297
Plants	5.44	0.78	−0.01	0.04	1.414
Invertebrates	5.65	0.87	−0.03	0.31	1.409
Vertebrates	5.60	0.89	−0.18	0.25	1.394
All kingdoms	5.49	0.81	−0.26	0.63	1.406

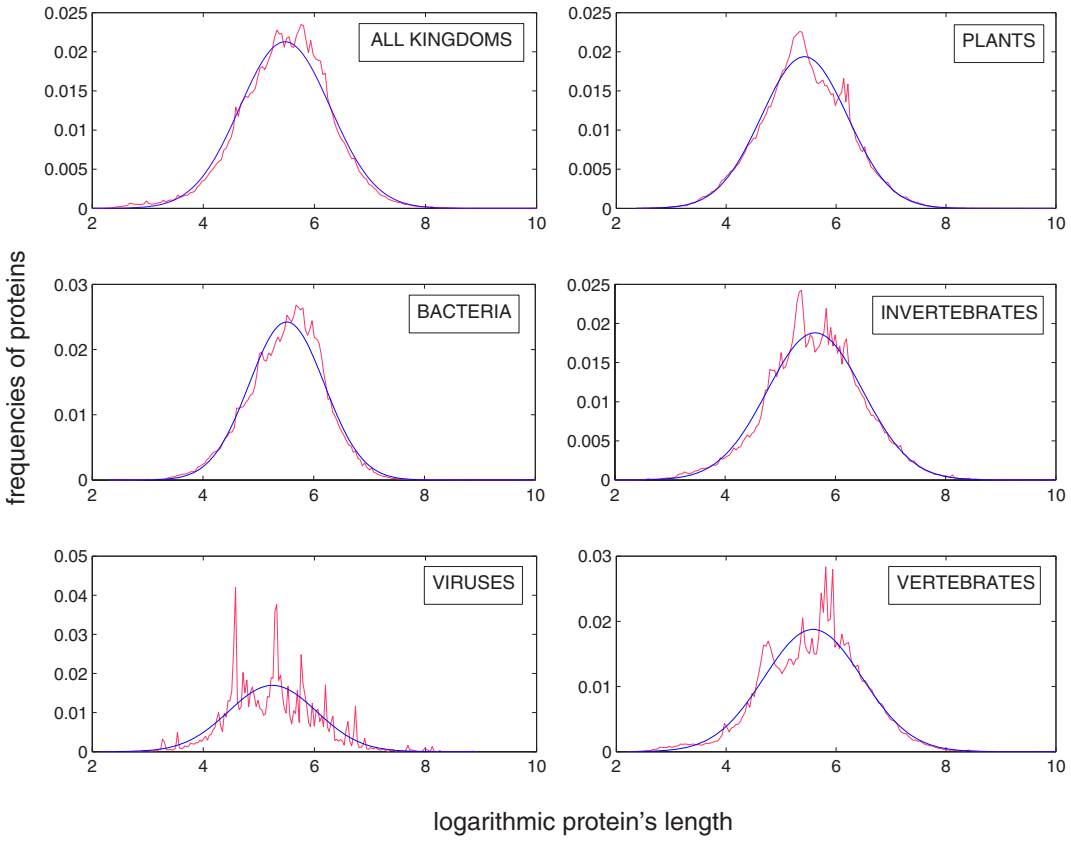

FIG. 4. (Color online) Length distributions of SIMAP proteins in the (natural) logarithmic length space. These (unnormalized) distributions have been obtained through a uniform gridding with 200 intervals over $x=\ln \ell$. Before the change of variable from ℓ to x, we scattered the protein length values from integer to real to avoid the introduction of spurious fluctuations in the logarithmic space.

the total existing in nature, we do not know what the total number might actually be. So we cannot fix the precise value of the total discrete time of the stochastic process. This does not lead to large uncertainties, though, since the evolution is only logarithmic in this discrete time.

Second, the one-step PDF $\mathcal{W}(x)$ governing the model can hardly have any precise quantitative relation with the multitude of microscopic and macroscopic effects that drive the evolution. So one could not ascribe any particular value to a specific functional form of $\mathcal{W}(x)$, whose most relevant free parameters were determined from data fitting. Rather, we must restrict ourselves to very general properties valid for a wide class of one-step PDF's. This argument applies also to the lowest moments of $\mathcal{W}(x)$, which might very well differ in distinct independent processes.

Let us examine therefore in more detail to what extent the model agrees with the observed distributions.

First of all, according to the model, the length distribution of a large set of proteins belonging to a *single* long evolutionary history must be almost a Gaussian in logarithmic space, that is, log normal over the lengths. As we have just seen, this agrees quite well with the SIMAP distributions. The approximate log normality of protein sizes was observed several years ago on much smaller data sets [26].

Next, there are the two scales of the fluctuations of the two parameters of a Gaussian, namely, the mean and the standard deviation, which were denoted as μ_t and σ_t in the previous section. Once a process simulation is fine tuned to produce the correct average values of μ_t and σ_t, their fluctuations have a scale which depends mildly on the higher moments of $\mathcal{W}(x)$, does not depend on t in the case of μ_t, and depends at most as $1/\log t$ in the case of σ_t. These scales agree fairly well with those observed over the SIMAP kingdom distributions (see Tables I and III), at least when $\mathcal{W}(x)$ has skewness and kurtosis not too large. Notice that also the lower cutoff $\ell_{\min}$ on the possible lengths acts as a constraint on the higher cumulants of $\mathcal{W}(x)$. For instance, an excessively negative skewness in $\mathcal{W}(x)$ would typically lead to large left tails also in the length distributions which are then abruptly cut off at $\ell_{\min}$; such abrupt cuts are absent in the observed data, as evident from Fig. 4.

Then there are the systematic deviations from Gaussianity. The model predicts a positive kurtosis for any $\mathcal{W}(x)$, and the SIMAP data agree very well with this. Also, the entropy is very close to the expected values, except for the viruses, whose protein distribution is the least abundant, and has the smallest mean length and the largest relative fluctuations. However, the data in Table III also always show negative skewness, again except for the viruses. This characteristic cannot be accounted for too easily in the model.

Indeed, as we have already noted, it is natural to assume that the average protein length has been growing in time. This requires that the first moment μ of the one-step PDF $\mathcal{W}(x)$ is positive, that is, that positive shifts in x space prevail

over negative ones. Now, from the average relation (23) and the numerical analysis reported in the preceding section, a prevalently negative skewness requires that the third moment μ_3 of $\mathcal{W}(x)$ is negative. The simplest Gaussian one-step PDF used to produce the results in Fig. 2 has by definition μ and μ_3 with the same sign, but other PDF's with positive μ and negative μ_3 are certainly possible. However, this would mean even more negative skewness in $\mathcal{W}(x)$, causing problems with the lower cutoff $\ell_{\min}$, unless very large unrealistic values of the initial lengths are assumed, which in turn would typically spoil the overall scale fitting. Thus, after all, there is tension on the model in spite of its many free parameters.

Our simple model for the protein length distributions is based on RRT embedded in the natural numbers, with the assumption of almost scale invariance for the transition probability $W(\ell|\ell')$. By this, we mean that the stochastic process is uniform in continuous logarithmic space, with translation invariance broken only by length discreteness and the lower cutoff $\ell_{\min}$, in both cases with negligible effects. This is an idealization suggested by simplicity (it translates the intuitive idea that longer proteins can change more than shorter ones throughout evolution) and ease of analytic investigation of average properties. It has some difficulties in accounting for negative skewness (viruses apart), but it is in overall good agreement with observations, especially for the positiveness of the kurtosis. This suggests keeping to a minimum the modifications in more realistic models for $W(\ell|\ell')$. We describe one minimal change in the next section.

VI. INTRINSIC SMOOTH CUTOFF ON LARGE LENGTHS

In the stochastic framework we have considered up to now, the vanishing of the length distribution for large ℓ is determined by the slowly drifting and diffusing character of the process, with the assumption of relatively small initial lengths.

On the other hand, there are reasons to expect that very long proteins are intrinsically less probable than shorter ones, in the sense that the "microscopic" mechanisms that determine, upon countless repetitions, the production of longer and longer proteins are eventually limited by simple stability and biosynthesis criteria: very long proteins, to be stable against thermal fluctuations in the natural environment, must fold in the biologically active form more "tightly" than shorter ones, as could be measured by their growing spectral dimension [28]; but this requires more and more complex stereoscopic orderings while the building blocks (amino acids at the lowest level and larger structures at the second and third levels) are limited in number and typology, thus decreasing the unfolding stability and/or increasing the misfolding probability [29]; moreover, there exists a higher biosynthetic cost for longer proteins that might determine an evolutionary constraint on the expression of long proteins [30]. We could therefore expect some form of smooth cutoff on long lengths, parametrized by a stability scale ℓ_s.

The minimal change on the model, as anticipated above, could therefore be the following:

$$W(\ell|\ell') = \ell^{-1} g(\ell/\ell_s) \mathcal{W}(\log(\ell/\ell')) \tag{27}$$

where $\mathcal{W}(x)$ is the usual one-step PDF in logarith space and $g(u)$ a smooth function which is almost constant for $u \lesssim 1$ and monotonically decreases to zero for large u.

The random recursion corresponding to Eq. (28) is a simple modification of Eq. (24):

$$\ell = \text{interger part of } e^x \ell(n_t);$$

$$\text{if } \ell \geqslant \ell_{\min} \text{ and } r < g(\ell/\ell_s) \text{ then } \ell(t+1) = \ell, \tag{28}$$

where, as before, n_t is a random integer from 1 to t, and x is extracted from $\mathcal{W}(x)$, while the new random number r is extracted uniformly in the interval $(0, g_{\max})$, with $g_{\max} = g(\ell_{\min}/\ell_s)$ the assumed largest value of the function g.

Another possibility could be to introduce an explicit ℓ dependence in $\mathcal{W}(x)$, in such a way that length reductions $(x<0)$ become more probabale than length growths $(x>0)$ for large enough lengths. In this case, the random recursion would be the same as Eq. (24); the only change is that x is extracted in a weakly non-scale-invariant way from a one-step PDF $\mathcal{W}(x;\ell)$.

Once some specific form for $\mathcal{W}(x)$ and $g(u)$, or for $\mathcal{W}(x;\ell)$, is chosen, simulations with weakly broken scale invariance can be performed as easily as before. Since there are now more tunable parameters, it is almost obvious that data fitting can be improved. From a purely quantitative point of view, these better fits have little significance. Instead, we want to stress the main new qualitative aspects: the smooth cutoff typically induces shorter right tails in the simulated distribution, thus slightly reducing both skewness and kurtosis. If the one-step PDF has the right characteristics, it is possible to obtain almost always length distributions with still positive kurtosis but negative skewness after a few million steps. The cutoff prevents the formation of proteins that are too long, thus allowing us to reproduce the observed mean length and length variance. Typically, ℓ_s, which by construction provides the scale of the rightmost tail, needs to be chosen between 5000 and 10 000, depending on other details of the model.

It is also interesting to observe that the positive skewness of the length distribution of viruses does not constitute a problem for the above scenario, since the overall size of this distribution is smaller than the others, and might very well be too small to feel the effects of the smooth cutoff on higher lengths.

It should also be noticed that the upper cutoff significantly reduces the variance of the length distribution, allowing an easier fit of the observed distributions with t of the order of 10^7. This value indeed fixes $1/\log t$ of the order of the absolute values of the non-Gaussianity indicators, and of the fluctuations of μ_t and σ_t observed in the SIMAP data. On the contrary, a straightforward application of Eq. (20), which ignores both fluctuations and the cutoff, would typically lead to much larger (and most likely unrealistic) values of t.

VII. CONCLUSIONS AND OUTLOOK

In this work, we have described a simple stochastic framework for the theoretical modeling of the evolution of protein lengths. It is based on the idea of recursive random trees over the set of natural numbers. RRTs represent the

simplest formal implementation of the main feature of the evolutionary process: new biological material is produced through modifications of the biological material already existing. In the case of proteins, the full space over which the RRT grows is that of all amino acid sequences, but it can be reduced to more tractable spaces when only specific observables are considered, as is the case of the protein lengths.

Of course, the details of the stochastic process, as encoded in the conditional probabilities, are out of reach in practice, due to the multitudes of natural causes ranging from biochemical interactions to selection mechanisms in varying environments. The relevance of the stochastic framework is based therefore on the concept of universality; namely, that, under the law of large numbers, statistical coarse-grained observations tend to take universal forms that depend only on a few fundamental features of the stochastic process. In the case at hand, the main features are the autoaveraging property of RRTs and the approximate scale invariance of the one-step transition probability; they imply the universal properties that protein length distributions are almost log normal, with positive kurtosis, and a specific scale for the overall deviations from exact Gaussianity.

There are several routes for improvements. First of all, the choice of RRTs (which have a uniform probability over all nodes of the tree for the attachment of the new node) is in itself an ideal simplification. In a more realistic setup, one should consider differently weighted nodes in order to mimic certain aspects of the evolutionary process, such as selection and differentiation. Then there are many more observables other than the distribution length in protein databases such as SIMAP. The global statistical analysis of the SIMAP protein homology network carried out in Ref. [25] shows several interesting features, which deserve to be studied within some generalization of the stochastic process described here.

ACKNOWLEDGMENT

We are thankful to Thomas Rattei for his kind permission to access the SIMAP database.

[1] M. Drmota and H.-K. Hwang, SIAM J. Discrete Math. **19**, 19 (2005); Adv. Appl. Probab. **37**, 321 (2005).
[2] A. Meir and J. W. Moon, Can. J. Math. **30**, 997 (1978).
[3] P. Bialas, Z. Burda, J. Jurkiewicz, and A. Krzywicki, Phys. Rev. E **67**, 066106 (2003).
[4] H. Flyvbjerg, T. E. Holy, and S. Leibler, Phys. Rev. Lett. **73**, 2372 (1994).
[5] J. Ferkinghoff-Borg, M. H. Jensen, J. Mathiesen, P. Olesen, and K. Sneppen, Phys. Rev. Lett. **91**, 266103 (2003).
[6] J. Soding and A. N. Lupas, BioEssays **25**, 837 (2003).
[7] T. Ohta, Genome **31**, 304 (1989).
[8] K. H. Wolfe and D. C. Shields, Nature (London) **387**, 708 (1997).
[9] T. J. Vision, D. G. Brown, and S. D. Tanksley, Science **290**, 2114 (2000).
[10] A. McLysaght, K. Hokamp, and K. H. Wolfe, Nat. Genet. **31**, 200 (2002).
[11] M. Lynch and J. S. Conery, Science **290**, 1151 (2000).
[12] G. C. Conant and A. Wagner, Genome Res. **13**, 2052 (2003).
[13] S. Pascarella and P. Argos, J. Mol. Biol. **224**, 461 (1992).
[14] S. A. Benner, M. A. Cohen, and G. H. Gonnet, J. Mol. Biol. **229**, 1065 (1993).
[15] B. Snel, P. Bork, and M. Huynen, Trends Genet. **16**, 9 (2000).
[16] B. Snel, P. Bork, and M. Huynen, Proc. Natl. Acad. Sci. U.S.A. **99**, 5890 (2002).
[17] W. F. Doolittle, Nature (London) **272**, 581 (1978).
[18] C. A. Voigt, C. Martinez, Z-G. Wang, S. L. Mayo, and F. H. Arnold, Nat. Struct. Biol. **9**, 553 (2002).
[19] G. C. Conant and A. Wagner, Adv. Genome Biol. **6**, R50 (2005).
[20] R. Aroul-Selvam, T. Hubbard, and R. Sasidharan, J. Mol. Biol. **338**, 633 (2004).
[21] J. Weiner, F. Beaussart, and E. Bornberg-Bauer, FEBS J. **273**, 2037 (2006).
[22] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science* (Addison-Wesley, Reading, MA, 1989).
[23] *Handbook of Mathematical Functions: With Formulas, Graphs and Mathemathical Tables*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972).
[24] R. Arnold, T. Rattei, P. Tischler, M. Truong, V. Stümpflen, and W. Mewes, Bioinformatics **21**, ii42 (2005); http://mip_sgfs.de/genre/proj/simap
[25] C. Miccio and T. Rattei, eprint- arXiv:q-bio.QM/0703053.
[26] S. S. Sommer and J. E. Cohen, J. Mol. Evol. **15**, 37 (1980).
[27] A. N. Lupas, C. P. Ponting, and R. B. Russell, J. Struct. Biol. **134**, 191 (2001).
[28] R. Burioni, D. Cassi, F. Cecconi, and A. Vulpiani, Proteins **55**, 529 (2004).
[29] U. Bastolla and L. Demetrius, Protein Engineering Design Selection **18**, 405 (2005).
[30] J. Warringer and A. Blomberg, BMC Evol. Biol. **6**, 61 (2006).