Genetic Correlations in Mutation Processes

E. Ben-Naim and A. S. Lapedes

Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos NM 87545

We study the role of phylogenetic trees on correlations in mutation processes. Generally, correlations decay exponentially with the generation number. We find that two distinct regimes of behavior exist. For mutation rates smaller than a critical rate, the underlying tree morphology is almost irrelevant, while mutation rates higher than this critical rate lead to strong tree-dependent correlations. We show analytically that identical critical behavior underlies all multiple point correlations. This behavior generally characterizes branching processes undergoing mutation.

PACS numbers: 87.10+e, 87.15.Cc, 02.50.-r, 87.23.Kg

I. INTRODUCTION

Biological evolution is influenced by a number of processes including population growth, mutation, extinction, and interaction with the environment, to name a few [1]. Genetic sequences are strongly affected by such processes and thus provide an important clue to their nature. The ongoing effort of reconstructing evolution histories given the incomplete set of mapped sequences constitutes much of our current understanding of biological evolution.

However, this challenge is extraordinary as it involves an inverse problem with an enormous number of degrees of freedom. Statistical methods such as maximum likelihood techniques coupled with simplifying assumptions on the nature of the evolution process are typically used to infer the structure of the underlying evolutionary tree, *i.e.*, the phylogeny [2–5].

Genetic sequences such as RNA/DNA or amino acid sequences can be seen as words with letters taken from an alphabet of 4 or 20 symbols, respectively. Generally, there are nontrivial intra-sequence correlations that influence the evolution of the entire sequence. Additionally, the structure of the evolutionary tree plays a role in this process as one generally expects that the closer sequences are on this tree, the more correlated they are [6]. In this study, we are interested in describing the influence of the latter aspect, namely the phylogeny, on the evolution of sequences. Specifically, we examine correlations between sequences, thereby complementing related studies on changes in fluctuations and entropy due to the phylogeny [7–9]. To this end, we consider particularly simple sequences and focus on a model that mimics the competition between the fundamental processes of mutation and duplication.

The rest of this paper is organized as follows. In Sec. II, the model is introduced, and the main result is demonstrated using the pair correlations. Correlations of arbitrary order are obtained and analyzed asymptotically in Sec. III. To examine the range of validity of the results, generalizations to stochastic tree morphologies and sequences with larger alphabets are briefly discussed in Sections IV and V. Sec. VI discusses implications for multiple site correlations in sequences with independently evolving sites. We conclude with a summary and a discussion in Sec. VII.

II. PAIR CORRELATIONS

Let us formulate the model first. The sequences are taken to be of unit length, i.e., single symbol, and the corresponding alphabet consists of two letters. The numeric values $\sigma = \pm 1$ are conveniently assigned to these letters. We will focus on binary trees where the number of children equals two. This structure is deterministic in that both the number of children and the generation lifetime are fixed. Nevertheless, the results apply qualitatively to stochastic tree morphologies as well. Finally, the mutation process is implemented as follows: with probability $1-p$ a child equals his predecessor while with probability p a mutation occurs, as illustrated in Fig. 1. The mutation process is invariant under the transformation $\sigma \to -\sigma$ and $p \to 1-p$, and we restrict our attention to the case $0 \le p \le 1/2$ without loss of generality.

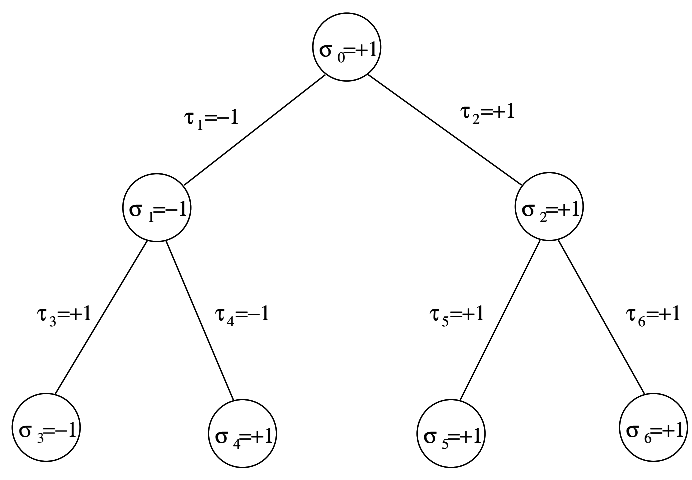

Fig.1 The mutation process on a two-generation binary tree. The multiplicative variable τ indicates whether a mutation occurred.

A natural question is how correlated are the various leafs (or nodes) of a tree in a given generation (or equivalently, time)? Consider $G_2(k)$ the average correlation between two nodes at the kth generation

$$G_2(k) = \langle\langle \sigma_i \sigma_j \rangle\rangle. \tag{1}$$

The first average should be taken over all realizations for a fixed pair of nodes $i \neq j$, while the second average is taken over all different pairs belonging to the same generation. The above average gives equal weights to all pairs and thus it corresponds to a uniformly chosen sample. For example, consider this quantity at the second generation (see Fig. 1), $G_2(2) = [\langle\sigma_3\sigma_4\rangle + \langle\sigma_3\sigma_5\rangle + \langle\sigma_3\sigma_6\rangle]/3$. One index ($i = 3$) may be fixed since all nodes in a given generation are equivalent.

To evaluate averages, it is useful to assign a multiplicative random variable $\tau_i = \pm 1$ to every branch of the tree such that $\sigma_i = \sigma_j \tau_i$ with j the predecessor of i. One has $\tau_i = 1$ (-1) with probability $1-p$ (p), and consequently,

$$\langle\tau\rangle \equiv \langle\tau_i\rangle = 1 - 2p. \tag{2}$$

Pair correlations are readily calculated using the τ variables: writing $\sigma_3 = \sigma_0\tau_1\tau_3$ and similarly for σ_4 gives $\langle\sigma_3\sigma_4\rangle = \langle\sigma_0\tau_1\tau_3\sigma_0\tau_1\tau_4\rangle = \langle\sigma_0^2\tau_1^2\tau_3\tau_4\rangle$. Since $\sigma_i^2 = \tau_i^2 = 1$, this correlation simplifies, $\langle\sigma_3\sigma_4\rangle = \langle\tau_3\tau_4\rangle$. Furthermore, mutation processes on different branches are independent and consequently $\langle\tau_i\tau_j\rangle = \langle\tau_i\rangle\langle\tau_j\rangle$ when $i \neq j$. Thus, $\langle\sigma_3\sigma_4\rangle = \langle\tau\rangle^2$ and similarly $\langle\sigma_3\sigma_5\rangle = \langle\sigma_3\sigma_6\rangle = \langle\tau\rangle^4$. The overall picture becomes clear: when calculating two-point correlations, the path to the tree root is traced for each node. As $\tau^2 = 1$, doubly counted branches cancel. Only branches that trace the path to the first common ancestor are relevant. In other words,

$$\langle\sigma_i\sigma_j\rangle = \langle\tau\rangle^{d_{i,j}} \tag{3}$$

with $d_{i,j}$ the "genetic distance" between two points, the minimal number of branches that connect two nodes. Indeed, at the second generation $d_{3,4} = 2$, $d_{3,5} = d_{3,6} = 4$ and consequently $G_2(2) = (\alpha^2 + 2\alpha^4)/3$ with the shorthand notation $\alpha = \langle\tau\rangle = 1 - 2p$. This generalizes into a geometric series $G_2(k) = (\alpha^2 + 2\alpha^4 + \cdots + 2^{k-1}\alpha^{2k})/(2^k - 1)$. Evaluating this sum gives the pair correlation

$$G_2(k) = \frac{\alpha^2}{2\alpha^2 - 1}\frac{(2\alpha^2)^k - 1}{2^k - 1}. \tag{4}$$

Interestingly, pair correlations are not affected by the initial state, *i.e.*, the value of the tree root.

For sufficiently large generation numbers, the leading order of the pair correlation decays exponentially with the generation number. However, different constants characterize this decay, depending on the mutation probability

$$G_2(k) \simeq \begin{cases} \frac{\alpha^2}{2\alpha^2-1}\,\alpha^{2k} & p < p_c; \\ \frac{\alpha^2}{1-2\alpha^2}\,2^{-k} & p > p_c. \end{cases} \tag{5}$$

As seen from Eq. (4), the transition between the two different behaviors occurs when $2\alpha^2 = 1$ or alternatively at the following mutation probability

$$p_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right). \tag{6}$$

Although in general correlations decay exponentially $G_2(k) \sim \beta^{2k}$, the decay constant β exhibits two distinct behaviors which depend on the mutation probability p. When the mutation probability is smaller than the critical one $p < p_c$ then $\beta = \alpha$ while in the complementary case $\beta = 1/\sqrt{2}$.

As a reference, it is useful to consider the decay of the average node value $G_1(k) = \langle\sigma\rangle$. At the kth generation, the path to each node involves k branches and thus, $G_1(k) = G_1(0)\alpha^k$ with $G_1(0) = \langle\sigma_0\rangle$. Writing $G_1(k) \sim \beta^k$ then $\beta = \alpha$ for all mutation probabilities, in contrast with the asymptotic behavior of $G_2(k)$. Below the critical mutation rate, $G_2(k) \propto [G_1(k)/G_1(0)]^2$, indicating that knowledge of the one-point average suffices to characterize correlations.

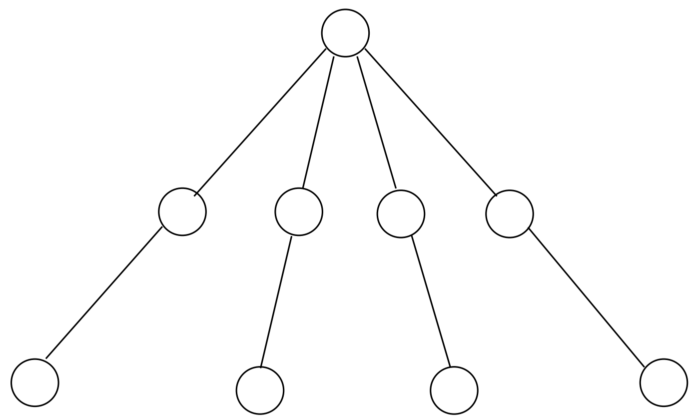

Fig. 2 The trivial "star" phylogeny. The path connecting two nodes always contains the tree root.

In fact, the above behavior can be attributed to the tree morphology. To see that, it is useful to consider a structureless morphology where the only ancestor shared by two nodes is the tree root itself (see Fig. 2). Using the notation G^* to denote correlations on this "star" morphology, we see that the average remains unchanged $G_1(k) = G_1^*(k) = G_1^*(0)\alpha^k$. The star morphology is trivial in that all genetic distances are equal: $d_{i,j} = 2k$ when $i \neq j$. Thus, pair correlations are immediately obtained from the average $G_2^*(k) = [G_1^*(k)/G_1^*(0)]^2 = \alpha^{2k}$. As branches in the star morphology do not interact, no correlations develop.

In contrast, nontrivial phylogenies do induce correlations. Indeed, $G_2(k) > G_2^*(k)$ when $p > 0$. Interestingly, when $p < p_c$, merely the asymptotic prefactor $\alpha^2/(2\alpha^2 - 1) > 1$ in Eq. (5) is enhanced and $G_2(k) \propto G_2^*(k)$. As the critical point is approached, this constant diverges thereby signaling the transition into a second regime. When $p > p_c$, the decay constant itself is enhanced and the ratio $G_2(k)/G_2^*(k)$ grows exponentially. The mutation probability affects only the asymptotic prefactor, and the decay constant $\beta = 1/\sqrt{2}$ is determined by the tree morphology. We conclude that the nontrivial phylogeny generates significant correlations for larger than critical mutation probabilities.

This behavior can be understood and partially rederived using a heuristic argument. Genetically close nodes are highly correlated, while distant pairs are weakly correlated, as indicated by Eq. (3). On the other hand, distant pairs are more numerous. Both effects are magnified exponentially for large generation numbers, and their competition results in a critical point. Different mechanisms dominate on different sides of this point. Specifically, the number of minimal genetic distance pairs ($d = 2$) is 2^{k-1}, while the number of maximal distance pairs ($d = 2k$) is $2^{2(k-1)}$. The rule (3) gives the relative contributions of these two terms to the overall two-point correlation: $2^{k-1}\alpha^2$ versus $2^{2(k-1)}\alpha^{2k}$. These are simply the first and last terms in the geometric series that led to Eq. (4). Comparing these two terms in the limit $k \to \infty$ correctly reproduces the most relevant aspects, *i.e.*, the location of the critical point (6) and the decay constants of Eq. (5). We conclude that competition between the multiplicity and the degree of correlation of close and distant nodes underlies the transition.

III. HIGHER ORDER CORRELATIONS

The above analysis gives useful intuition for the overall qualitative behavior. Yet, it can be generalized into a more complete treatment that addresses correlations of arbitrary order. This set of quantities is helpful in determining the extent to which this picture applies, and in particular, whether the transition is actually a phase transition.

Multiple point correlations obey a rule similar to Eq. (3). For example, consider the four-node average $\langle\sigma_3\sigma_4\sigma_5\sigma_6\rangle$ in Fig. 1. Using the τ variables, we rewrite $\langle\sigma_3\sigma_4\sigma_5\sigma_6\rangle = \langle\sigma_0^4\tau_1^2\tau_2^2\tau_3\tau_4\tau_5\tau_6\rangle$, and since $\sigma^2 = \tau^2 = 1$ we get $\langle\sigma_3\sigma_4\sigma_5\sigma_6\rangle = \langle\tau_3\tau_4\tau_5\tau_6\rangle = \langle\tau\rangle^4$ or $\langle\sigma_3\sigma_4\sigma_5\sigma_6\rangle = \langle\sigma_3\sigma_4\rangle\langle\sigma_5\sigma_6\rangle$. The four point average equals a product of two-point averages with the indices chosen as to minimize the total number of branches. This can also be seen by tracing the path of each node to the tree root and canceling doubly counted branches. Thus, Eq. (3) generalizes as follows:

$$\langle\sigma_i\sigma_j\sigma_k\sigma_l\rangle = \langle\tau\rangle^{d_{i,j,k,l}}, \tag{7}$$

with the four-point genetic distance

$$d_{i,j,k,l} = \min\{d_{i,j} + d_{k,l}, d_{i,k} + d_{j,l}, d_{i,l} + d_{j,k}\}. \tag{8}$$

Similarly, the law for arbitrary order averages is $\langle\tau\rangle$ raised to a power equal to the n-point genetic distance. Such distance is obtained by considering all possible decompositions into pairs of nodes. The genetic distance is the minimal sum of the corresponding pair distances. Averages over an odd number of nodes can be obtained by adding a "pseudo" node at the root of the tree and using the convention $d_{i,\mathrm{root}} = k$ when i belongs to the kth generation. The average $\langle\sigma_0\rangle$ is generated by the root and this factor multiplies all odd order correlation. Since even order correlations are independent of the root value, and odd correlations are simply proportional to $\langle\sigma_0\rangle$, we set $\langle\sigma_0\rangle = 1$ in what follows without loss of generality.

The average n-point correlation is defined as follows

$$G_n(k) = \langle\langle\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_n}\rangle\rangle, \tag{9}$$

where the averages are taken over all realizations and over all possible choices of n distinct nodes at the kth generation. For the trivial star phylogeny, the n-point genetic distance is constant and equals a product of the correlation order and the generation number, $d = nk$. Consequently, all averages are trivial as knowledge of the one-point average immediately gives all higher-order averages, $G_n^*(k) = [G_1^*(k)]^n$, or explicitly

$$G_n^*(k) = \alpha^{nk}. \tag{10}$$

When the tree morphology is nontrivial, the minimal-sum rules (7)-(8) imply that such factorization no longer holds. For binary trees, it is possible to obtain these correlations recursively. Let us assign the indices $1, 2, \ldots, 2^k$ to the kth generation nodes and order them as follows $1 \le i_1 < i_2 < \cdots < i_n \le 2^k$. As the average over the realizations is performed first, the average correlation requires a summation over all possible choices of nodes

$$F_n(k) = \sum_{1\le i_1<i_2<\cdots<i_n\le 2^k} \langle\sigma_{i_1}\sigma_{i_2}\cdots\sigma_{i_n}\rangle. \tag{11}$$

Proper normalization gives the n-node correlation

$$G_n(k) = F_n(k)\Big/\binom{2^k}{n}. \tag{12}$$

Consider a group of n nodes taken from the kth generation. They all share the tree root as a common ancestor. The two first generation nodes naturally divide this group into two independently evolving subgroups. This partitioning procedure allows a recursive calculation of the correlations. Formally, a given choice of nodes $1 \le i_1 < i_2 < \cdots < i_n \le 2^k$ is partitioned into two subgroups as follows $1 \le i_1 < \cdots < i_m \le 2^{k-1}$ and $2^{k-1} + 1 \le i_{m+1} < \cdots < i_n \le 2^{k-1} + 2^{k-1}$. These subgroups involve different τ variables, so their correlations factorize

$$\langle\sigma_{i_1}\cdots\sigma_{i_n}\rangle \propto \langle\sigma_{i_1}\cdots\sigma_{i_m}\rangle\langle\sigma_{i_{m+1}}\cdots\sigma_{i_n}\rangle. \tag{13}$$

The proportionality constant depends upon the parity of m and $n - m$. Even correlations are independent of the tree root, while odd correlations are proportional to the average value of the tree root. This extends to sub-trees as well, and since $\sigma_0 = 1$, the average value of the root of both sub-trees is $\langle\tau\rangle$. This factor accompanies all odd correlations. Substituting Eq. (13) into Eq. (11) shows that the summation factorizes as well. Using $F_m(k-1)=\sum_{1\le i_1<\cdots<i_m\le 2^{k-1}}\langle\sigma_{i_1}\cdots\sigma_{i_m}\rangle$ reduces the problem to two sub-trees that are one generation shorter, and a recursion relation for $F_n(k)$ emerges

$$F_n(k)=\sum_{m=0}^{n}F_m(k-1)B_mF_{n-m}(k-1)B_{n-m}, \qquad (14)$$

with the boundary conditions $F_n(0)=\delta_{n,0}+\delta_{n,1}$. The summation corresponds to the $n+1$ possible partitions of a group of n nodes into two subgroups. The weight of the odd correlations is accounted for by B_n

$$B_n=\begin{cases}1 & n=2r;\\ \langle\tau\rangle & n=2r+1.\end{cases} \qquad (15)$$

Using the definition (11), the sums $F_n(k)$ vanish whenever $n>2^k$. This behavior emerges from the recursion relations as well. Additionally, one can check that the sums are properly normalized in the no mutation case ($\alpha=1$), $F_n(k)=\binom{2^k}{n}$ when $n\le 2^k$.

For sufficiently small n, it is possible to evaluate the sums explicitly using Eqs. (14). The average correlations are then found using Eq. (12)

$$\begin{aligned}
G_0(k)&=1,\\
G_1(k)&=\alpha^k,\\
G_2(k)&=\frac{\alpha^2}{2\alpha^2-1}\frac{(2\alpha^2)^k-1}{2^k-1},\\
G_3(k)&=\frac{3\alpha^{k+2}}{2\alpha^2-1}\frac{\frac{(4\alpha^2)^k-(4\alpha^2)}{4\alpha^2-1}-(2^k-2)}{(2^k-1)(2^k-2)}.
\end{aligned} \qquad (16)$$

Indeed, these quantities agree with the previous results for $n=1,2$ and equal unity when $p=0$. We see that correlations involve a sum of exponentials. Furthermore, it appears that the condition $2\alpha^2=1$ still separates two different regimes of behaviors. However, calculating higher correlations explicitly is not feasible as the expressions are involved for large n. Instead, we perform an asymptotic analysis that more clearly exposes the leading large generation number behavior.

Let us consider first the regime $p<p_c$ or equivalently $2\alpha^2>1$. From Eq. (16), we see that the leading large k behavior of the average correlation satisfies $G_n(k)\sim\alpha^{nk}$ for $n=0,1,2,$ and 3. We will show below that this behavior extends to higher order correlations, *i.e.*,

$$G_n(k)\simeq g_n\alpha^{nk}. \qquad (17)$$

In other words, the following limit $\alpha=\lim_{k\to\infty}[G_n(k)]^{1/nk}$ exists and is independent of n. As correlations are larger when the phylogeny is nontrivial, one expects that $G_n(k)\ge G_n^*(k)$ or in terms of the prefactors, $g_n\ge g_n^*=1$. Combining Eq. (12) with the leading behavior of the combinatorial normalization constant $\binom{2^k}{n}\sim 2^{nk}/n!$ gives the the asymptotic behavior of the sums

$$F_n(k)=f_n(2\alpha)^{nk}, \quad \text{with} \quad f_n=\frac{g_n}{n!}. \qquad (18)$$

Substituting Eq. (18) into the recursion relation (14) eliminates the dependence on the generation number k, and a recursion relation for coefficients f_n is found

$$f_n(2\alpha)^n=\sum_{m=0}^{n}f_mB_mf_{n-m}B_{n-m}, \qquad (19)$$

with B_n of Eq. (15). These recursion relations are consistent with the conditions $f_0=f_1=1$. The case $n=2$ reproduces the coefficient $f_2=\alpha^2/[(2\alpha)^2-2]$. The divergence at $2\alpha^2=1$ indicates that the ansatz (17) breaks down at the critical point. To show that the ansatz holds in the entire range $0\le p<p_c$, one has to show that the coefficients f_n are positive and finite for all n. Rewriting the recursion (19) explicitly $f_n[(2\alpha)^n-2B_n]=\sum_{m=1}^{n-1}f_mB_mf_{n-m}B_{n-m}$ allows us to prove this. Since $f_0=1>0$, then to complete a proof by induction one needs to show that a positive f_{n-1} implies a positive f_n. The right hand side of the recursion is clearly positive and thus the positivity of f_n hinges on the positivity of the term $(2\alpha)^n-2B_n$. When $2\alpha^2>1$, then $\alpha>1/\sqrt{2}$ and certainly $2\alpha>1$. Combining this with the inequality $(2\alpha)^2>2>2B_n$ shows that $(2\alpha)^n-2B_n>0$ when $n\ge 2$. Hence f_n is positive and finite for all n, which validates the ansatz (17) in the regime $p<p_c$.

In principle, the coefficients can be found by introducing the generating functions

$$f(z)=\sum_n f_nz^n. \qquad (20)$$

Multiplying Eq. (19) by z^n and summing over n yields the following equation for the generating functions

$$f(2\alpha z)=\left[\frac{f(z)+f(-z)}{2}+\alpha\frac{f(z)-f(-z)}{2}\right]^2. \qquad (21)$$

This equation reflects the structure of the recursion relations. A factor α is generated by each odd-index coefficient and as a results, the odd part of the generating functions $[f(z)-f(-z)]/2=f_1z+f_3z^3+\cdots$ is multiplied by α. Although a general solution of this equation appears rather difficult, it is still possible to obtain results in the limiting cases. It is useful to check that when $\alpha=1$, the above equation reads $f(2z)=f^2(z)$ which together with the boundary conditions $f_0=f_1=1$ gives $f(z)=\exp(z)$ or $f_n=\frac{1}{n!}$. As $g_n\to 1$, the trivial correlations are recovered, $G_n\to G_n^*$ indicating that role played by the tree morphology diminishes in the no mutation limit.

In the limit $p\to p_c^-$ it is possible to extract the leading behavior of the asymptotic prefactors. Here, it is sufficient to keep only the highest powers of the diverging term $1/(2\alpha^2-1)$. The calculation in this case is identical to the one detailed below for the case $p>p_c$ and we simply quote the results

$$G_n(k) \to \begin{cases} \frac{2r!}{r!}\left[\frac{\alpha^2}{2(2\alpha^2-1)}\right]^r \alpha^{nk} & n=2r; \\ \frac{(2r+1)!}{r!}\left[\frac{\alpha^2}{2(2\alpha^2-1)}\right]^r \alpha^{nk} & n=2r+1. \end{cases} \quad (22)$$

In this limit, the odd order correlations simply follow from their even counterparts and for example $f_{2r+1}=f_{2r}$.

In the complementary case $p>p_c$, it proves useful to rewrite the recursion relations (19) for the even and odd correlations separately

$$\begin{aligned} F_{2r}(k) &= \sum_{s=0}^{r} F_{2s}(k-1)F_{2r-2s}(k-1) \\ &+ \alpha^2 \sum_{s=0}^{r-1} F_{2s+1}(k-1)F_{2r-2s-1}(k-1) \quad (23) \\ F_{2r+1}(k) &= 2\alpha \sum_{s=0}^{r} F_{2s}(k-1)F_{2r-2s+1}(k-1). \end{aligned}$$

The leading asymptotic behavior of Eq. (16) implies $F_0(k)=f_0$, $F_1(k)\simeq f_0(2\alpha)^k$, $F_2(k)\simeq f_2 2^k$, and $F_3(k)\simeq f_2 2^k(2\alpha)^k$ with $f_0=1$ and $f_2=\alpha^2/[2-(2\alpha)^2]$. Let us assume that this even-odd pattern is general

$$\begin{aligned} F_{2r}(k) &= f_{2r}2^{rk}, \quad (24) \\ F_{2r+1}(k) &= f_{2r}2^{rk}(2\alpha)^k. \end{aligned}$$

Substituting this ansatz into Eq. (23) shows that the second summation in the recursion for the even correlations is negligible asymptotically. Both equations reduce to

$$f_{2r}2^r = \sum_{s=0}^{r} f_{2s}f_{2r-2s}, \quad (25)$$

and therefore the pattern (24) holds when $p>p_c$. It is seen that odd correlators are enslaved to the even ones.

To obtain the coefficients, we introduce the generating functions $f(z)=\sum_r f_{2r}z^{2r}$ which satisfies $f(0)=1$, $f'(0)=0$ and $f''(0)=f_2=\alpha^2/[2(1-2\alpha^2)]$. The recursion relation translates into the following equation for $f(z)$

$$f\left(\sqrt{2}z\right)=[f(z)]^2. \quad (26)$$

Its solution is $f(z)=\exp[(\alpha z)^2/2(1-2\alpha^2)]$. Thus, $f_{2r}=\frac{1}{r!}[f_2]^r$. From Eqs. (17)-(18), the leading asymptotic behavior in the regime $p_c<p<1/2$ is found

$$G_n(k) \simeq \begin{cases} \frac{2r!}{r!}\left[\frac{\alpha^2}{2(1-2\alpha^2)}\right]^r 2^{-kr} & n=2r; \\ \frac{(2r+1)!}{r!}\left[\frac{\alpha^2}{2(1-2\alpha^2)}\right]^r \alpha^k 2^{-kr} & n=2r+1. \end{cases} \quad (27)$$

Using the Stirling formula $n!\simeq\sqrt{2\pi n}\,n^n e^{-n}$ it is seen that the coefficients g_{2r} have nontrivial r behavior as $g_{2r}=g_{2r+1}/(2r+1)\simeq\sqrt{2}[2\alpha^2/(1-2\alpha^2)]^r r^r$.

The even order correlations have identical asymptotic behavior to the two point correlation: $\lim_{k\to\infty}[G_{2r}(k)]^{1/2rk}=\frac{1}{\sqrt{2}}$ for all r. The odd order correlations behave differently, however, as this limit depends on the correlation order: $\lim_{k\to\infty}[G_{2r+1}(k)]^{1/(2r+1)k}=\frac{1}{\sqrt{2}}(\sqrt{2}\alpha)^{1/2r+1}$. Thus, only in the limit $r\to\infty$ do the even and odd order correlations agree. However, this conclusion is misleading since the decay rate of the (properly normalized) odd order correlations $G_{2r+1}(k)/G_1(k)\sim G_{2r}(k)$ is identical to that of the even order correlations. We conclude that the decay rate of two-point correlations characterizes the decay of all higher order correlations.

From Eqs. (22) and (27), we see that the coefficients diverge according to

$$f_{2r}=f_{2r+1}\sim|p_c-p|^{-r} \quad (28)$$

as the critical point is approached, $p\to p_c$. Since the correlations must remain finite, this indicates that the purely exponential behavior must be modified when $p=p_c$. Indeed, evaluating Eq. (16) at $p=p_c$ yields $F_2(k)\simeq f_2 2^k$ and $F_3(k)=f_2 2^{3k/2}$ with $f_2=k/4$, *i.e.*, the even-odd pattern of Eq. (24) is reproduced. Furthermore, the value of f_2 shows that the diverging quantity $1/|1-2\alpha^2|$ is simply replaced by k. This implies that the coefficients become generation dependent, $f_n\to f_n(k)$. Assuming the pattern Eq. (24), substituting it into Eq. (25), and following the steps that led to Eq. (27) yields the critical behavior

$$G_n(k) \simeq \begin{cases} \frac{2r!}{r!}\left[\frac{k}{4}\right]^r 2^{-kr} & n=2r; \\ \frac{(2r+1)!}{r!}\left[\frac{k}{4}\right]^r 2^{-k(r+1/2)} & n=2r+1. \end{cases} \quad (29)$$

Generally, the diverging quantity $1/|1-2\alpha^2|$ is replaced with the finite (but ever growing) quantity k. The algebraic modification to the leading exponential behavior in Eq. (29) is reminiscent of the logarithmic corrections that typically characterize critical behavior in second order phase transitions [10].

IV. STOCHASTIC TREE MORPHOLOGIES

The question arises: how general is the behavior described above? The binary tree considered was particularly simple as it involved a fixed number of children and a fixed generation lifetime. Below we show that relaxing either of these conditions does not affect the nature of the results.

Let us first consider tree morphologies with a varying number of children, *i.e.*, the trees are generated by a stochastic branching process where with probability P_r there are r children. This probability sums to unity $\sum_r P_r=1$, and the average number of children is given by $\langle r\rangle=\sum_r rP_r$. As a result, the average number of nodes at the kth generation is $\langle r\rangle^k$, indicating that the tree "survives" only if $\langle r\rangle > 1$, a classical result of branching processes theory [11]. The rule (3) is independent of the tree morphology, and therefore, one can repeat the heuristic argument in Sec. II. The extreme contributions to the average pair correlations have the relative weights $\langle r\rangle^{k-1}\alpha^2$ and $\langle r\rangle^{2(k-1)}\alpha^{2k}$. Comparing these two terms asymptotically shows that the critical point is a simple generalization of Eq. (6)

$$p_c = \frac{1}{2}\left(1 - \sqrt{\frac{1}{\langle r\rangle}}\right). \tag{30}$$

The critical mutation rate varies from 0 to 1/2 as the average ancestry size varies between 1 and ∞. This indicates that correlations are significant over a larger range of mutation rates for smaller trees. The heuristic argument also gives the decay constant β, and the leading asymptotic behavior of Eq. (5) is generalized by simply replacing 2 with $\langle r\rangle$. A more complete treatment of this problem is actually possible and closely follows Eq. (4). Again, the ancestry size $\langle r\rangle$ replaces the deterministic value 2. As both the results and the overall behavior closely follow the deterministic case, we do not detail them here.

A second possible generalization is to morphologies with a varying generation lifetime. Such tree morphologies can be realized by considering a continuous time variable. Branching is assumed to occur with a constant rate ν. For such tree morphologies, the number of nodes $n(t)$ obeys $\dot{n}(t) = \nu n(t)$ which gives an exponential growth $n(t) = e^{\nu t}$. Similarly, the mutation process is assumed to occur with a constant rate γ. A useful characteristic of this process is the autocorrelation $A(t) = \langle\sigma(0)\sigma(t)\rangle$. To evaluate it's evolution, we note that $A(t+dt) = (1-\gamma dt)A(t) - \gamma dt A(t)$ when $dt \to 0$. Therefore, $\dot{A}(t) = -2\gamma A(t)$ and one finds $A(t) = e^{-2\gamma t}$. The quantities $n(t)$ and $A(t)$ allow calculation of the average pair correlation.

Let us pick two nodes at time t and denote their values by $\sigma_i(t)$ and $\sigma_j(t)$, and let the genetic distance between these two nodes be τ. Using their first common ancestor $\sigma_c(t-\tau) = \sigma_i(t-\tau) = \sigma_j(t-\tau)$ and the identity $\sigma^2 = 1$, their correlation can be evaluated as follows $\langle\sigma_i(t)\sigma_j(t)\rangle = \langle\sigma_i(t)\sigma_c(t-\tau)\sigma_c(t-\tau)\sigma_j(t)\rangle = \langle\sigma_i(t)\sigma_i(t-\tau)\rangle\langle\sigma_j(t)\sigma_j(t-\tau)\rangle = A^2(\tau)$. Integrating over all possible genetic distances gives the average pair correlation

$$G_2(t) = \frac{\int_0^t d\tau\, n(\tau)A^2(\tau)}{\int_0^t d\tau\, n(\tau)}. \tag{31}$$

The factor $n(\tau)/\int_0^t d\tau\, n(\tau)$ accounts for the multiplicity of pairs with genetic distance τ. Using $A(t) = e^{-2\gamma t}$ and $n(\tau) = e^{\nu t}$, the average pair correlation is evaluated

$$G_2(t) = \frac{\nu}{\nu - 4\gamma}\frac{e^{(\nu-4\gamma)t} - 1}{e^{\nu t} - 1}. \tag{32}$$

For the star phylogeny the genetic distance is always t and therefore $G_2^*(t) = e^{-4\gamma t}$. Here the relevant parameter is the normalized mutation rate $\omega = \gamma/\nu$. Again, there exists a critical point $\omega_c = 1/4$. For smaller than critical mutation rates, $\omega < \omega_c$, correlations due to the tree morphology are not pronounced, $G_2(t) \propto G_2^*(t)$. On the other hand, when $\omega > \omega_c$, strong correlations are generated and $G_2(k) \sim e^{-\nu t}$ is exponentially larger than $G_2^*(t)$. We conclude that the behavior found for the deterministic case is robust.

V. MULTISTATE SEQUENCES

We now consider larger alphabets. Previously, the two states satisfied $\sigma^2 = 1$. A natural generalization is to $\sigma^n = 1$, i.e, the nth order roots of unity $\sigma = e^{i2\pi l/n}$ with $l = 0, 1, \ldots, n-1$. Previously, with probability p the mutation $\sigma \to \tau\sigma$ occurred with $\tau = e^{i\theta}$ and $\theta = \pi$. We thus impose the same transition but with $\theta = 2\pi/n$. This can be viewed as a clockwise rotation in the complex plane by an angle θ. Since the states are now complex, the definition of the pair correlation is now

$$G_2(k) = \langle\langle\bar{\sigma}_i\sigma_j\rangle\rangle, \tag{33}$$

with $\bar{\sigma}$ the complex conjugate of σ. The real part of $\bar{\sigma}_i\sigma_j$ gives the inner product of the two-dimensional vectors corresponding to σ_i and σ_j, respectively.

Consider the average $\langle\bar{\sigma}_3\sigma_4\rangle$ in Fig. 1. Using the τ variables and $\bar{\tau}\tau = \bar{\sigma}\sigma = 1$ one has $\langle\bar{\sigma}_3\sigma_4\rangle = \langle\bar{\sigma}_0\bar{\tau}_1\bar{\tau}_3\sigma_0\tau_1\tau_4\rangle = \langle\bar{\tau}_3\tau_4\rangle = \langle\bar{\tau}_3\rangle\langle\tau_3\rangle = \overline{\langle\tau\rangle}\langle\tau\rangle = |\langle\tau\rangle|^2$. All of our previous results hold if one replaces the average $\langle\tau\rangle$ with its magnitude $\alpha = |\langle\tau\rangle| = |1 - p(1-e^{i\theta})| = \sqrt{1 - 2p(1-p)(1-\cos\theta)}$. Furthermore, it is sensible to consider arbitrary phase shifts $0 < \theta < 2\pi$ since the identity $\bar{\sigma}\sigma = \bar{\tau}\tau = 1$ rather than $\sigma^n = \tau^n = 1$ was used to evaluate correlations.

The critical point is determined from the condition $\langle r\rangle\alpha^2 = 1$. This equation has a physical solution only when $2\varphi < \theta < 2(\pi - \varphi)$ with the shorthand notation

$$\varphi = \cos^{-1}\sqrt{\frac{1}{\langle r\rangle}}. \tag{34}$$

In terms of the number of states, this translates to

$$\frac{\pi}{\pi - \varphi} < n < \frac{\pi}{\varphi}. \tag{35}$$

Hence, the transition may or may not exist depending on the details of the model, in this particular "clock" model case, the number of states. As we have seen before, correlations become less pronounced when the number of ancestors increases. Indeed, the transition always exists in the limit $\langle r\rangle \to 1$, while the transition is eliminated in the other extreme $\langle r\rangle \to \infty$. When the transition do occur, the following critical mutation probability is found

$$p_c = \frac{1}{2}\left[1 - \sqrt{1 - \frac{\sin^2\varphi}{\sin^2\frac{\theta}{2}}}\,\right]. \tag{36}$$

Indeed, Eq. (30) is reproduced in the two state case ($\theta = \pi$). This turns out to be the minimal critical point, $p_c \geq (1 - \sqrt{1/\langle r\rangle})/2$, reflecting the fact that transition $\sigma \to -\sigma$ provides the most effective mutation mechanism. In effect, increasing the number of states reduces the mutation rate, and this mechanism is responsible for eliminating the transition.

Interestingly the transition is restored when both the mutation and the duplication processes occur continuously in time. In this continuous description duplication occurs with rate ν and the mutation $\sigma \to e^{i\theta}\sigma$ occurs with rate γ. The autocorrelation $A(t) = \langle\sigma(0)\bar{\sigma}(t)\rangle = \exp\left[-\gamma(1 - e^{-i\theta})t\right]$ is found from its time evolution $\dot{A}(t) = -\gamma(1 - e^{-i\theta})A(t)$. It can be easily shown from the definition of the pair correlation (33) that $A^2(\tau)$ should be replaced with $|A(\tau)|^2 = \exp\left[-2\gamma(1 - \cos\theta)\tau\right]$ in the integral (31). Comparing with the previous section results, we see that the effective mutation rate is now $\gamma(1 - \cos\theta)/2$. As a result, the location of the critical point is increased by a factor $2/(1 - \cos\theta)$. Using the normalized mutation rate $\omega = \gamma/\nu$ one finds

$$\omega_c = \frac{1}{2(1 - \cos\theta)}. \tag{37}$$

This critical point increases with the number of states, and it diverges according to $\omega_c \simeq (n/2\pi)^2$ when $n \to \infty$. This behavior is intuitive as one expects that mutations between a large number of states diminishes correlations and consequently, phylogenetic effects.

VI. TWO-SITE CORRELATIONS

When sequences are not of unit length, *i.e.*, when there are two or more sites per sequence, the results can be used to characterize a correlation measure quantifying the interaction between sites. Assume there are two or more sites per sequence and that the sites evolve independently of each other. Denote the state of position a in sequence i as σ_i^a, and similarly denote the state of position b in sequence i as σ_i^b. If the sequences were not related by a phylogenetic tree, but instead were independent samples drawn from a given distribution, then the following quantity defined on a finite set of $N = 2^k$ samples specifies a two-site correlation measure:

$$\rho = \frac{1}{N}\sum_i \sigma_i^a\sigma_i^b - \frac{1}{N^2}\sum_i \sigma_i^a \sum_j \sigma_j^b. \tag{38}$$

Correlation between sites a and b is indicated by a non-zero value of ρ. Related correlation measures, defined over all sequences at a given generation, such as the mutual information are used in analysis of biological sequences.

The quantity ρ is well defined also when the sequences are related by a phylogenetic tree. Due to the assumption of independent positions, the mean of ρ over all realizations vanishes $\langle\rho\rangle = 0$. This behavior is independent of the tree morphology. To see the effects of the phylogeny, one needs to consider fluctuations, *i.e.*, the variance $\Delta\rho = \langle\rho^2\rangle$

$$\begin{aligned}
\Delta\rho &= \Big\langle\Big(\frac{1}{N}\sum_i \sigma_i^a\sigma_i^b\Big)^2\Big\rangle - \Big\langle\Big(\frac{1}{N}\sum_i \sigma_i^a\Big)^2\Big(\frac{1}{N}\sum_i \sigma_i^b\Big)^2\Big\rangle \\
&= \frac{1}{N^2}\sum_{ij}\langle\sigma_i^a\sigma_j^a\rangle\langle\sigma_i^b\sigma_j^b\rangle - \frac{1}{N^4}\sum_{ij}\langle\sigma_i^a\sigma_j^a\rangle\sum_{kl}\langle\sigma_k^b\sigma_l^b\rangle \\
&= \frac{1}{N^2}\Big[N + \sum_{i\neq j}\langle\tau\rangle^{2d_{i,j}}\Big] - \frac{1}{N^4}\Big[N + \sum_{i\neq j}\langle\tau\rangle^{d_{i,j}}\Big]^2.
\end{aligned}$$

The first equality in the above equation was obtained by rewriting Eq. (38) as $\rho = \rho_1 - \rho_2$ and noting that $\langle\rho_1\rho_2\rangle = \langle\rho_2^2\rangle$. The final expression can be simplified using $\sum_{i\neq j}\langle\tau\rangle^{2d_{i,j}} = N(N-1)G_2(\alpha^2, k)$ with $G_2(\alpha^2, k)$ the pair correlation of Eq. (4), considered as a function of α^2. The following expression for the variance is obtained

$$\Delta\rho = \Big[\frac{1}{N} + \big(1 - \frac{1}{N}\big)G_2(\alpha^2, k)\Big] - \Big[\frac{1}{N} + \big(1 - \frac{1}{N}\big)G_2(\alpha, k)\Big]^2. \tag{39}$$

For the star morphology the leading order of the fluctuations is independent of the mutation rate and it scales as the familiar N^{-1}. For the binary tree morphology, there are again two regimes, characterized by $p > p_c$ or $p < p_c$ where p_c is now defined by $2\alpha^4 = 1$, *i.e.*,

$$p_c = \frac{1}{2}\left(1 - \left(\frac{1}{2}\right)^{\frac{1}{4}}\right). \tag{40}$$

When $p < p_c$, the phylogeny plays a significant role and the variance is exponentially enhanced $\Delta\rho \sim \alpha^{4k}$, while when $p > p_c$, the variance is still statistical in nature $\Delta\rho \simeq A\Delta^*\rho$ with $A > 1$. Hence, it is more likely to observe large values of ρ in the tree morphology than it is in the star morphology, even when the sites evolve independently. Since correlations and variances play opposite roles, they are influenced in different ways by the phylogeny.

VII. SUMMARY

In summary, we have studied the influence of the phylogeny on correlations between the tree's nodes. In general, for sufficiently small mutation rates, the morphology plays a minor role. For sufficiently high mutation rates large correlations that can be attributed to the phylogeny may occur. The transition between the two regimes of behavior is sharp and is marked by a critical mutation rate. Below this critical point all correlations are well described by the average, while above it, correlations decay much slower than the average. Underlying this transition is the competition between the multiplicity and the the degree of correlations between genetically close and distant leafs. This competition also leads to larger fluctuations in the correlation between different sites, even when these evolve independently.

We have also seen that this behavior is robust and appears to be independent of many details of the model. While the overall behavior generally holds, specific details such as the location of the critical point and the decay rate in the regime $p > p_c$ depend on a specific tree dependent parameter: the average number of children.

The above results can be extended in several directions. It will be interesting to see whether the recursive methods can be generalized to stochastic tree morphologies and in particular to the continuous time case. This methods should still be applicable even when the mutation rates are time dependent or disordered. In such cases it will be interesting to determine which parameters determine the critical point, the decay constants, etc.

Correlations can serve as useful measure of the diversity of a system since small correlations indicate large diversity and vice versa. If the diversity can be measured in an experiment where the phylogeny is controlled, its time dependence can be used to infer the mutation probability. Similarly, if the mutation probability can be controlled, than the degree of correlation/diversity can be used to infer characteristics of the phylogeny. Thus, our results may be useful for inferring statistical properties of actual biological systems.

This research is supported by the Department of Energy under contract W-7405-ENG-36.

[1] M. Kimura, *The natural theory of evolution* (Cambridge Press, Cambridge 1983).

[2] M. S. Waterman, *Introduction to Computational Biology: Maps, sequences and genomes*, (Chapman & Hall, London, 1995).

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*, (Cambridge Press, Cambridge, 1998).

[4] R. Levine and M.Tribus, *The maximum Entropy Formalism*, (MIT press, Cambridge MA, 1979).

[5] H. Kishino, T. Miyata, and M. Hasegawa, *J. Mol. Evol.* **31**, 151 (1990).

[6] S. Altschul, R. Carroll, D. Lipman, *J. Mol. Biol.* **207**, 647 (1989).

[7] W. Li, *Phys. Rev. A* **43**, 5240 (1991)

[8] Lapedes, A.S., Giraud, B.G., Liu, L.C. and Stormo, G.D., *Correlated mutations in protein sequences: phylogenetic and structural effects*, Proceedings of the AMS/SIAM Conference: Statistics in Molecular Biology (Seattle, WA 1997).

[9] B. G. Giraud, A. S. Lapedes, and L. C. Liu, *Phys. Rev. E*, **58**, 6312 (1998).

[10] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena*, (Oxford, Press, Oxford, 1971).

[11] T. E. Harris, *The theory of branching processes*, (Dover, New York, 1989).