Phylogenetic Correlations in Mutation Processes

E. Ben-Naim and A. S. Lapedes

CNLS & Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545

The influence of phylogenetic trees on correlations in mutation processes is investigated. 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. An identical critical behavior underlies all multiple point correlations. This behavior generally applies to branching processes undergoing mutation.

Goals

- Time evolution of inter-sequence correlations.
- Influence of Phylogeny (family tree) on sequence evolution.

The Mutation-Duplication Model

- **Sequences:** Simplest possible: alphabet size=2, length=1. The numeric values $\sigma_i = \pm 1$ are attached to the two states.
- **Random mutation process:** Poisson Statistics: $\sigma \rightarrow -\sigma$ (mutation event) with p the mutation probability.
- **Binary tree phylogeny:** Every parent has two children.

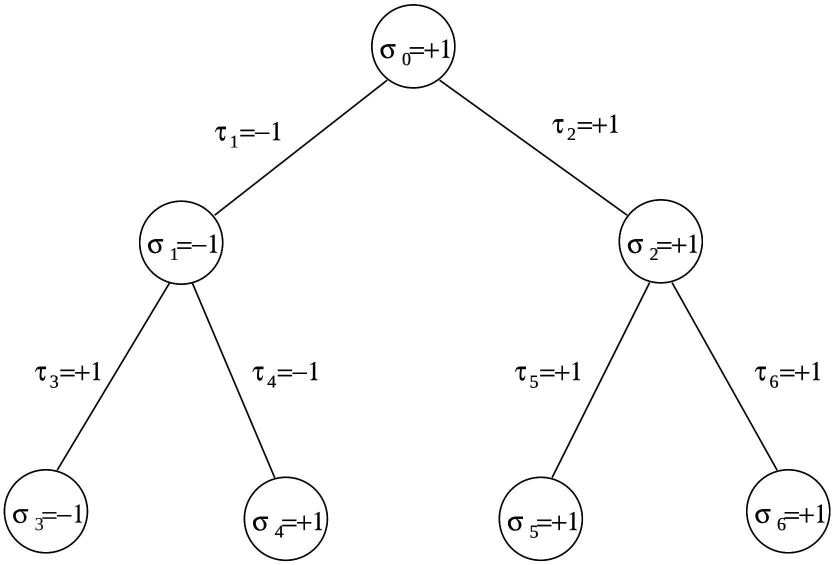

Fig. 1. The mutation process on a two-generation tree. The multiplicative variable τ ($\sigma_i = \tau_i \sigma_j$, j parent of i) indicates whether a mutation occurred.

Law for Correlations

- **Pair correlations** $\langle\sigma_i\sigma_j\rangle$. Example (Fig. 1):

$$\langle\sigma_3\sigma_4\rangle = \langle\sigma_0\tau_1\tau_3\sigma_0\tau_1\tau_4\rangle = \langle\sigma_0^2\tau_1^2\tau_3\tau_4\rangle =^{(*)} \langle\tau_3\tau_4\rangle =^{(**)} \langle\tau\rangle^2$$

$$\langle\sigma_3\sigma_5\rangle = \langle\sigma_3\sigma_6\rangle = \langle\tau\rangle^4 \qquad \text{(following similar calculation)}$$

(*) two-state symmetry: $\sigma^2 = \tau^2 = 1$

(**) identical, independent random variables: $\langle\tau_i\tau_j\rangle = \langle\tau_i\rangle\langle\tau_j\rangle$

- In general, let $d_{i,j}$ be the **genetic distance**, the minimal number of bonds connecting the nodes i, j ($d_{3,4} = 2$, $d_{3,5} = 4$)

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

- **Multiple point correlations** follow a similar rule,

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

- $d_{i,j,k,l}$ the generalized 4-point genetic distance, the minimal number of nodes connecting the nodes i, j, k, l:

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

- Generally, n-point correlations $= \langle\tau\rangle^{d_n}$. The n-point genetic distance d_n = minimal number of bonds connecting the n-nodes.

Average Correlations

- **Average pair correlation** at k generation

$$G_2(k) = \langle\langle \sigma_i \sigma_j \rangle\rangle$$

Average taken over all (i) realizations (ii) i, j from kth generation.

- For example, at second generation (Fig. 1): $G_2(2) = [\langle \sigma_3 \sigma_4 \rangle + \langle \sigma_3 \sigma_5 \rangle + \langle \sigma_3 \sigma_6 \rangle]/3 = (\langle \tau \rangle^2 + 2\langle \tau \rangle^4)/3$. In general, the geometric series $G_2(k) = (\langle \tau \rangle^2 + 2\langle \tau \rangle^4 + \cdots + 2^{k-1}\langle \tau \rangle^{2k})/(2^k - 1)$, or

$$G_2(k) = \frac{\langle \tau \rangle^2}{2\langle \tau \rangle^2 - 1} \frac{(2\langle \tau \rangle^2)^k - 1}{2^k - 1}.$$

- **Two asymptotic behaviors** marked by $p_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$

$$G_2(k) \sim \begin{cases} \langle \tau \rangle^{2k} & p < p_c; \\ 2^{-k} & p > p_c. \end{cases}$$

- **Low mutation rates:** Phylogeny is marginally relevant. Exponential decay of correlations has same constant as for trivial star phylogeny. Only overall prefactor (> 1) is enhanced.

- **High mutation rates:** Phylogeny generates strong correlations. Exponential decay constant is enhanced. It depends on tree morphology rather than the mutation probability.

The Star Phylogeny

- The trivial star phylogeny serves as a reference.

- All genetic distances are equal $d_{i,j} = 2k$.

- Factorizing (mean-field) correlations

$$G_2^*(k) = [G_1^*(k)]^2 = \langle \tau \rangle^{2k}$$

- Compare tree with star in $k \to \infty$ limit

$$\frac{G_2(k)}{G_2^*(k)} \to \begin{cases} \text{const.} & p < p_c; \\ \infty & p > p_c. \end{cases}$$

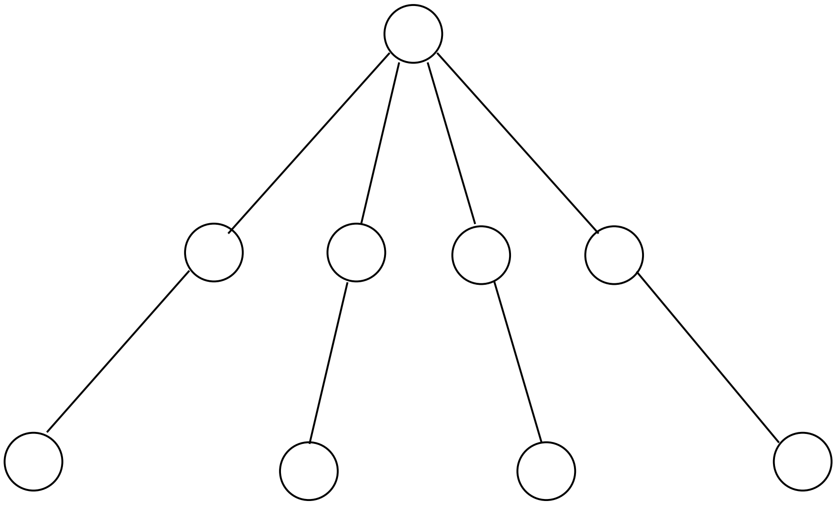

Fig. 2. The path connecting two nodes always contains the tree root.

Higher Order Correlations

- **Average n-node correlation** at kth generation

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

- Obtain from $G_n(k) = F_n(k)/\binom{2^k}{n}$ using the sums

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

- Corresponding **recursion relations** reflect the binary tree

$$F_n(k) = \sum_{m=0}^{n} F_m(k-1)F_{n-m}(k-1)B_m B_{n-m}$$

- Boundary conditions $F_n(0) = \delta_{n,0} + \delta_{n,1}$

- Tree root counted for odd order correlators, as reflected by B_n

$$B_n = \begin{cases} 1 & n = 2r \\ \langle\tau\rangle & n = 2r+1. \end{cases}$$

- Analysis in the asymptotic ($k \to \infty$ limit) is possible using **generating functions** techniques.

Asymptotic Behavior

Low mutation rates $p < p_c$: All correlations behave similarly as $\lim_{k\to\infty}[G_n(k)]^{1/nk} = \langle\tau\rangle$. For the star case $G_n^*(k) = \langle\tau\rangle^{nk}$ and thus, correlations are only marginally enhanced due to phylogeny as $G_n(k)/G_n^*(k) = A_n > 1$.

$$G_n(k) \simeq A_n\langle\tau\rangle^{nk}$$

High mutation rates $p > p_c$: Phylogeny generates significant correlations. Odd correlations are enslaved to even ones $G_{2r+1}(k) = (2r+1)\langle\tau\rangle^k G_{2r}(k)$. Correlations decay slower and the decay rate depends on tree morphology only. Ratio with trivial star phylogeny diverges $G_n(k)/G_n^*(k) \to \infty$ for $n > 1$.

$$G_{2r}(k) \simeq C_r 2^{-rk}$$

Nature of transition: can be understood heuristically. Near relatives are strongly correlated but exponentially rare. Far relatives are abundant but exponentially weak. When $p < p_c$ far relatives dominate, and when $p > p_c$ near relatives dominate.

Same critical point underlies all correlators

Generalizations

Stochastic tree morphologies: Only relevant parameter is the average number of children $\langle k \rangle$. Critical point shifts, role of phylogeny diminishes for larger trees

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

Continuous time formulation: Mutation occurs with rate γ, birth with rate ν. In terms of the dimensionless mutation rate $\theta = \gamma/\nu$ the transition occurs at $\theta_c = 1/4$.

Larger alphabets: consider n-state "clock" model $\sigma = \exp(i2\pi j/n)$ with $j = 0, 1, \ldots, n$ with the rotation mutation $\sigma \to \sigma \exp(i2\pi/n)$. Critical point shifts, role of phylogeny diminishes with increasing the number of states

$$\theta_c = \frac{1}{2(1 - \cos\frac{2\pi}{n})}.$$

$$\boxed{\textbf{Nature of transition remains the same}}$$

Conclusions

- Correlations decay exponentially with time, genetic distance.
- Phylogeny matters only when the mutation rates is high.
- Transition is critical in nature: all correlations behave similarly.
- Results apply to a large class of mutation/duplication processes.
- Role of phylogeny decreases as alphabet size, tree size increases.

References

1. S. Altschul, R. Carroll, D. Lipman, *J. Mol. Biol.* **207**, 647 (1989).
2. B G. Giraud, A. S. Lapedes, and L. C. Liu, *PRE* **58**, 6312 (1998).
3. E. Ben-Naim and A. S. Lapedes, *Phys. Rev. E* **59**, 7000 (1999).