Existence of Critical Phenomenon in Mutation Processes

Dissertation submitted by **Parag Ghosh**

Date: 31.10.2001

Introduction

The field of molecular evolution owes most of its existence to the possibility of sequencing proteins and nucleic acids. Molecular sequences provide us with precisely comparable characters observed at or near the level of the gene, which can be examined in diverse organisms. The amount of data is very large and rising rapidly. It enables us to work in two modes: we can either use our knowledge in the evolutionary history of the species to examine the mechanisms of evolution of the molecules, or we can use knowledge of the evolution of the molecules to infer the evolutionary history of the species.

Inferring phylogenetic relationships from molecular data requires the selection of an appropriate method. Phylogenetic analysis is often treated as a black box into which data are fed and out of which "The Tree" springs. But in this kind of approach one often overlooks certain conceptual intricacies one of which is the rate dependent correlation among species.

Aim

In this essay the role of phylogenetic trees on correlations is studied. The dependence of tree morphology on mutation rates is analyzed by formulating a model based on binary trees and later on generalizing the binary tree to stochastic tree morphologies.

In a phylogenetic tree the genetically close species are highly correlated while the distant pairs are weakly correlated. On the otherhand the distant pairs are more in number. These two effects increase exponentially for large

generation numbers and their competition results in critical point. It is this critical behavior that we are interested in. The model described below identifies that it is the rate of mutation which is characterizes such critical behavior.

Fig. 1 Binary tree showing the mutation process in two generations.

Suppose we are comparing sequences which are of unit length; there are only two possibilities which we denote by the numeric values $\sigma = \pm 1$. As we know for a binary tree each parent has two children. The tree is deterministic since the number of children and the generation lifetime are fixed. We assume that the probability of a mutaion to occur is p and hence the probability that a child equals her predecessor is $1-p$. The mutation process is invariant under the transformation $\sigma \to -\sigma$ and $p \to 1-p$ and we restrict ourselves to the case $0 \geq p \leq 1/2$ without the loss of generality.

Next we define the average correlation between two nodes at the k^{th} generation as:

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

where the first average is over all realizations for a fixed pair of nodes $i \neq j$, while the secind average is taken over all different pairs belonging to the same

generation. For example $G_2(2) = [\langle\sigma_3\sigma_4\rangle + \langle\sigma_3\sigma_5\rangle + \langle\sigma_3\sigma_6\rangle]/3$. The relation between each parent and child is drawn by assigning a multiplicative random variable $\tau_i = \pm 1$ to each branch of the tree such that $\sigma_i = \tau_i\sigma_j$ where j is the parent of i. The probability for $\tau = 1(-1)$ is $1 - p(p)$ and hence

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

To obtain pair correlations we proceed as follows:

$$\begin{aligned} \sigma_3 &= \sigma_0\tau_1\tau_3 \\ \sigma_4 &= \sigma_0\tau_1\tau_4 \\ \langle\sigma_3\sigma_4\rangle &= \langle\sigma_0^2\tau_1^2\tau_3\tau_4\rangle. \end{aligned}$$

Since $\sigma_i^2 = \tau_i^2 = 1$ we have $\langle\sigma_3\sigma_4\rangle = \langle\tau^2\rangle$. We next assume that the mutations on different branches are not correlated and hence $\langle\tau_i\tau_j\rangle = \langle\tau_i\rangle\langle\tau_j\rangle$, when $i \neq j$. Thus $\langle\sigma_3\sigma_4\rangle = \langle\tau\rangle^2$ and similarly $\langle\sigma_3\sigma_5\rangle = \langle\tau\rangle^4$ and $\langle\sigma_3\sigma_6\rangle = \langle\tau\rangle^4$. So the general process of calculating the correlations between two nodes is to trace back the path that connects the two nodes. Also since $\tau^2 = 1$ so only the path which connects to the common ancestor is relevant. We can define a generic distance $d_{i,j}$ bewteen two nodes i and j as the minimum number of branches required to connect them, so that

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

Now lets define a new variable $\alpha = \langle\tau\rangle$. Then pair correlation at the second generation is given by $G_2(2) = (\alpha^2 + 2\alpha^4)/3$. This result when generalized to the k^{th} level gives rise to a geometric series:
$G_2(k) = (\alpha^2 + 2\alpha^4 + ... + 2^{k-1}\alpha^{2k})/(2^k - 1)$
Evaluating the sum

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

We are now in a position to study the effect of mutation rates on the pair correlation. For sufficiently large generation number k, we have $G_2(k) = \frac{\alpha^2}{2\alpha^2-1}2^k\alpha^{2k}$. Now depending on the value of α or p we have:

For $\alpha > 1/\sqrt{2}$ or $p < p_c$:

$$G_2(k) \approx \frac{\alpha^2}{2\alpha^2 - 1}\alpha^{2k} \tag{5}$$

For $\alpha < 1/\sqrt{2}$ or $p > p_c$:

$$G_2(k) \approx \frac{\alpha^2}{1-2\alpha^2} 2^{-k} \tag{6}$$

where the mutation probability $p_c = \frac{1}{2}\left\{1 - \frac{1}{\sqrt{2}}\right\}$ characterizes the transition. It can shown that below the critical mutation rate, $G_2(k) \propto [G_1(k)/G_1(0)]^2$ which implies that the knowledge about one-point average is enough to characterize correlations. It also indicates that below the critical mutation rate the behavior is trivial and independent of the tree morphology. It is the behavior above the critical point which depends on tree morphology and nontrivial phylogenies do induce correlations.

The results discussed above can be generalized to higher order correlations. One observes logarithmic corrections that typically characterizes critical behavior in second order phase transitions.[1]

For tree morphologies with $\langle r \rangle$ number of children the trees are generated by a stochastic branching process where the probablility for having r children is P_r. Then $\langle r \rangle = \sum_r r P_r$. The average number of nodes at the k^{th} generation is $\langle r \rangle^k$. Since the rule given by eqn(3) is independent of tree morphologies, one can work out the same procedure (as has been done for binary trees) and obtain:

$$P_c = \frac{1}{2}\left\{1 - \sqrt{\frac{1}{\langle r \rangle}}\right\} \tag{7}$$

So we can see that the results we obtained for binary tree can be generalized to stochastic tree morphologies by replacing the deterministic factor 2 by $\langle r \rangle$.

Conclusion

In this essay the effect of phylogeny on correlations is studied. It is observed that the rate of mutation plays an important role in determining the behavior of correlation functions. For mutation rates less than a critical value, all correlations are well described by the average and the tree morphology does not play any role. Above the critical rate non-trivial phylogenies can give rise to strong correlation among species. The transition is sharp and bear resemblance with the characteristics of second order transition. The studied behavior can also be generalized to stochastic tree morphologies.

References

[1] Ben-Naim E., Lapedes A.S., Preprint cond-mat/**9812184**

[2] Woese C.R., *PNAS* **97** (2000) 8392,

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