Origin of life and evolution: phylogenetic tree

The variation of the **Object's** position in **Space** within a **time** period is the essence of dynamics of physical world. The replacement of **Nucleotide** at different **sites of DNA** within a **time** period are the milestones of evolution of life. That shed light on our understanding of the evolution of life on the Earth and the Universe. The study of life evolution contribute to our understanding of different disciplines from geology to astronomy, wherever biology intervene without forgetting medicine and epidmeiology in fighting diseases through drug and vaccine development, mapping transmission networks, etc.

Module: Constructing phylogenetic tree

The uniqueness of DNA (RNA) from one lineage to another allows to reconstuct the evolutionary history through a phylogenetic tree and get information on the past evolutionary processes. A phylogenetic tree is composed of branches (edges) and nodes. Branches connect nodes; a node is the point at which two (or more) branches diverge. Branches and nodes can be internal or external (terminal). An internal node corresponds to the hypothetical last common ancestor (LCA) of everything arising from it. Terminal nodes correspond to the sequences (DNA/RNA) from which the tree was constructed. The lengths of the branches correspond to the amount of evolution (roughly, percent sequence difference) between the two nodes they connect.

Mathematical background: Markov process & autonomous dynamic systems

The DNA (RNA) evolution can be seen as a continuous-time Markov process since at each site of the DNA (RNA) must be one of the four bases (Guanine, Cytosine, Adenine and Thymine or Uracile if it is RNA). Therefore, the variation of DNA (RNA) from one lineage to another is associated to the probability of replacement of bases at each site. Thus, with a DNA sequence of a fixed length L (it has L sites)evolving in time by base replacement. Assuming that the processes followed by the L sites are Markovian independent, identically distributed and that the process is constant over time. For a given site of the DNA, if we denote P(t) = (p_A(t), p_G(t), p_C(t), p_T(t)), the probabilities of states A, G, C and T at time t.

For each nucleotide, its frequency at time t+dt is equal to the frequency of the nucleotide at time t, minus the frequency of the its lost plus the frequency of the newly created in that small time interval. Thus, the changes in the probability distributions (p_A(t), p_G(t), p_C(t), p_T(t)) for small time is given by P(t+dt) = P(t) + QP(t)dt, where Q is a matrix of substitution rates between bases.

$$\begin{pmatrix} P_A(t+\Delta t) \\ P_G(t+\Delta t) \\ P_C(t+\Delta t) \\ P_T(t+\Delta t) \end{pmatrix} = \begin{pmatrix} P_A(t) \\ P_G(t) \\ P_C(t) \\ P_T(t) \end{pmatrix} + \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix} \begin{pmatrix} P_A(t) \\ P_G(t) \\ P_C(t) \\ P_T(t) \end{pmatrix} \Delta t$$

That leads to an autonomous system P'(t) = QP(t), which is an ordinary differential equation. Most of the evolutionary models are based on the above mathematics and they differe only in terms of assumptions and values of the rate matrix.

Thus, taking into account the susbtitution of nucleotides, the amount of sequence divergence provides information about the number of changes that have occurred along the path separating the sequences, and subsequently the lengths of branches of the phylogenetic tree.

Interaction/Activities: Visitors can simulate sequences and contruct themselves phylogenetic trees in R enviroment.

Author: David Niyukuri