2022

Certificate Authority Cup International Mathematical Contest Modeling

http://mcm.tzmcm.cn

Problem B (MCM)
The Genetic Process of Sequences

Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life[1]. Homology among DNA, RNA, or proteins is typically inferred from their nucleotide or amino acid sequence similarity. Significant similarity is strong evidence that two sequences are related by evolutionary changes from a common ancestral sequence[2].

Consider the genetic process of a RNA sequence, in which mutations in nucleotide bases occur by chance. For simplicity, we assume the sequence mutation arise due to the presence of change (transition or transversion), insertion and deletion of a single base. So we can measure the distance of two sequences by the amount of mutation points. Multiple base sequences that are close together can form a family, and they are considered homologous.

Your team are asked to develop a reasonable mathematical model to complete the following problems.

1. Please design an algorithm that quickly measures the distance between two sufficiently long($> 10^3$ bases) base sequences.
2. Please evaluate the complexity and accuracy of the algorithm reliably, and design suitable examples to illustrate it.
3. If multiple base sequences in a family have evolved from a common ancestral sequence, design an efficient algorithm to determine the ancestral sequence, and map the genealogical tree.

References

[1] Koonin EV. "Orthologs, paralogs, and evolutionary genomics". Annual Review of Genetics. 39: 30938, 2005.

[2] Reeck GR, de Han C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, et al. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell. 50 (5): 667, 1987.