BULL. AUSTRAL. MATH. SOC.
VOL. 41 (1990) [159–160]

05C05, 60C05, 92A12

Distributions on bicoloured evolutionary trees

MICHAEL A. STEEL

A central and challenging problem in contemporary biology is how to reconstruct evolutionary trees from DNA sequence data accurately. This thesis addresses three themes from this endeavour — comparison, consistency and confidence intervals — by analysing distributions arising from phylogenetic trees.

Toward the first theme, the distribution of the symmetric difference metric on pairs of binary and phylogenetic trees is studied, and a number of new results obtained. These theorems, as well as a result of another tree metric answer previous conjectures in this area. Also under the theme of comparison, we analyse distributions on bicoloured trees arising from the principle of parsimony. A streamlined proof is given of an elegant theorem which allows an efficient comparison of how much better a maximum parsimony tree fits given data than a randomly-chosen tree. A dual distribution, where the tree is fixed and the data varies is also analysed, answering a recent unsolved problem.

We then consider the theoretical accuracy of tree-building methods, concentrating on the statistical property of consistency. Under a simple stochastic model on bicoloured trees, conditions for the consistency of frequently-used methods based on parsimony and compatibility are examined. It is shown that even in "best possible" conditions both methods can be inconsistent, though a strong sufficient condition for compatibility is given. The analysis is extended for a molecular clock.

Finally, procedures are described for placing confidence intervals around phylogenies, and limitations on the sort of confidence intervals possible are given. Ways to implement these procedures efficiently are then considered – in particular, approximate methods, applications to sets of taxa of size four, and simplifications under a molecular clock.

The rate that sequence data must grow as a function of the number of taxa for confidence intervals to converge to a single tree is also considered.

The arguments in this thesis are primarily combinatorial and stochastic. In the hope that their implications will also interest biologists, some space has been given to

Received 25 July, 1989. Thesis submitted to Massey University, March 1989. Degree approved June 1989. Supervisors: Dr M.D. Hendy and Dr. D. Penny.

motivating and explaining the biological relevance of the results presented.

Department of Mathmatics and Statistics
Massey University
Palmerston North
NEW ZEALAND