Cladistics (1996) 12:65–82

ON COMBINING PROTEIN SEQUENCES AND NUCLEIC ACID SEQUENCES IN PHYLOGENETIC ANALYSIS: THE HOMEOBOX PROTEIN CASE

Donat Agosti[1], David Jacobs[2] and Rob DeSalle[1]

[1]*Department of Entomology, American Museum of Natural History, New York, NY 10024, U.S.A. and* [2]*Department of Biology, University of California, Los Angeles CA, U.S.A.*

Received for publication 28 June 1995; accepted 20 December 1995

Abstract — Amino acid encoding genes contain character state information that may be useful for phylogenetic analysis on at least two levels. The nucleotide sequence and the translated amino acid sequences have both been employed separately as character states for cladistic studies of various taxa, including studies of the genealogy of genes in multigene families. In essence, amino acid sequences and nucleic acid sequences are two different ways of character coding the information in a gene. Silent positions in the nucleotide sequence (first or third positions in codons that can accrue change without changing the identity of the amino acid that the triplet codes for) may accrue change relatively rapidly and become saturated, losing the pattern of historical divergence. On the other hand, non-silent nucleotide alterations and their accompanying amino acid changes may evolve too slowly to reveal relationships among closely related taxa. In general, the dynamics of sequence change in silent and non-silent positions in protein coding genes result in homoplasy and lack of resolution, respectively. We suggest that the combination of nucleic acid and the translated amino acid coded character states into the same data matrix for phylogenetic analysis addresses some of the problems caused by the rapid change of silent nucleotide positions and overall slow rate of change of non-silent nucleotide positions and slowly changing amino acid positions. One major theoretical problem with this approach is the apparent non-independence of the two sources of characters. However, there are at least three possible outcomes when comparing protein coding nucleic acid sequences with their translated amino acids in a phylogenetic context on a codon by codon basis. First, the two character sets for a codon may be entirely congruent with respect to the information they convey about the relationships of a certain set of taxa. Second, one character set may display no information concerning a phylogenetic hypothesis while the other character set may impart information to a hypothesis. These two possibilities are cases of non-independence, however, we argue that congruence in such cases can be thought of as increasing the weight of the particular phylogenetic hypothesis that is supported by those characters. In the third case, the two sources of character information for a particular codon may be entirely incongruent with respect to phylogenetic hypotheses concerning the taxa examined. In this last case the two character sets are independent in that information from neither can predict the character states of the other. Examples of these possibilities are discussed and the general applicability of combining these two sources of information for protein coding genes is presented using sequences from the homeobox region of 46 homeobox genes from *Drosophila melanogaster* to develop a hypothesis of genealogical relationship of these genes in this large multigene family.

© 1996 The Willi Hennig Society

Introduction

When molecular information for protein coding regions is used in phylogenetic analysis either nucleic acid or amino acid sequences are used as the source of information for character coding. These two types of sequence are simply two different ways of character coding the same information. In many studies, variations on these two themes are employed, such as the use of character transformation

0748-3007/96/010065+18/$18.00/0 © 1996 The Willi Hennig Society