M. Feldman

ABSTRACT FROM: German-Israeli Symposium Computer Science, Aspects of
Molecular Biology, Tel Aviv University School of Mathematics
26-28 November, 1995.

ECOLOGICAL BASIS OF MOLECULAR SEQUENCE DIVERGENCE: *RUBISCO* LARGE SUBUNIT GENE AS AN EXAMPLE

Valery A. Ugrinovskii, Abraham B. Korol, Eviatar Nevo
Institute of Evolution, University of Haifa
Mount Carmel, Haifa 31905, Israel,
E-mail: rabi307@uvm.haifa.ac.il

This paper reports some results that bring into light the role of ecological factors in the primary organization of molecular genetic sequences at the DNA level. It is shown that some structural aspects of sequence organization are correlated with the ecological range, geographical distribution and some other life history traits.

As a tool to answer the questions on the role of ecological factors in sequence organization, we exploit the so called indices of the relative abundance of oligonucleotides. This abundance (*e.g.*, for di-, tri- and tetranucleotide words) was reported in the literature as a reasonable basis for distance measure and multidimensional comparisons and classifications of sequences from different groups of organisms. Based on the abundance indices, some distance measures were also proposed to estimate the closeness between different sequences. We discuss here the application of these measures to problems of classification with respect to ecological traits.

The following example demonstrates the efficiency of this approach for the ecological grouping of several hundreds of plant species based on the chloroplast ***rubisco*** gene (*rbcL*). The question was whether variations in the life history traits affect the *rbcL* oligonucleotide organization. 716 plant species were grouped according to their life form: ***herbaceous***, ***nonherbaceous*** (bushes and trees), and ***liana***. The learning group included ***herbaceous*** and ***nonherbaceous*** subgroups. Alternate small groups of ***bushes*** and ***trees*** were involved to verify the classification. The following are the results of classification using the dinucleotide abundance indices (as a method for grouping, the stepwise discriminant analysis has been applied).

Classification matrix

Group	Correct (%)	Cases classified into group	
		Herbaceous	Nonerbaceous
Herbaceous	67.1	241	118
Nonherbaceous	79.0	44	166
Bushes	–	1	7
Trees	–	0	5
Liana	–	13	14

Similar results have been obtained for classification with respect to geographic life zones. It is noteworthy, that in spite of its relatively small size (about 1430 bps), the *rbcL* sequence has allowed also to get a high precision (not less than 85–90%) taxonomic classification.