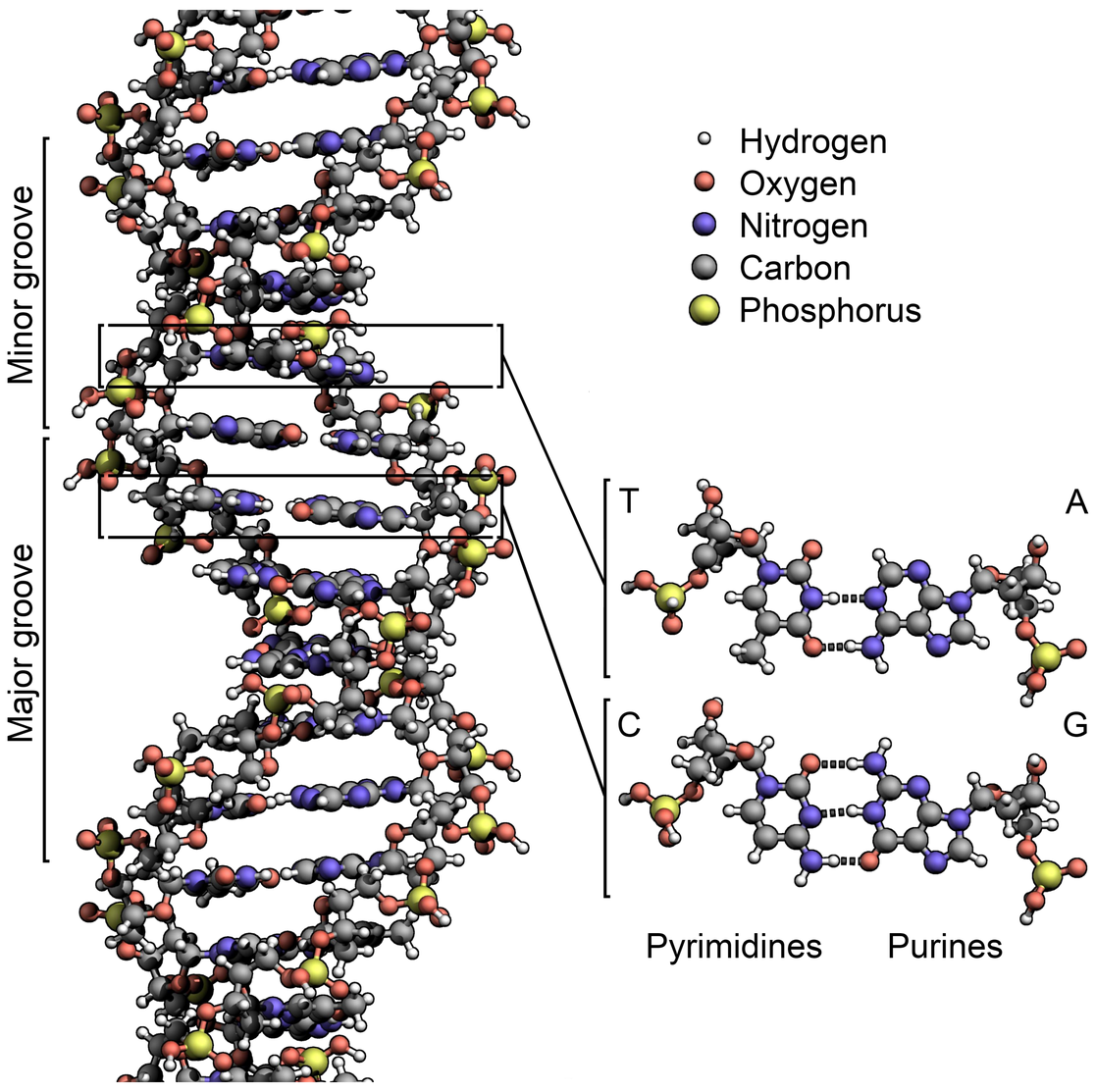
Evolution of nucleotide sequences over passing time

ABSTRACT

All DNA sequences contain four types of nucleotides, which in turn hold all genetic information inherited by an organism. However, DNA can mutate while replicating itself, which means that it is possible to lose a number of nucleotides and/or to gain different fragments of the original sequence; in other words, the initial DNA sequence can differ from its duplicate. Knowing this, DNA sequence over passing time can be depicted as a discrete-time homogeneous Markov chain, while sequence evolution in space can be described as an action which depicts new element addition to the sequence. Theoretically, evolution in space simulates DNA sequence formation. In the stationary case, the distribution of a random sequence does not depend on the fixed time moment. It is hard to find any data regarding DNA sequence evolution in space – usually, only one sequence can be found. It is possible to reconstruct the transition matrix or the properties of that matrix from the stationary distribution of the Markov chain during the evolution over passing the time, yet this problem is ill-posed. In general, said the problem has a lot of solutions which could be found only by using some additional assumptions and regularization methods. However, the solution could be found more easily using the local balance equation if the DNA sequence is reversed and the transition matrix only depends on a relatively small number of unknown parameters.

DATA AND METHODS



Nucleotide triplets of the sequences are analyzed: every second nucleotide from the sequence is selected, which is the middle value of the triplet. Nucleotides are neighboring to it from the left and right. The lateral nucleotide of two adjacent triplets is common: for one, it is a neighbor from the left, for the other, it is from the right.
Each nucleotide is characterized by two properties:

(p) whether it is pyrimidine or purine,

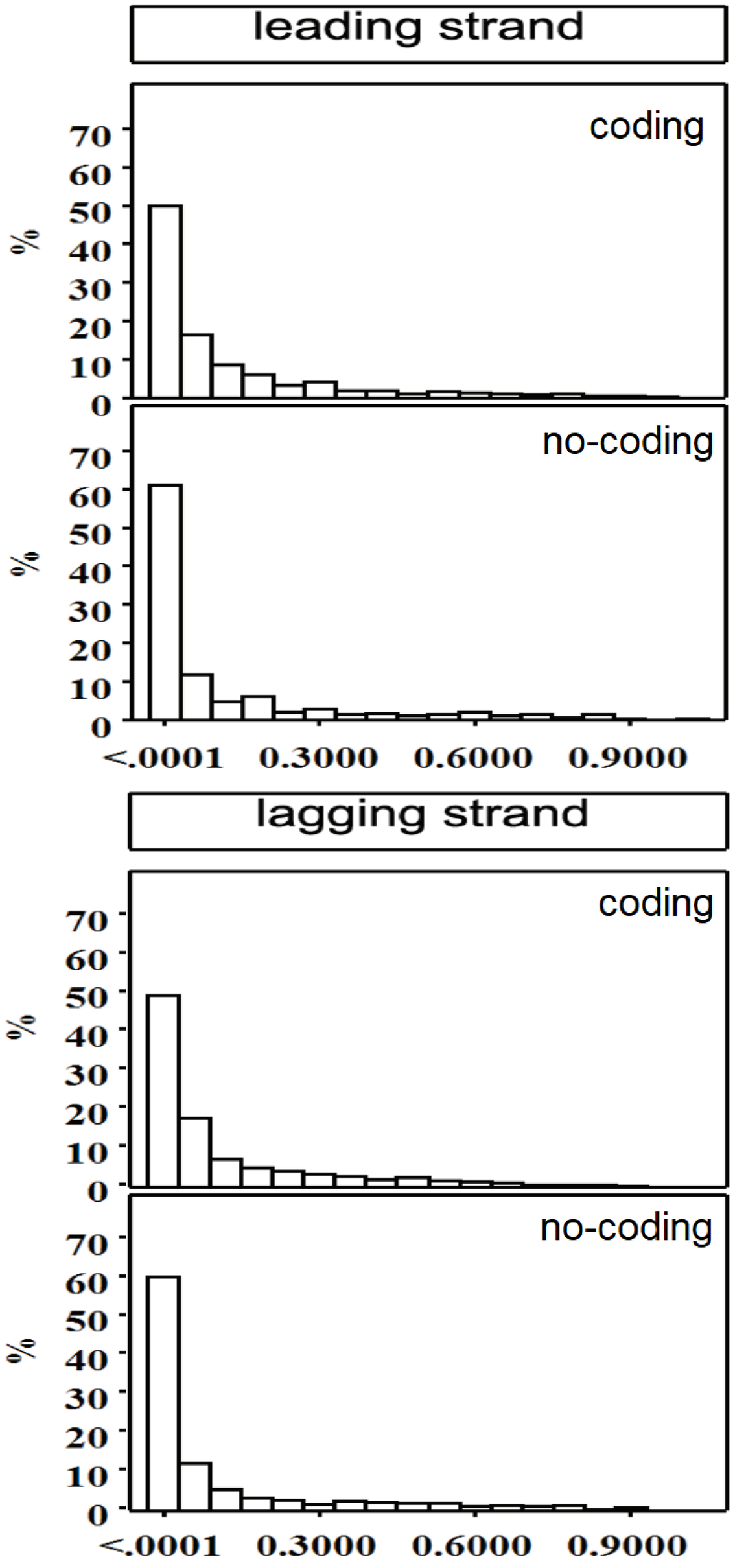
(j) whether it has two or three joints.

This makes it possible to examine the properties (p), (j) of each neighboring nucleotide and the influence of their interaction (p*j) on the middle nucleotide. Evaluation is performed on all DNA primary and secondary sequences, both coding and non-coding, and then the same model is applied to each sequence separately.
First, a generalized logit model was estimated to test the first-order Markov property for all, both coding and non-coding DNA primary and secondary strand sequences.
Whether a chain is first-order Markov is determined by the interaction coefficients of the left and right nucleotides. Their significance is determined by Likelihood Ratio statistics, which compares the evaluated model with the saturated model.

RESULTS

Bacteria *Escherichia coli*

Distribution of p-values of Likelihood Ratio statistic for testing the Markovity hypothesis



CONCLUSION

In almost all cases, the hypothesis of no differences between the estimated and saturated models was rejected. Thus, for both coding and non-coding DNA sequences (in the specific case of the bacterium Escherichia coli), the first-order Markov property does not hold.

AUTHORS

K. Jablonskaitė, T. Ruzgas

Department of Applied Mathematics, Kaunas University of Technology