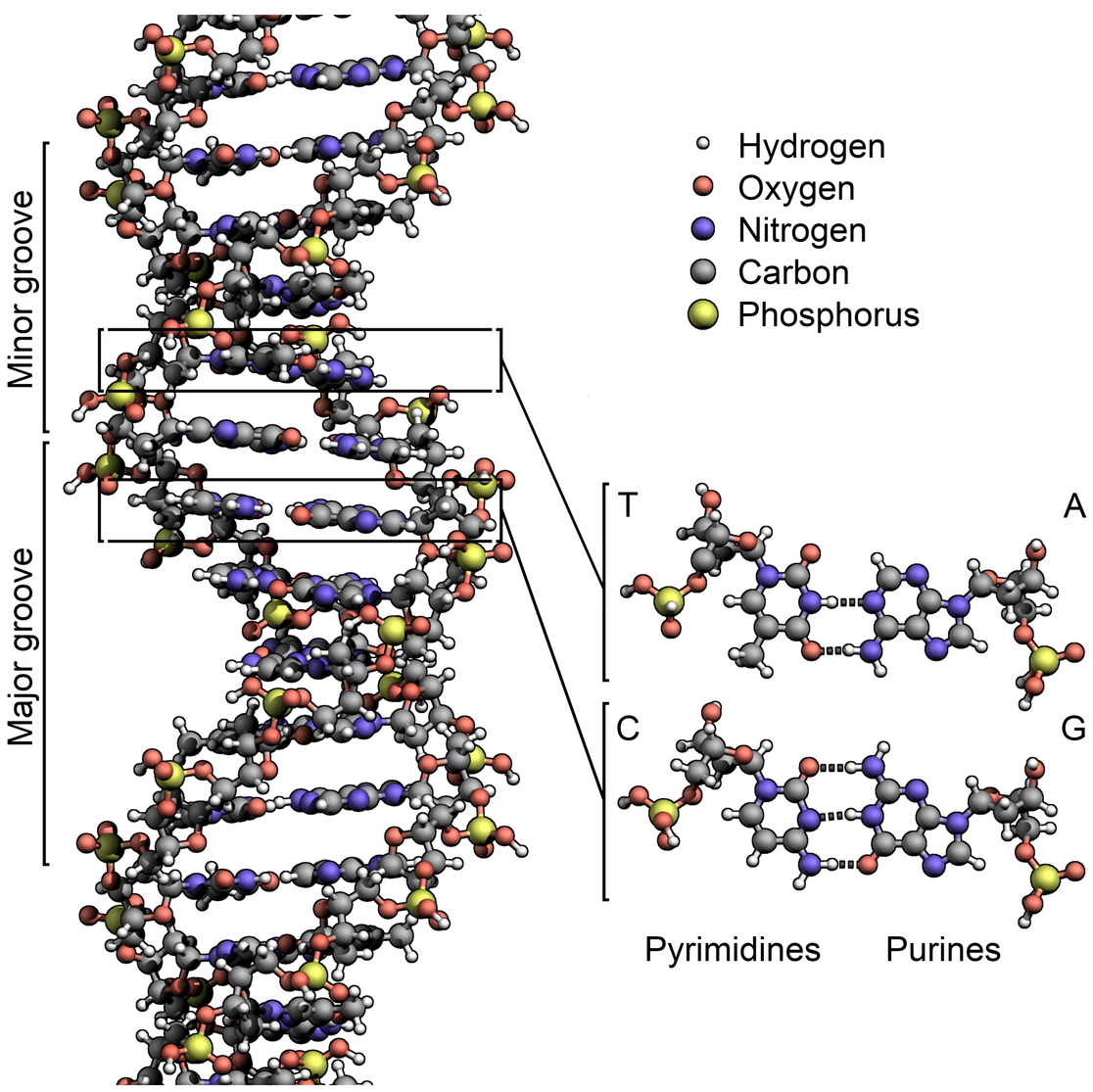
Evolutionary Model for Non-coding Nucleotide Sequences

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 sequence are studied: every second nucleotide is selected from the sequence which is the average value of the triplet. Nucleotides stand beside it from the left and right side. One nucleotide of two adjacent triplets is common: for one it is a neighbor from the left, for the other – a neighbor from the right.
Each nucleotide has two properties:
(p) whether it is a pyrimidine or a purine,
(b) whether it has two or three bonds.
This allows to study the properties (p) and (b) of each adjacent nucleotide and the influence of their interaction (p*b) on the middle nucleotide. The 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 for a first-order Markov property across all, both coding and non-coding, primary and secondary DNA sequences.
Whether a chain is a first-order Markov chain 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 full (saturated) model.

RESULTS

Bacteria *Escherichia coli*

Distribution of p-values of Likelihood ratio statistic for testing the Markovity hypothesis are shown in Figures 1, 2 and 3.

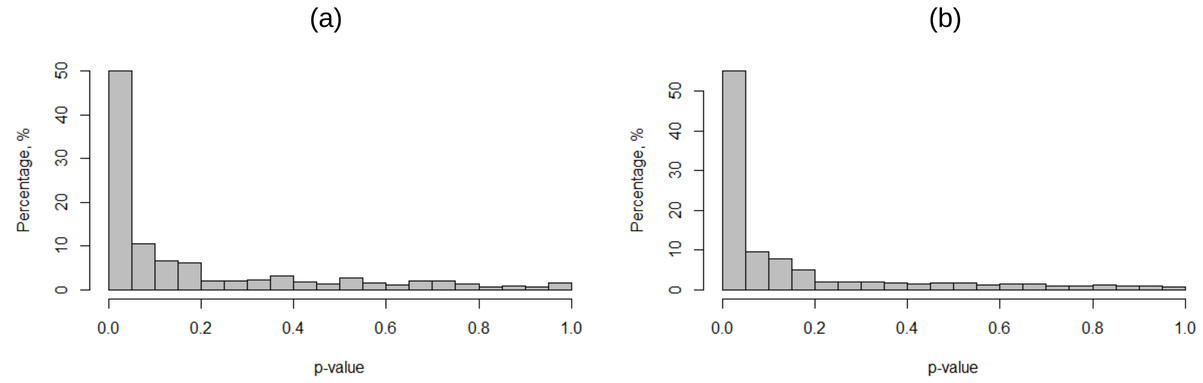


Figure 1 Nucleotide dependency on its neighbor from the left side; (a) shows dependency in non-coding part of the sequence, (b) shows dependency in coding part of the sequence

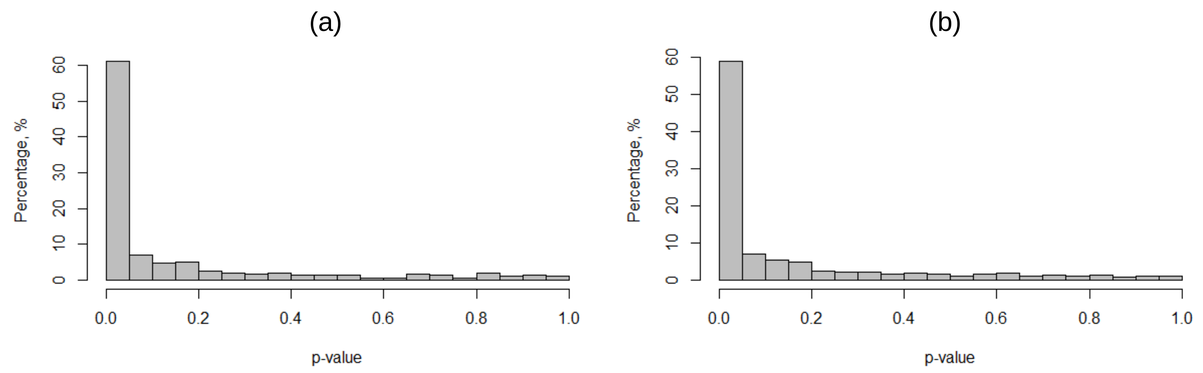


Figure 2 Nucleotide dependency on its neighbor from the right side; (a) shows dependency in non-coding part of the sequence, (b) shows dependency in coding part of the sequence

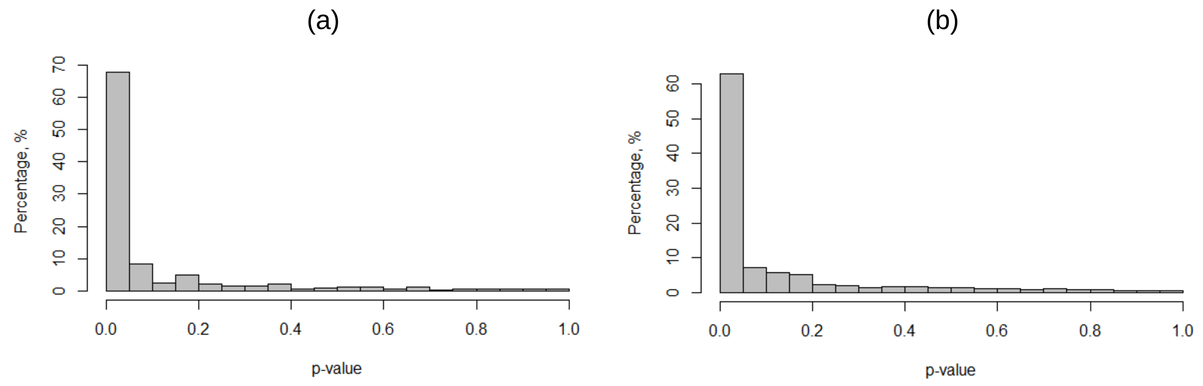


Figure 3 Nucleotide dependency on its neighbor from both sides; (a) shows dependency in non-coding part of the sequence, (b) shows dependency in coding part of the sequence

P-values for continuous DNA sequence (not splitted into coding and non-coding parts and splitted into them and combined) for every dependency (left, right and both sides) are shown in Table 1.

Dependency / DNA sequence	Left side	Right side	Both sides
Whole (continuous)	< 0.01	< 0.01	< 0.01
Non-coding	< 0.01	< 0.01	< 0.01
Coding	< 0.01	< 0.01	< 0.01

Table 1 Nucleotide dependency on its neighbor from left, right and both sides; p-value calculated for the whole sequence, non-coding and coding parts separately

All calculations were done for both primary (main) and secondary (complementary) strands; results for the secondary strand are not shown, as they are almost identical to the calculations for the primary strand shown above (in Figures 1, 2 and 3 and Table 1). This is mostly because the structure of secondary DNA strand is identical to the main one, only with the bases changed (example is shown below).

It is also worth mentioning that some errors of calculation might have occurred, as the dataset used is a shotgun sequence – that means that before continuous DNA strand is created, it is broken apart and many sequence reads are generated from the separate DNA pieces. Naturally, this means that there might be some overlaps, repeated parts or missing parts altogether.

CONCLUSION

In all cases, the hypothesis of no differences between the estimated and full 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

M. Frolovaitė[1], E. Lebon[2], T. Ruzgas[1]

[1] Kaunas University of Technology, Lithuania
[2] University of Angers, France