Małgorzata Wańczyk, Paweł Błażej and Paweł Mackiewicz
Faculty of Biotechnology, Department of Genomics, University of Wrocław

Comparison of two algorithms based on Markov chains applied in recognition of protein coding sequences in prokaryotes

Methods which are based on the theory of Markov chains are one of the most commonly used in the recognition of protein coding sequences. However, they require big learning sets to thoroughly fill up transition probability matrices describing a dependence between nucleotides in analyzed sequences. In this paper we tested the efficiency (in term of true positive rate) of Markov chain methods depending on the size of learning set and the order of these models. Besides the 'classical' GeneMark algorithm, which used three-periodic non-homogeneous Markov chain, we also studied an algorithm (called PMC) that considers six independent homogeneous Markov chains to describe transition between nucleotides for each of three codon positions in two DNA strands separately. The PMC algorithm of different chain orders is more stable and generally outperformed the GeneMark algorithm. The PMC algorithm seems promising in recognition of protein coding genes, especially when the learning sets are small. The application of the PMC algorithm showed that substantial fraction of annotated ORFs in the analyzed *Escherichia coli* 536 genome are likely false and do not code proteins.