APPLICATIONES MATHEMATICAE
40,4 (2013), pp. 447–457

Małgorzata Grabińska, Paweł Błażej and
Paweł Mackiewicz (Wrocław)

TWO ALGORITHMS BASED ON MARKOV CHAINS AND THEIR APPLICATION TO RECOGNITION OF PROTEIN CODING GENES IN PROKARYOTIC GENOMES

Abstract. Methods based on the theory of Markov chains are most commonly used in the recognition of protein coding sequences. However, they require big learning sets to fill up all elements in transition probability matrices describing dependence between nucleotides in the analyzed sequences. Moreover, gene prediction is strongly influenced by the nucleotide bias measured by e.g. G+C content. In this paper we compare two methods: (i) the classical GeneMark algorithm, which uses a three-periodic non-homogeneous Markov chain, and (ii) an algorithm called PMC that considers six independent homogeneous Markov chains to describe transition between nucleotides separately for each of three codon positions in two DNA strands. We have tested the efficiency (in terms of true positive rate) of these two Markov chain methods for the model bacterial genome of *Escherichia coli* depending on the size of the learning set, uncertainty of ORFs' function annotation, and model order of these algorithms. We have also applied the methods with different model orders for 163 prokaryotic genomes that covered a wide range of G+C content. The PMC algorithm of different chain orders turns out to be more stable in comparison to the GeneMark algorithm. The PMC also outperforms the GM algorithm giving a higher fraction of coding sequences in the tested set of annotated genes. Moreover, it requires much smaller learning sets than GM to work properly.

1. Introduction. DNA contains genetic information about coded proteins and consists of two strands, which are long chains of four simpler units

2010 *Mathematics Subject Classification*: Primary 92D20, 60J10; Secondary 60J20, 62P10.
Key words and phrases: Markov chains, DNA sequence, recognition of protein coding sequences, open reading frame.

DOI: 10.4064/am40-4-5

called nucleotides (adenine, guanine, cytosine, and thymine). Protein coding sequences (genes) are not random assortments of nucleotides but consist of continuous stretches of nucleotides, arranged in triplets called codons. Each codon encodes either (i) one of 20 amino acids which are linearly arranged elements of proteins, or (ii) one of three stop translation signals informing about the ending of protein biosynthesis. These coding requirements together with mutational pressure influence a specific nucleotide composition and usage of codons, hexamers or other sequence 'words' in protein coding sequences. This specific composition can be successfully used in gene recognition in genomic sequences [11], [10]. The first step in gene identification is scanning the genomic sequence for *open reading frames* (ORFs), which are sequences beginning with a start translation codon and ending with a stop translation codon. However, not all ORFs are coding, therefore gene prediction in prokaryotes (including archaea and bacteria) consists mainly in discriminating the real coding ORFs from random or spurious ones. This recognition uses characteristic statistical patterns in nucleotide composition resulting from coding capacity.

Among many approaches in gene recognition (see for review [1], [16]), one of the most commonly used are Markov chain models describing the gene sequence by transitions from one sequence 'word' (state) to another [5], [6], [13], [17]. In spite of advance in these methods, they still need huge learning sets of known coding sequences in their training stage to estimate all elements in matrices describing dependencies between the sequence words.

It is also well known that the efficiency of gene finding algorithms is strongly dependent on the composition of the analyzed genome described for example by their G+C content ([2], [18]). Skovgaard et al. [18] observed that abundance of long non-coding ORFs is correlated with the G+C level. This clearly results from the nucleotide composition of stop codons (TAA, TAG, and TGA) which are poor in G+C and A+T-rich. Therefore these triplets occur with low frequency in genomes with high G+C content. Consequently, relatively long random open reading frames can be easily generated in such genomes.

Here we compare two Markov chain approaches in the context of their relationship with the size of the learning set, the order of chains, the functional annotations of ORFs, and the G+C content using 163 completely sequenced prokaryotic genomes. Besides the classical GeneMark algorithm [6] considering sequence words as consecutive states, we also analyze a model regarding particular three codon positions separately [3], [4]. This second approach, called the PMC algorithm, uses specific correlations in nucleotide composition observed in the first, second, and third codon positions [10], [8] and does not require a high chain order to work properly. The algorithm described in [4] is a more advanced version of that presented in [3]. The extended algo-

rithm includes 216 positional patterns whereas the simpler version takes into account 27 positional patterns.

2. Material and methods

2.1. General characteristics of tested algorithms. The algorithms under consideration are an example of supervised learning because they consist of two stages: the learning step and the testing step. During the learning step all necessary parameters are computed from a learning set which usually consists of selected protein coding sequences with annotated function. These algorithms are based on the assumption that every DNA sequence is a realization of a suitable Markov chain. In the case of the classical GeneMark algorithm we assume that every DNA sequence is realized by a three-periodic non-homogeneous Markov chain, whereas in the PMC algorithm we consider six independent homogeneous Markov chains to describe transition between nucleotides for each of three codon positions (three chains for the direct DNA strand and three for the complementary DNA strand).

2.2. Learning and tested sets. First, we have tested these two algorithms on 2773 open reading frames with annotated function coded in the *Escherichia coli* 536 genome. This set was divided into two parts:

1. the learning set consisted of 1000 ORFs,
2. the tested set contained the remaining 1773 ORFs.

To test the efficiency of these algorithms in dependence on the size of the learning set, we chose randomly from the learning set subsets consisting of an increasing number of ORFs $(100, 200, \ldots, 900)$. These subsets were also used to assess the coding signal in two additional tested subsets including:

1. 1309 ORFs annotated as hypothetical (i.e. without assigned function),
2. 578 ORFs with ascribed putative function.

All the tests were repeated 20 times and the resulting values were averaged. To further assess the effectiveness of the tested algorithms, we considered the set of 163 prokaryotic genomes that had over 2000 sequences annotated as protein coding. For every genome we prepared learning subsets containing 1000 protein coding sequences, whereas the remaining sequences formed the tested subsets. Model orders of $h = 1, 2, 3, 4$ were used in this analysis. We checked the coding prediction for each genome for the two algorithms and different model orders. The analyzed genomes were selected to represent a wide range of G+C *content* (from 28% to 75%), i.e. the fraction of guanine and cytosine in the whole genome sequence. They were also used to examine the influence of the nucleotide bias on coding signal prediction. The

genome sequences and their annotations were downloaded from GenBank (www.ncbi.nlm.nih.gov).

2.3. The PMC algorithm. Let $S = \{s_1, \ldots s_n\}$ (where $n = 3k,\ k \in \mathbb{N}$) be a DNA sequence. We assume that every protein coding sequence can be modeled by six independent homogeneous Markov chains [4]. During the learning step we calculate, for a given model order h and every position i in codon:

(i) the initial probabilities $P(S_i^h)$ of h nucleotides situated in the same codon position i,
(ii) the transition probability matrices $(M_1, \ldots, M_6)$.

The matrices M_1, M_2, M_3 concern the direct DNA strand of ORFs in the learning set whereas M_4, M_5, M_6 are based on the complementary strand and are useful for a model of 'shadow' coding regions. In the next step we use these processes to obtain the positional pattern frequencies distribution, which is used to detect a coding signal in the analyzed ORFs.

Definition 1. The *positional pattern frequencies distribution* for a given protein coding sequence is the distribution of three dimensional vectors which are obtained in the following way:

(i) The DNA sequence is divided into moving windows with a fixed length (e.g. 96 nt) and a fixed window shift (e.g. 12 nt).
(ii) For each window, a vector of digits (d_1, d_2, d_3) is determined, by using the maximum likelihood approach, in the following way:
  (a) the probabilities $P_{M_i},\ i = 1, \ldots, 6$, are calculated for each of three codon positions by using the previously trained matrices;
  (b) if $P_{M_j} = \max(P_{M_i} :\ i = 1, \ldots, 6)$ (for the fixed codon position), then $d_i = j$ and the positional pattern (d_1, d_2, d_3) is obtained.
(iiii) Finally, this procedure is applied to obtain the frequency for each positional pattern calculated from all the analyzed windows by scanning the learning set sequences in all their reading frames.

It is evident that there are 216 possible positional patterns (Fig. 1). As a result of the training step we obtain six transition probability matrices and six distributions of the positional pattern frequencies.

During the testing step we divide the sequence S of a given ORF into windows and by using transition probability matrices we obtain a positional pattern for every window. It is obvious that for a given reading frame the distribution of the positional pattern could be treated as the conditional probability of obtaining a given sequence S under the condition that S is coding $(P(S \,|\, \mathrm{fr}),\ \mathrm{fr} = 1, 2, \ldots, 6)$. In addition, we assume a non-coding reference set with uniform distribution of positional patterns, which is expressed

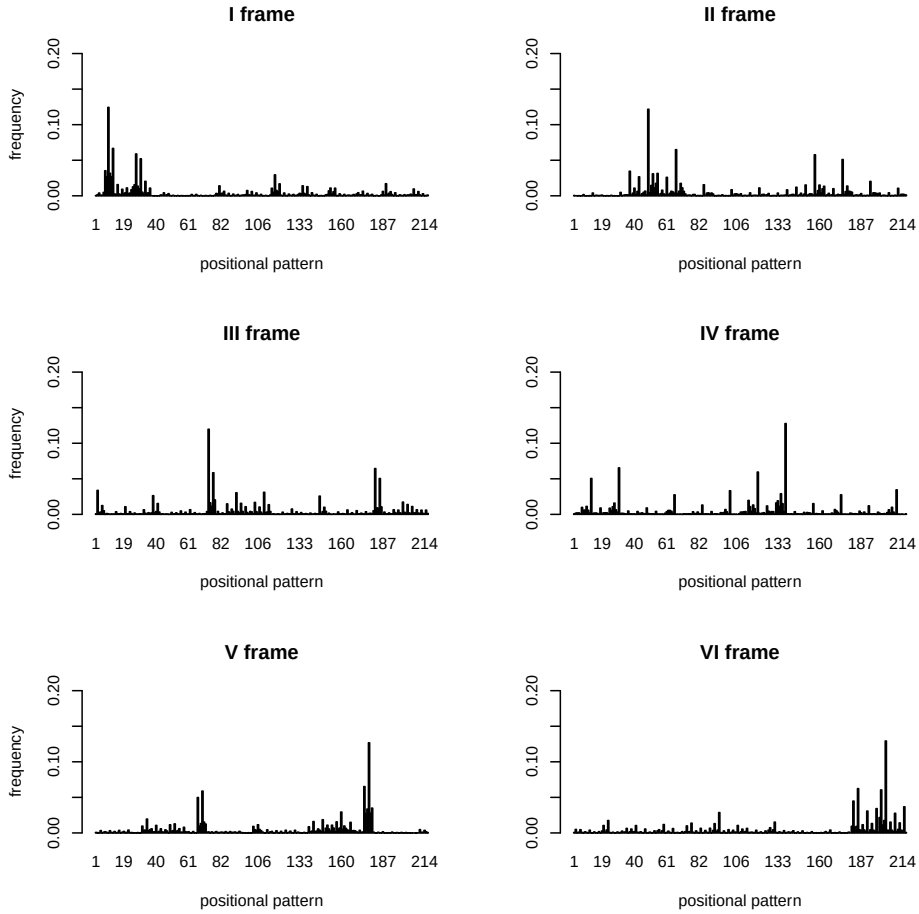

Fig. 1. Histograms of positional pattern frequencies computed for the training set of ORFs with annotated function from the *Escherichia coli* 536 genome for each of six reading frames.

by $P_7 = 1/216$. Next we use the Bayes theorem to obtain the a posteriori probabilities ($P(\mathrm{fr} = i \,|\, S),\ i = 1, \ldots, 7$) according to the equation

$$P(\mathrm{fr} = i \,|\, S) = \frac{P(S \,|\, \mathrm{fr} = i)}{\sum_{j=1}^{6} P(S \,|\, \mathrm{fr} = j) + P_7}, \tag{2.1}$$

where the a priori probability is $P(\mathrm{fr} = i) = 1/7,\ i = 1, \ldots, 7$. We say that a given sequence is *in coding frame* i when the a posteriori probability $P(\mathrm{fr} = i \,|\, S)$ achieves its highest value.

The algorithm can be used without any modifications in recognition of protein coding sequences in eukaryotic genomes that have uninterrupted genes, i.e. without introns. This method can also be applied in the case

of interrupted genes, provided information about boundary between exons and introns is included.

2.4. The GeneMark algorithm. We have used the main part of the algorithm described in [6], [5]. This algorithm assumes that every protein coding sequence is treated as a realization of a three-periodic non-homogeneous Markov chain for the given model order h. The non-homogeneous periodic Markov chain model is used because it describes more precisely the three-step periodicity of the coding sequence [12]. During the learning step the initial probabilities $(P(s_i^h),\ i = 1, \ldots, 6)$ and also six transition probability matrices $(M_i :\ i = 1, \ldots, 6)$ are calculated. The non-protein coding DNA sequences are described by a homogeneous Markov chain (with transition probability matrix M_7). During the testing step for a given DNA sequence S the conditional probability of being in a given reading frame under the protein coding condition, $P(S|M_i)$ for $i = 1, \ldots, 6$, is calculated and the a posteriori probability (by using the Bayes theorem) is computed according to

$$P(M_i \,|\, S) = \frac{P(S \,|\, M_i)P(M_i)}{\sum_{j=1}^{7} P(S \,|\, M_j)P(M_j)}, \tag{2.2}$$

where

$$P(M_i) = \begin{cases} 1/12, & i = 1, \ldots, 6, \\ 1/2, & i = 7, \end{cases}$$

is the a priori probability of each of the seven events specified by M_i. We say that a given sequence is *in coding frame* i when the a posteriori probability $P(M_i \,|\, S)$ achieves its highest value.

3. Results and discussion

3.1. Relationship between coding signal prediction and the size of the learning set. To evaluate the efficiency of the two methods based on Markov chains, we have measured the true positive rate (i.e. sensitivity, TPR) for different model orders h. The results for the two algorithms applied to the *Escherichia coli* 536 genome are compared in Fig. 2. The true positive rate for the PMC algorithm of different model orders shows much less variation than that for the GeneMark algorithm. The TPR values for the PMC algorithm also tend to increase with the model order, a feature not observed for the GeneMark algorithm. The TPR value for the PMC algorithm ranges from 0.93 for model order $h = 1$ to 0.945 for $h = 3$. Interestingly, the TPR for the GeneMark algorithm for $h = 1$ is about 0.76, for $h = 3$ it reaches 0.91, and it achieves 0.94 only for $h = 2$. The very low TPR value for $h = 1$ is probably due to the simplicity of this model whereas the lower value for $h = 3$ than for $h = 2$ results probably from too low size of the learning set

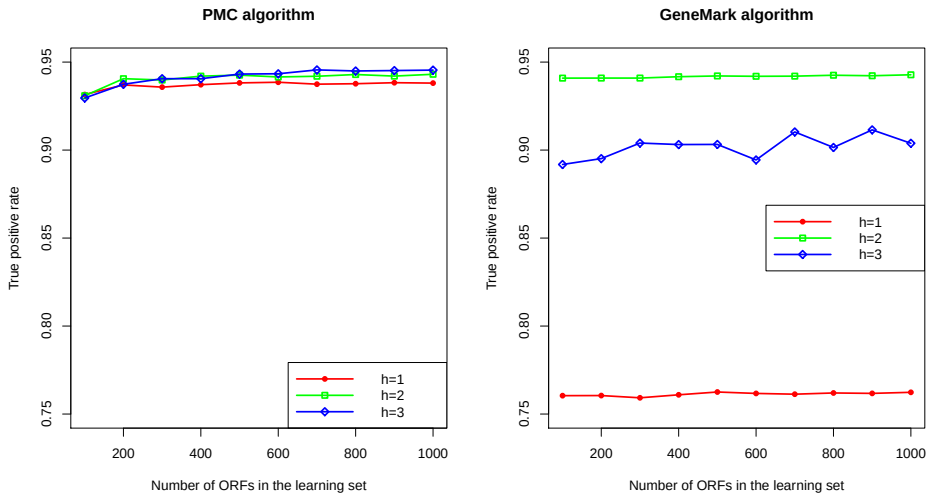

Fig. 2. Relationship between the true positive rate and the size of the learning set for PMC and GeneMark algorithms

for much larger transition probability matrices in the more complex model. For all model orders in the case of PMC, and for $h = 3$ in the case of the GeneMark algorithm the TPR increases slightly with the number of ORFs in the learning set.

The results clearly indicate that the PMC algorithm is much less dependent on the model order and is more stable in gene recognition than the GeneMark algorithm. The PMC algorithm considers three codon positions separately, which allows estimating the transition probability matrices effectively even for a low chain order and quite small learning sets because this algorithm retains information on dependence between nucleotides in DNA sequences on relatively long distances.

3.2. Coding signal prediction for different sets of ORFs. Figs. 3 and 4 enable us to compare the coding signal prediction by the PMC and GM algorithms in three sets of ORFs identified in the *Escherichia coli* 536 genome:

(i) with annotated function,
(ii) with putative function,
(iii) hypothetical, i.e. without assigned function.

We observe that for both algorithms the fraction of ORFs recognized as protein coding decreases with the uncertainty of function annotation, For $h = 3$, the highest TPR (about 0.94 in the case of PMC and 0.9 in the case of GM) is for the first set of ORFs, which seems obvious because these two

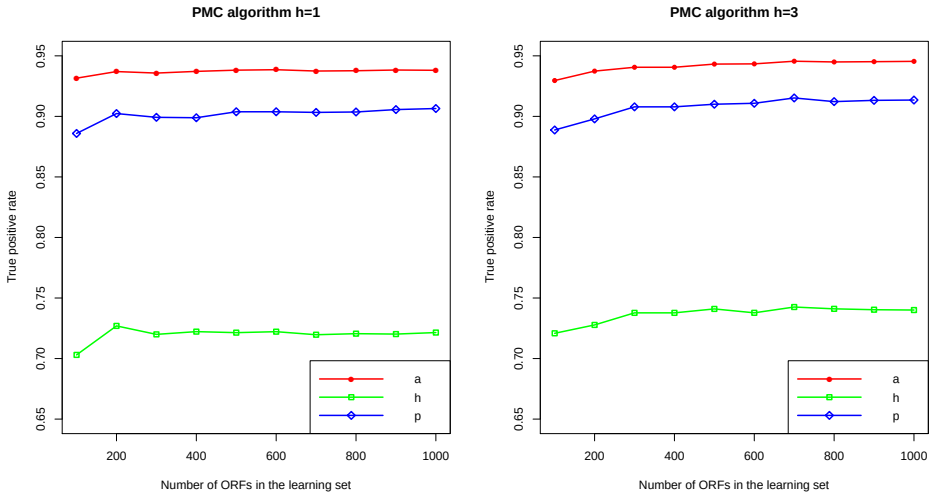

Fig. 3. The true positive rate of the PMC algorithm for model orders $h = 1$ and $h = 3$, for three sets of ORFs: with annotated function (a), without assigned function, i.e. hypothetical (h), and with putative function (p).

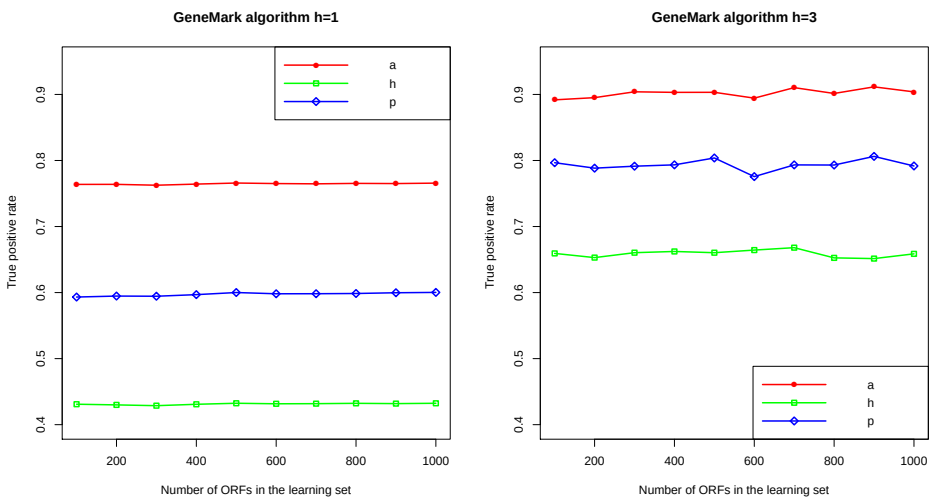

Fig. 4. True positive rate of the GM algorithm for model orders $h = 1$ and $h = 3$, for three sets of ORFs: with annotated function (a), without assigned function, i.e. hypothetical (h), and with putative function (p).

algorithms were trained on sequences with known functions. However, the TPR values diminish (to about 0.90 in the case of PMC and 0.8 in the case of GM) for putative ORFs and are the lowest (about 0.74 in the case of PMC and about 0.65 in the case of GM) for the hypothetical ORFs. The coding

signal prediction by the PMC algorithm for $h = 1$ is very similar to that for $h = 3$. In contrast, the GM algorithm gives much lower TPR values for $h = 1$ than for $h = 3$ for all the three sets of ORFs.

The lower fraction of ORFs predicted as coding in the last two sets of sequences may result from the inefficiency of the algorithms applied or non-representativeness of known annotated protein coding sequences. On the other hand, these results can also indicate that substantial fractions of these ORFs are in fact spurious frames which probably do not code proteins, as was found for other genomes by applying other methods [8], [9]. Such non-coding ORFs can be generated by real coding genes in alternative reading frames [14], [15] as a result of peculiar features of the genetic code and properties of protein coding sequences [7].

3.3. Influence of G+C content on coding prediction. Fig. 5 illustrates the stability of coding prediction by the PMC and GM algorithms in dependence on the G+C level for 163 prokaryotic genomes. We have consid-

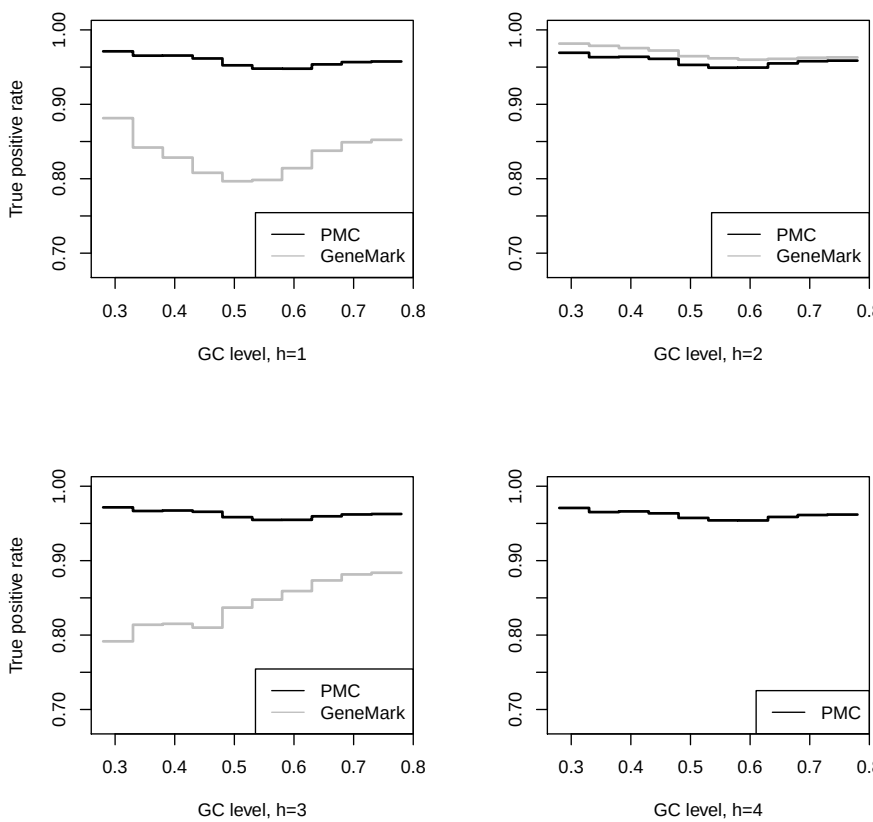

Fig. 5. Relationship of the true positive rates computed in the GM and PMC algorithms to the G+C level.

ered four model orders in this analysis. Interestingly, the fraction of correctly predicted coding sequences by the PMC algorithm is very weakly influenced by the G+C composition of the genomes under study. Moreover, this method gives nearly constant TPR values usually higher than 0.95. On the other hand, predictions made by the GM algorithm strongly depend on the G+C level and the model order.

This algorithm works comparably with PMC only in the case of $h = 2$, whereas it is much worse for other orders for which it gives TPR values much lower than 0.90. The GM method with $h = 1$ shows the lowest efficiency for genomes with about $50 - 60\%$ G+C content whereas for $h = 3$ the TPR is the lowest for G+C-poor genomes and increases linearly with the G+C level. The last result is probably influenced by not fully determined transition probability matrices used in this model order. Although the selection genomes code more than 2000 protein genes with assigned function, the learning set turned out to be too small to entirely fill up all elements in the matrices constructed. Consequently, it was not possible to apply the GM algorithm for $h = 4$, and therefore only the results for the PMC method are presented in Fig. 5.

References

[1] R. K. Azad, *Genes in prokaryotic genomes an their computational prediction*, in: Computational Methods for Understanding Bacterial and Archeal Genomes, Series of Advances in Bioinformatics and Computational Biology, Y. Xu and J. P. Gogarten (eds.), Vol. VII, College Press, 2003, 39–74.

[2] R. K. Azad and M. Borodovsky, *Effects of choice of DNA sequence model structure on gene identification accuracy*, Bioinformatics 20 (2004), 993–1005.

[3] P. Błażej, P. Mackiewicz and S. Cebrat, *Using genetic coding wisdom for recognizing protein coding sequences*, in: Proc. 2010 International Conference on Bioinformatics & Computational Biology BIOCOMP 2010, Las Vegas, Vol. I, 2010, 302–305.

[4] P. Błażej, P. Mackiewicz and S. Cebrat, *Algorithm for finding coding signal using homogeneous Markov chains independently for three codon positions*, in: Proc. 2011 Int. Conference on Bioinformatics and Computational Biology (ICBCB 2011) (Haikou, 2011), 20–24.

[5] M. Borodovsky, Y. A. Sprizhitskii, E. I. Golovanov and A. A. Aleksandrov, *Statistical patterns in primary structures of the functional regions of the genome in Escherichia coli*, Molecular Biol. 20 (1986), 826–840, 1144–1150.

[6] M. Borodovsky and J. McIninch, *GeneMark: parallel gene recognition for both DNA strands*, Comput. Chem. 17 (1993), 122–133.

[7] S. Cebrat and M. R. Dudek, *Generation of overlapping reading frames*, Trends in Genetics 12 (1996), 12.

[8] S. Cebrat, M. R. Dudek, P. Mackiewicz, M. Kowalczuk and M. Fita, *Asymmetry of coding versus non-coding strand sequences of different genomes*, Microbial and Comparative Genomics 2 (1997), 259–268.

[9] S. Cebrat, M. R. Dudek and P. Mackiewicz, *Is there any mystery of ORPHANs?*, J. Appl. Genetics 38 (1997), 365–372.

[10] S. Cebrat, M. R. Dudek and P. Mackiewicz, *Sequence asymmetry as a parameter indicating coding sequence in Saccharomyces cerevisiae genome*, Theory in Biosci. 117 (1998), 78–89.

[11] J. W. Fickett and C. S. Tung, *Assessment of protein coding measures*, Nucleic Acids Res. 20 (1992), 6441–6450.

[12] J. Kleffe and M. Borodovsky, *First and second moment of counts of words in random texts generated by Markov chains*, Comput. Appl. Biosci. 8 (1992), 433–441.

[13] A. Krogh, I. S. Mian and D. Hausler, *A hidden Markov model that finds genes in E. coli DNA*, Nucleic Acids Res. 22 (1994), 4768–4778.

[14] P. Mackiewicz, M. Kowalczuk, A. Gierlik, M. R. Dudek and S. Cebrat, *Origin and properties of noncoding ORFs in the yeast genome*, Nucleic Acids Res. 27 (1999), 3503–3509.

[15] P. Mackiewicz, M. Kowalczuk, A. Gierlik, D. Szczepanik, A. Nowicka, M. R. Dudek and S. Cebrat, *No mystery of ORFans in genomics—generation of ORFans in the antisense of coding sequences*, in: Proc. Second International Conference on Bioinformatics of Genome Regulation and Structure, BGRS'2000 (Novosibirsk, 2000), 38–41.

[16] W. H. Majoros, *Methods for Computational Gene Prediction*, Cambridge Univ. Press, 2007.

[17] S. L. Salzberg, A. L. Delcher, S. Kasif and O. White, *Microbial gene identification using interpolated Markov models*, Nucleic Acids Res. 23 (1998), 544–548.

[18] M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery and A. Krogh, *On the total number of genes and their length distribution in complete microbial genomes*, Trends in Genetics 17 (2001), 425–428.

Małgorzata Grabińska, Paweł Błażej, Paweł Mackiewicz
Department of Genomics
Faculty of Biotechnology
University of Wrocław
Przybyszewskiego 63/77
51-148 Wrocław, Poland
E-mail: ewan@smorfland.uni.wroc.pl
blazej@smorfland.uni.wroc.pl
pamac@smorfland.uni.wroc.pl

Received on 20.12.2011;
revised version on 29.4.2013 (2116)