New Approach in Coding Signal Recognition Using Homogeneous Markov Chains Independently for Three Codon Positions

Paweł Błażej, Paweł Mackiewicz, and Stanisław Cebrat

***Abstract*—Many currently used algorithms for protein coding sequences require large learning sets consisting of real genes to estimate sensible values for used parameters and make the prediction reasonable. They also fail in recognition of short genes because their sequencescontain usually very weak coding signal. To overcome these problems, we worked out a new algorithm for finding protein coding potential in prokaryotic genomes. This algorithm uses homogeneous Markov chain for modeling nucleotide transition between fixed positions in codons thereby reduces the order of Markov chain retaining simultaneously information on dependence between nucleotides in sequence on relatively long distances. We tested performance of this algorithm in relationship to size of the learning set calculating true and false positive rates for different model orders. We also made some comparisons between our algorithm and commonly used GeneMark. The presented algorithm seems to work better than GeneMark especially for smaller learning sets.**

***Index Terms*—Gene finding, markovchains, ORF, protein coding sequence**

I. Introduction

Although many algorithms using different measures [1] for predicting protein coding sequences in prokaryotic genomes have been developed (see [2] and [3] for recent reviews), there is still an unsolved problem to distinguish true and false coding sequences among short open reading frames (ORFs) fewer than 300 bp. Though majority of these ORFs are spurious, some short genes are likely present in this set. They may encode peptides important for cell functioning, e.g. fulfilling regulatory or signalling functions.

The number of small ORFs (smORFs) increases exponentially with the decrease in their length [4], which hampers to recognize real genes among false frames. Recognition of these genes is also difficult because their coding signal is disturbed by statistical fluctuations coming out from their short sequences. As a result of this, gene predicting programs that achieve very high rates of gene detection, accept simultaneously quite a lot of false positives. Moreover, many of these algorithms rely only on large learning sets of true genes which are necessary to make reliable estimation of used parameters. Therefore, they are not optimal for small bacterial genomes that encode smaller sets of real genes. Then, to reach the proper size of learning sets, some non-coding ORFs are probably included in the training procedure. It may additionally increase the false positive rate in the stage of gene recognition. Furthermore, more general models which are assumed to be universal for a wide range of genomes are not appropriate for some, especially small genomes which are characterized by a specific nucleotide or codon bias.

To avoid these problems we developed a suitable statistical model which can be useful for detection a protein coding signal. This model utilizes specific properties of protein coding sequences related to correlations in nucleotide composition in particular codon positions, which was observed both in prokaryotic [5] and eukaryotic genomes [6]. Our algorithm uses homogeneous Markov chains to analyse this coding information on long distances in particular codon positions (separately for the first, the second and the third) and does not require high chain order to work properly. The new method was compared with commonly used GeneMark also based on Markov chains [7].

Manuscript received April 23, 2012; revised May 22, 2012.

The authors are with the Department of Genomics, Faculty of Biotechnology, University of Wrocław, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland (tel.: +48-071-3756-303; fax: +48-71-3756-234; e-mail: blazej@smorfland.uni.wroc.pl, pamac@smorfland.uni.wroc.pl, cebrat@ smorfland.uni.wroc.pl).

II. Algorithm For Finding A Coding Signal

The most common gene finders use Markov chain approach for modelling dependences between occurrence of nucleotides in protein coding sequences [7], [8]. Our method uses six homogeneous Markov chains forsuch sequences to determine the positional pattern frequencies which are next employedin detectionof coding signal in analysed sequences. This algorithm consists of two stages: the training step and the analysis step.

A. Training Step

The main task of this step is to compute model parameters which are calculated from a learning set of nucleotide sequences. For a given genome, such a set is a collection of annotated ORFs with ascribed function in GenBank database [9], excluding ORFs that were described as questionable or hypothetical.

B. Construction of Transition Matrices

Let us consider $S = \{S_{i1}, S_{i2}, \dots, S_{in}\}$ asequence of nucleotides extracted from fixed codon positions ($i=1,2,3$) in a protein coding sequence. We construct the initial probabilities $P(S_i^h)$ of$h$ nucleotides S_i situated in the same codon positions i (where h defines the model order) and also the probability transition matrices between nucleotides in the same codon position. Matrices M_1, M_2, M_3 concern to direct (i.e. sense) strands of training sequences whereas matrices M_4, M_5, M_6 are based on complementary strands of these

sequences (i.e. antisense strands). The matrices M_4, M_5, M_6 are useful for the model of "shadow" coding regions. Obviously,the matrices $M_1, \dots, M_6$ are transition matrices for homogeneous Markov chains.

C. Determination of Positional Pattern Frequencies

The obtained matrices are used to determine vectors of positional pattern frequencies in the learning set. The positional pattern is a vector of indices of matrices that give the highest value of total probability for a given codon position. In sum, there are 216 such potential patterns i.e. 111, 112, 113, etc. It is easy to notice that in this case we actually use a maximum likelihood approach. The frequencies of these vectors are obtained as follows:

1) Each sequence in every reading frame is analysed by moving windows with a fixed length (e.g. 96 nt) and a fixed window shift (e.g. 12 nt);
2) For each window, a vector of digits (d_1, d_2, d_3) called the positional pattern is determined in the following way:
   a) For each of three codon positions probabilities $P_{M_1}, P_{M_2}, P_{M_3}, P_{M_4}, P_{M_5}, P_{M_6}$ are calculated by using trained matrices $M_1, \dots, M_6$ respectively;
   b) if $P_{M_j} = max(P_{M_1}, P_{M_2}, P_{M_3}, P_{M_4}, P_{M_5}, P_{M_6})$, for fixed codon position i, then $d_i = j$ and finally a positional pattern (d_1, d_2, d_3) is obtained;
3) The frequency for each positional pattern is calculated from all analysed windows which aremade of reading framescoming from thelearning set (Fig. 1).

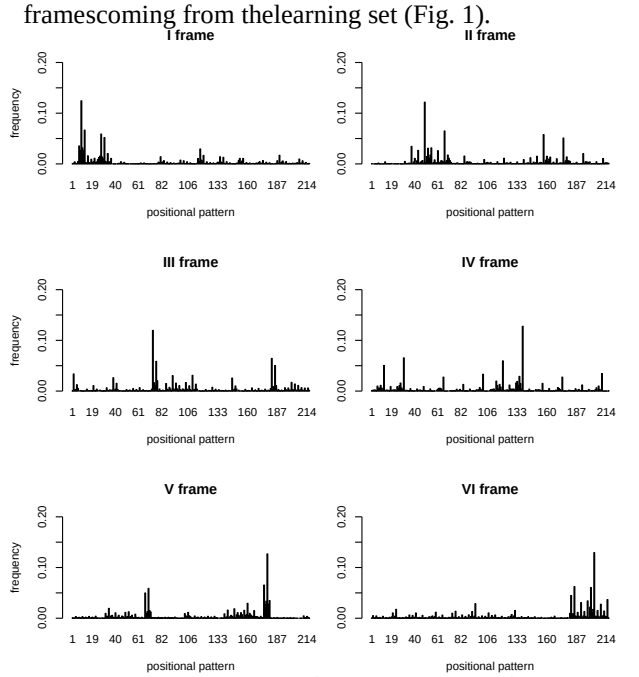

Fig. 1. Bar plots of positional pattern frequencies computed for the training set sequences from *Escherichia coli* genome for six reading frames.

D. Test or Analysis Step

The aim of this step is to detect the correct reading frame for an analysed DNA sequence.The first two steps are the same as in the determination of positional pattern frequencies (subsection II.*C.*):

1) Each sequence in every reading frame is analysed by moving windows with a fixed length (e.g. 96 nt) and a fixed window shift (e.g. 12 nt);
2) For each window, a vector of digits (d_1, d_2, d_3) called the positional pattern is determined in the following way:
   a. For each of three codon positions probabilities $P_{M_1}, P_{M_2}, P_{M_3}, P_{M_4}, P_{M_5}, P_{M_6}$ are calculated by using trained matrices $M_1, \dots, M_6$ respectively;
   b. if $P_{M_j} = max(P_{M_1}, P_{M_2}, P_{M_3}, P_{M_4}, P_{M_5}, P_{M_6})$, for fixed codon position i, then $d_i = j$ and finally a positional pattern (d_1, d_2, d_3) is obtained;
3) For a positional pattern $(d_1, d_2, d_3$)found for every window and every reading frame, we ascribe a respective frequency $P_1, P_2, P_3, P_4, P_5, P_6$ which were determined previously for the learning set;
4) As an additional non-coding reference, we assume the uniform distribution of positional pattern frequencies and introduce probability $P_7 = \frac{1}{216}$;
5) For every window we obtain a coding signal vector of frequencies for six reading frames plus the non-coding reference:
$$\left(\frac{P_1}{\sum_{i=1}^{7} P_i}, \frac{P_2}{\sum_{i=1}^{7} P_i}, \dots, \frac{P_7}{\sum_{i=1}^{7} P_i}\right)$$
6) Finally, the respective elements of the coding signal vector are averaged over all windows for a given sequence. The sequence is assumed to be coding in frame *i*, if the *i*position in the coding signal vectors has the highest value.

The idea of the presented algorithm is similar to the algorithm which was introduced in the paper [10]. The main difference is the extension of the set of possible positional pattern frequencies from 27 to 216. The new approach takes into account all possible frequencies obtained by using matrices $M_1, \dots, M_6$ at once. This approach gives better results especially in genomes with strong coding signal in the complementary (antisense) strand (e.g. in *Escherichia coli*genome).

III. Results

We have tested our algorithm on *Escherichia coli* 536 genome and have also analysed several small genomes of*Mycoplasma*. To evaluate efficiency of our algorithm we measured true positive rate (sensitivity) and false positive rate. For fixed model orders ($h = 2$ and $h = 4$) we also compared our results with scores obtained by GeneMark 2.5 for model orders $h = 2$ and $h = 5$. In this software,*Escherichia coli*was used as a reference genome.

A. Analysis of Escherichia Coli Genome

1) Estimation of true positive rate

The whole set of ORFs annotated as protein coding sequences including 2773 ORFs was divided into two parts:

1) training set (1000 ORFs);
2) test set (the rest 1773 ORFs).

Furthermore, from the training set, we chose randomly subsets containing increasing number of ORFs, i.e. 100, 200, ..., 1000 ORFs which were used as training sets. Our aim was to find dependences between true positive rate in the test set and the size of the learning set for the fixed model order ($h = 1, 2, 3, 4$). These results averaged on 20 simulations are presented in Fig. 2. The fraction of correctly recognized genes increases rapidly with the learning set size and stabilizes from the set of 300 or 400 ORFs. Interestingly, lower order models perform much better for smaller learning sets than the most complex one with the model order $h = 4$

which slightly surpasses the simpler models for larger learning sets.

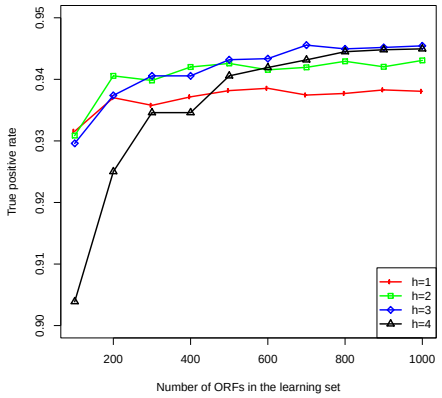

Fig. 2. Relationship between true positive rate and the size of training setcalculated for the new algorithm for different model orders ($h = 1, 2, 3, 4$).

2) Estimation of false positive rate

To estimate efficiency of the newalgorithm according to the false positive rate we used two test sets:

1) protein coding sequences that were read in incorrect (alternative) frame, i.e. 2, 3, 4, 5, and 6;
2) random sequences generated according to the genome nucleotide composition and the length distribution of real protein coding sequences.

The results averaged on 20 simulations are shown in Fig. 3. The relationships between false positive rate and the size of learning set are different for the two test sets. Generally, when ORFs in the incorrect reading frame are used as a test set, the false positive rate decreases with the size of the learning set. The rate is higher for the $h = 4$ model than for the simpler models when the smaller learning sets are considered.The opposite situation is for the larger learning sets. However, for random generated sequences,the rate increases with the learning set size. Moreover, the high order model $h = 4$in comparison to the simpler models receives the lowest false positive rate for all learning sets of random generated sequences.

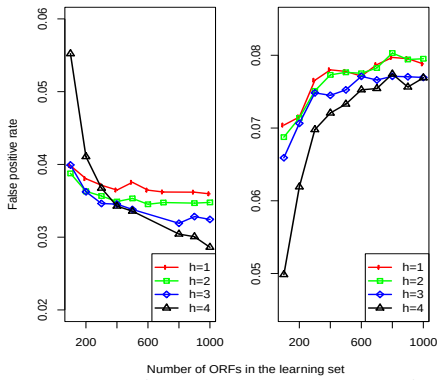

Fig. 3. Relationship between false positive rate and the size of the training set for: sequences in incorrect reading frame (in the left) and randomly generated sequences (in the right). Four model orders ($h = 1, 2, 3, 4$) were considered.

B. Comparison of the New Algorithm with Gene Mark

We used the same learning and test sets to make comparisonbetween the new algorithmand the popular softwareGeneMark.The new algorithm was applied with the orders$h = 2$and $h = 4$whereas GeneMarkwith the order$h = 2$ and $h = 5$. We chose the order of $h = 5$ for GeneMarkbecause it is the most common used order in the current GeneMark version 2.5. These two algorithms were compared according to true (Fig. 4 and 5) and false positive rates (Fig. 6 and 7) in the relationship to the size of learning set.

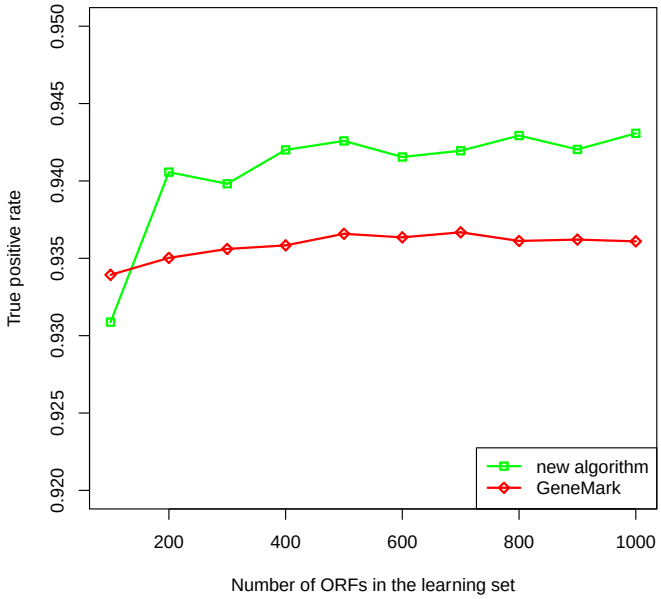

Fig. 4. Comparison of true positive rates between our algorithm (green) and GeneMark 2.5 (red) for $h = 2$model order in the relationship with the size of the learning set.

All algorithms achieve true positive rate higher than 0.93. For low order models (Fig. 4), the new algorithm receives the true positive rate higher than GeneMark, i.e. more than 0.94, for learning sets including more than 100 ORFs. Only for the smallest set consisting of 100 sequences, GeneMark performs slightly better. The increase in the true positive rate with the number of ORFs in the learning set is observed to about 500 sequences for the both algorithms whereas for more numerous sets, the rate stabilizes.

When more complex model orders are used (Fig. 5),the increase in the rate with the learning set size is more pronounced, especially for GeneMark. For the new algorithm, there is no change in the rate for learning sets including 800 and more sequences. The new algorithm still works better than GeneMark and obtains the true positive rate higher than 0.94for all learning sets. However, the difference between the two algorithms diminishes with the increase of the learning set size.The two algorithms converge for the set consisting of 1000 ORFs achieving true positive rate about 0.945.

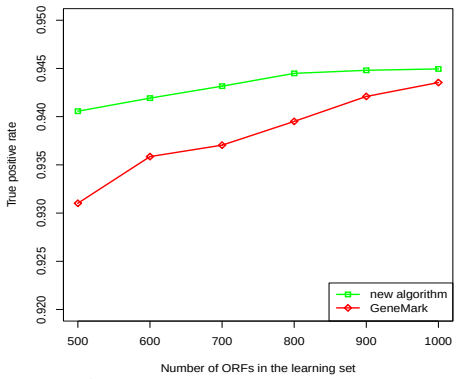

Fig. 5. Comparison of true positive rates between our algorithm (green) and GeneMark 2.5 (red) for model order $h = 4$ and $h = 5$, respectively, in the relationship with the size of the learning set.

Comparison of two methods regarding relationship between the false positive rate and the size of learning set is presentedin Fig. 6 and Fig. 7 for different model orders. The relationship is weaker than for true positive ratealthough in most cases the false positive rate decreases with the learning set size. Interestingly, performance of two algorithms depends on the tested set. The new algorithm has lower false positive rate for sequences read in incorrect frames with the $h = 2$model order and for random sequences with the order of $h = 4$ while GeneMark performs better in the case of incorrect reading frames with the order of $h = 5$ and for random sequences with the order of $h = 2$. By average, the two algorithms show similar 0.055 false positive rate. GeneMark achieves both the lowest and the highest false positive rate values.

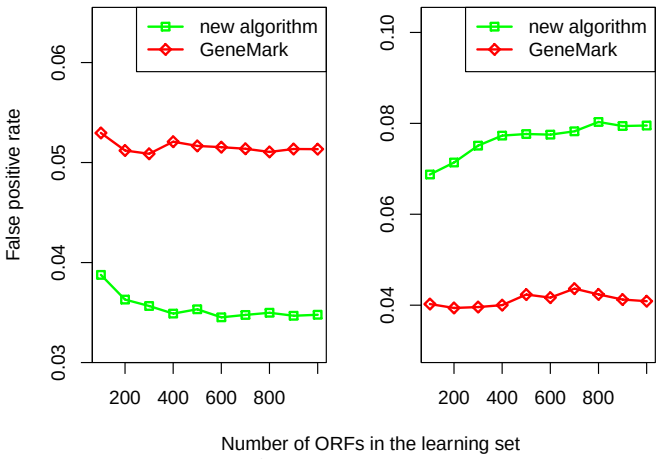

Fig. 6. Relationship between false positive rates calculated in our algorithm (green) and GeneMark (red) with the size of the learning set for: sequences read in incorrect frame (in the left) and random sequences (in the right). The model order of$h = 2$ was considered.

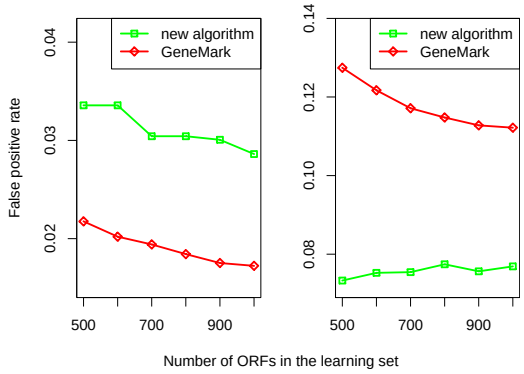

Fig. 7. Relationship between false positive rates calculated in our algorithm (green) for $h = 4$ and GeneMark (red) for$h = 5$with the size of the learning set for: sequences in incorrect reading frame (in the left) and random sequences (in the right).

C. Coding Signal Strength in Different Group of Sequences

The main task of our algorithm is to find a sequence with coding signal in a proper reading frame. In Fig. 8 we compared the strength of the coding signal for model order of $h = 2$ in different group of sequences: protein coding sequences, sequences in incorrect reading frame, and randomly generated sequences. The strength was described by empirical tail distribution functions (i.e. $1 - F(x) = P(X > x)$), where X is a random variable of the value with the strongest coding signal. The distribution for protein coding sequences is clearly shifted towards higher values of coding signal.Protein coding sequences with coding signal higher than 0.3 are over 91% while there are only 13% ofincorrect ORFs and almost no random sequences (0.9%) above this value.

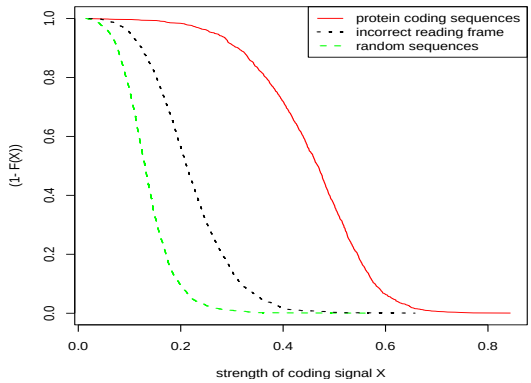

Fig. 8. Comparison empirical tail distribution functions $(1 - F(X))$ for: protein coding sequences (red solid line), sequences in incorrect reading frame (black dotted line), random sequences (green dashed line).

D. Small Genomes

As was mentioned in Introduction, one of the most important problems in the recognition of protein coding sequences is difficulty in obtaining a large enough training set when small genomes are analysed. Here, we tested the new algorithm in the case of genomes smaller than 1 Mbpassuming tiny learning sets (Table I). For every genome we chose randomly 200 annotated ORFs in the training set whereas the rest of ORFs was used to build the test set. Sets for calculation of the false positive rate were prepared similarly but were based on ORFs read in incorrect frames. The algorithm achieved true positive rate higher than 0.90 and false positive rate below 0.1.

TABLE I: TRUE POSITIVE RATE (TPR) FOR SMALL *MYCOPLASMA* GENOMES.

Mycoplasma strain (genome size)	TPR
M. agalatiae (0.88 Mbp)	0.97
M. arthihritidis 158L3 1 (0.82 Mbp)	0.96
M. mobile 163K (0.78 Mbp)	0.97
M. pulmonis UAB CTIP (0.96 Mbp)	0.94
M. synoviae 53 (0.8 Mbp)	0.97

IV. CONCLUSION

The presented algorithm describes nucleotide transition in three codon positions independently. Therefore, it reduces order of Markov chain retaining simultaneously the same coding information that is contained in higher order chains thatanalysethe dependence between nucleotides in subsequent codon positions of a studied sequence. The new algorithm achievedgood performance both for both small and large learning sets. In our test we obtained average the true positive rate over 0.90 and the false positive ratebelow 0.1. Models with lower order worked usually better for smaller learning sets but the most complex models were better for larger learning sets. However, the difference both in the true positive rate and the false positive rate between models of

different order was bigger for the small learning sets than for larger ones. Models with higher order showed stronger relationship with the size of learning set than simpler ones. The performed comparisons indicate that our algorithm is comparable with GeneMark algorithm according to the false positive rate but achieves the higher true positive rate. In addition to that, the new algorithm worksespecially well under low order models and seems useful in the recognition of protein coding sequences in tiny genomes with small coding capacity.

REFERENCES

[1] J. Fickett and C. Tung, "Assessment of protein coding measures," *Nucleic Acid Research*, vol. 20, pp. 6441-6450, 1992.

[2] W. H. Majoros, *Methods for computational gene prediction*, Cambridge University Press, 2007.

[3] R. K. Azad, "Genes in prokaryotic genomes and their computational prediction," in *Computational methods for understanding bacterial and* archeal *genomes, Series of advances in Bioinformatics and Computational Biology*, vol. VII, Y. Xu and J. P. Gogarten, Eds, College Press, pp. 39-74, 2008.

[4] A. Gierlik, P. Mackiewicz, M. Kowalczuk, M. R. Dudek, and S. Cebrat, "Some hints on Open Reading frame statistics – how ORF length depends on selection," *International Journal of Modern Physics C*, vol. 10, pp. 645-643, 1999.

[5] 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*, vol. 2, pp. 259-268, 1997.

[6] S. Cebrat, M. R. Dudek, and P. Mackiewicz, "Sequence asymmetry as a parameter indicating coding sequence in *Saccharomyces cerevisiae* genome," *Theory in Biosciences*, vol. 117, pp. 78-89, 1998.

[7] M. Borodovsky and J. McIninch, "Genemark: pararell gene recognition for both DNA strands," *Computers & Chemistry*, vol. 17, pp. 123-133, 1993.

[8] M. Y. Borodovsky, Y. A. Sprizhitskii, E. I. Golovanov, and A. A. Aleksandrow, "Statistical patterns in primary structures of the functional regions of the genome in *Escherichia coli*," *Molecular Biology*, vol. 20, pp. 826-840, 1144-1150, 1986.

[9] D. A. Benson, I. K. Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E W. Sayers, "GenBank," *Nucleic Acids Research*, 40(Database issue), pp. D48-53, 2012.

[10] P. Błażej, P. Mackiewicz, and S. Cebrat, "Using genetic coding wisdom for recognizing protein coding sequences," in *Proceedings of the 2010 International Conference on Bioinformatics &* Computational *Biology, BIOCOMP 2010*, Las Vegas Nevada, USA, vol. 1, pp. 302-305, 2010.