Using the genetic code wisdom for recognizing protein coding sequences

Paweł Błażej, Paweł Mackiewicz, Stanisław Cebrat
Department of Genomics, Faculty of Biotechnology, Wrocław University,
ul. Przybyszewskiego 63/77 51-148 Wrocław, Poland,
e-mail: blazej@smorfland.uni.wroc.pl

Abstract— *We have elaborated a new method of recognizing protein coding sequences in genomic sequences. The method is exploiting a specific way of genetic code degeneration and relations between mutational pressure and selection pressure shaping the amino acid usage in the proteomes. It is based on analyses of correlations in nucleotide occurrence separately in the first, the second and the third putative codon positions using only six matrices 4x4. Small sizes of matrices enable using only a few coding sequences for training the algorithm. The results of the new method were compared with Markov chain methods used in GeneMark for different genomes including DNA strand (leading/lagging) discrimination. There are no arbitrary "cut off" discriminating between coding and noncoding sequences, on the other hand there is a possibility to rank putative coding sequences according to their coding probability what is especially important in looking for small coding ORFs.*

Keywords: gene prediction, ORF, short genes, coding potential

1. Introduction

There are two main and different approaches for looking for protein coding sequences; the first one based on correlations in nucleotide sequences [1] and the second one based directly on correlations in potentially coded amino acid sequences [2], [3]. See [4] and [5] for recent reviews of different gene prediction methods. Unfortunately, both methods need many learning sequences, it is the best if the learning set is from the studied genome or even located at the defined DNA strand leading or lagging. A big number of learning sequences are necessary to fill up the very large matrices used in comparisons between the studied sequences with sequences considered as standard. On the other hand, it is known phenomenon that the mutational pressure responsible for nucleotide composition of DNA fits to selection pressure responsible mainly for amino acid composition of proteome while the criterion of fitness is the minimization of harmful effects of mutations into the protein coding sequences [6]. Thus, the genetic code plays the central role in the relations between the selection and the mutational pressure and particularly it is the way the code is degenerated. Degeneration of the genetic code should be understood not only as using the different codons for coding the same amino acid (mainly the third codon positions accepting so-called synonymous mutations) but also as some kind of degeneration of the first and second positions i.e. thymine in the second position coding for hydrophobic amino acid independently of the first and third positions or adenine in the second position coding for polar amino acids. It is also known that even in the same genome the amino acid composition of proteins depends on the mutational pressure operating on the leading and lagging strands which means that it is not only DNA which is asymmetric in its nucleotide composition but also fractions of proteomes coded by the two DNA strands are asymmetric in their amino acid compositions [7]. Correlations in the occurrence of amino acids of specific characters have been already considered in recognizing the coding sequences but the values describing the physical and chemical properties of amino acids were accepted disregarding the coding property of the genetic code. We have used directly this property of genetic code for recognizing the protein coding sequences looking for correlations in nucleotide composition separately for the first, the second and the third codon positions [8]. This method allows using very small matrices (six matrices 4x4) and, as a result, small number of learning sequences. The method discriminates very sharply between coding and noncoding ORFs (Open Reading Frames), shows the phase in which a given sequence could code or even indicate the DNA strand (leading/lagging) where the coding sequence is located.

2. Algorithm

The presented algorithm for finding protein coding sequences consists of two steps: the training step and the test (or analysis) step.

2.1 The training step

In this step, model parameters are computed based on nucleotide sequences of a learning set. For such set, ORFs annotated in GenBank with ascribed function are used for a given species, excluding ORFs that were described as hypothetical.

2.1.1 Construction of matrices

Let us consider $S = \{s_{i1}, s_{i2}, \ldots, s_{in}\}$ (where $i = 1, 2, 3$) which is a sequence of nucleotides extracted from fixed codon positions 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 (i.e. between nucleotides in the same codon position). Matrices M_1, M_2, M_3 concern to direct (sense) strands of training sequences whereas matrices M_4, M_5, M_6 are based on complementary strands of these sequences (antisense). Matrices M_4, M_5, M_6 are useful for a model of "shadow" coding regions. The idea of incorporation the "shadow" model was introduced by Borodovski and McIninch [9] who used it to avoid too many false positive predictions on the complementary strand.

2.1.2 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 probability for a given codon position. In sum, there are 27 such potential patterns e.g. 111, 112, 123, 121, 122, ..., etc. The frequencies of these vectors are obtained as follows:

1. Each sequence in every reading frame is analyzed by moving windows with a fixed length (e.g. 96 nt) and a fixed shift (e.g. 12 nt);
2. For each window a vector of digits (d_1, d_2, d_3) and (c_1, c_2, c_3) (called the positional pattern) is determined in the following way:
   - for each of three codon positions probabiblities P_{M_1},P_{M_2},P_{M_3} are calculated by using trained matrices M_1, M_2, M_3, respectively;
   - if $P_{M_j} = \max(P_{M_1}, P_{M_2}, P_{M_3})$ (for fixed codon position i), then $d_i = j$ and for each window a positional pattern (d_1, d_2, d_3) is obtained from such an analysis of all three codon positions;
   - pattern (c_1, c_2, c_3) for the complementary strand is obtained in the same way by using matrices M_4, M_5, M_6.
3. The frequency for each positional pattern are calculated from all analyzed windows and all sequences from the learning set for each reading frame (Fig. 1).

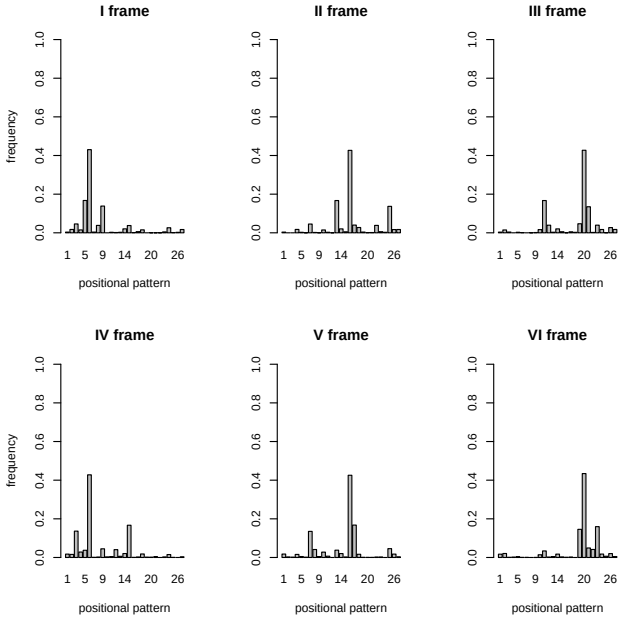

Fig. 1. Barplots of positional pattern frequencies computed for the training set from *Borrelia burgdorferi* genome for six reading frames.

2.2 The test or analysis step

The aim of this step is to find the correct reading frame for an analyzed DNA sequence. The first two steps are the same as in determination of positional pattern frequencies (subsection 1.1.2)

1. As 1 in 2.1.2
2. As 2 in 2.1.2
3. For a positional pattern (d_1, d_2, d_3) or (c_1, c_2, c_3) found for every window and every reading frame, we ascribe a respective frequency P_1, P_2, P_3 or P_4, P_5, P_6 which were determined previously for the learning set;
4. As an additional non-coding reference we assume uniform distribution of positional pattern frequencies and introduce $P_7 = 1/27$;
5. For every window we obtain a coding signal vector of probabilities for six reading frames plus the non-coding reference:
$$\left(\frac{P_1}{\sum_{i=1}^{7} P_i}, \frac{P_2}{\sum_{i=1}^{7} P_i}, \ldots, \frac{P_7}{\sum_{i=1}^{7} P_i}\right).$$
6. Finally, the respective elements of the coding signal vector are averaged over all windows for a given sequence. The sequence is coding in frame i if the i position in coding signal vectors has the highest value.

3. Results

3.1 Detection of the coding signal

To check how the presented algorithm works on a real DNA-sequence, two different genomes were chosen: *Borre-*

lia burgdorferi and *Escherichia coli*.

1) For a given organism the set of annotated genes was divided randomly for two subsets (the learning set and the test set);
2) After the training step algorithm (for model order $h = 1$) the tested sequences were analysed;
3) The results were compared with results obtained by GeneMark version 2.5f (for model order $h = 1$ and $h = 2$) which was also trained on the same learning set.

3.1.1 *B. burgdorferi* genes

The set of annotated 474 sequences was divided randomly for a learning set (200 sequences) and a test set (274 sequences). The results averaged over ten repetitions of learning and test steps are presented in Table 1. The percent of genes recognized as coding for the new method is equal to 93% which is higher than for GeneMark of the same order. However, the percentage increased significantly to 97% when we considered genes on the leading and lagging strands separately, i.e. both the learning set and the test set consisted of only genes from the same DNA strand. Interestingly, the learning set contained only 10 genes in this case.

Table 1: Averaged percent of genes recognized as coding in *B. burgdorferi*

New method ($h = 1$)	
- the whole set	93.4
- only lagging strand genes	97.0
- only leading strand genes	97.3
GenMark 2.5f	
- the whole set ($h = 1$)	91.0
- the whole set ($h = 2$)	97.6

3.1.2 *E. coli* genes

The set of annotated sequence (2774-sequences) are divided randomly for a learning set (1000-sequences) and a test set (1773-sequences). Table 2 shows the results averaged over ten repetitions of the whole procedure. The percent of genes recognized as coding for the new method is higher than for GeneMark of the same order and nearly equal to the GeneMark result of the second order.

Table 2: Averaged percent of genes recognized as coding in *E. coli*

New method ($h = 1$)	91.3
GenMark 2.5f ($h = 1$)	79.4
GenMark 2.5f ($h = 2$)	93.7

3.2 Small ORFs ranking

Recognition of the coding rhythm is especially difficult and important when looking for protein coding sequences among short ORFs. The number of such ORFs generated inside other longer ORFs or in the intergenic sequences is usually very high while the fraction of coding small ORFs is relatively low [10]. In the genome of *Streptomyces coelicolor*, there are 28917 ORFs of length 31 to 99 codons and only 555 ($\sim 2\%$) of them are annotated in the data base as coding. In the left plot of Fig 2 we have shown the distribution of all small ORFs from *S. coelicolor* genome according to their positions in the ranking of coding probability where the positions of ORFs with annotated functions are marked. In the right diagram of Fig 2 the numbers of annotated ORFs in the consecutive classes of ranked small ORFs are shown. The first 10% of the highest ranked small ORFs contains 93% of all annotated ORFs.

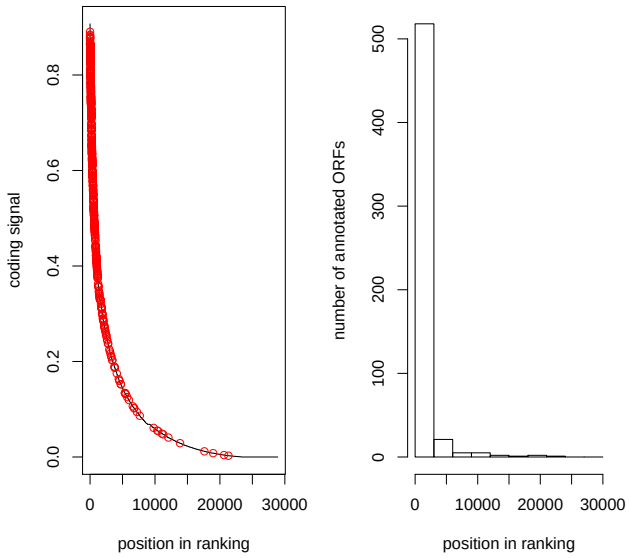

Fig. 2. Ranking of small ORFs. Left panel: distribution of all small ORFs from *S. coelicolor* genome according to their positions in the ranking of coding probability (line) and positions of ORFs with annotated functions (marked with open circles). Right panel: histogram of the numbers of annotated ORFs in the consecutive classes of ranked small ORFs.

4. Summary

The new method of protein coding sequences recognition works very efficiently with small number of learning sequences and could be preferentialy used in analysis of coding capacity in small genomes and short protein coding sequences.

References

[1] M.Yu. Borodovsky, Y.A. Sprizhitskii, E.I. Golovanov, A.A. Aleksandrov, "Statistical Patterns in Primary Structures of the Functional

Regions of the Genome in *Escherichia coli*", *Molecular Biology*, 20, 826-833, 833-840, 1144-1150, 1986.

[2] P. McCaldon, P. Argos, "Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences", *PROTEINS: Structure, Function and Genetics*, 4, 99-122, 1988.

[3] J.W. Fickett, C.S. Tung, "Assessment of protein coding measures", *Nucleic Acids Res.* 20(24):6441-6450, 1992.

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

[5] R.K. Azad, "'Genes in prokaryotic genomes and their computational prediction", in: Y. Xu, J.P. Gogarten (eds), *Computational methods for understanding bacterial and archaeal genomes*, Series on Advances in Bioinformatics and Computational Biology - Vol. 7:39-74, Imperial College Press, 2008.

[6] P. Mackiewicz, P. Biecek, D. Mackiewicz, J. Kiraga, K. Bączkowski, M. Sobczyński, S. Cebrat, "Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code", *Lecture Notes in Computer Science* 5103, 100-109, 2008.

[7] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, "DNA Asymmetry and the Replicational Mutational Pressure", *Journal of Applied Genetics* 42 (4), 553-577, 2001.

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

[9] M. Borodovsky, J. Mcininch, "GenMark: pararell gene recognition for both DNA strands", *Comput. Chem.* 17 123-133, 1993.

[10] A. Gierlik, P. Mackiewicz, M. Kowalczuk, M.R. Dudek, S. Cebrat, "Some hints on Open Reading Frame statistics - how ORF length depends on selection", *Int. J. Modern Phys. C* 10(4), 635-643, 1999.