Evaluation of first and second Markov chains sensitivity and specificity as statistical approach for prediction of sequences of genes in virus double strand DNA genomes

Jalal Farzami, Ebrahim Hajizadeh*, Gholamreza Babaei-Rochee, Anoshirvan Kazemnejad

Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, P.O.Box: 14115-331, Tehran, I.R. Iran

Abstract
Growing amount of information on biological sequences has made application of statistical approaches necessary for modeling and estimation of their functions. In this paper, sensitivity and specificity of the first and second Markov chains for prediction of genes was evaluated using the complete double stranded DNA virus. There were two approaches for prediction of each Markov Model parameter, initial probability and transition matrix, which together with the first and second Markov chains resulted in development of eight algorithms for gene prediction. In order to compare the algorithms, a sensitivity and specificity repeated measure with 3 factors (Markov model, type of selection and estimation of transition probabilities) were utilized. Results significantly revealed that the second order Markov chain had more sensitivity and specificity than the first order Markov chain, with "p-Value" < 0.001. By adding the covariates, the number of annotated genes per length of genome as well as the A & T and C & G contents of genomes in the repeated measure showed an insignificant difference between the sensitivities of the two Markov models (0.407, 0.071 and 0.120, respectively). It was also proved that gene base-pairs per genome length and A & T contents of the genome, as model covariates, resulted in significant differences between the specificities of the Markov models.
Keywords: Gene prediction; Markov chain; Virus genome.

*Correspondence to: **Ebrahim Hajizadeh,** Ph.D.
Tel: +98 21 88011001; Fax: +98 21 88013030
E-mail: hajitm@yahoo.com

INTRODUCTION

Currently, there are large number of disciplines in the field of biological sciences, which are growing exponentially due to rapid advances in biotechnology, requiring and involving an extensive amount of quantitative data and information. The large quantities of data, which are conventionally archived in biological data-bases, make statistical analyses, data-mining and information retrieval a requisite. Information on the primary structure of biological macromolecules such as DNA sequences is one type of the so called datasets. One common problem in this area is the determination of relationships between the physical structure of these molecules and their biological functions (Borodovsky *et al*., 1986), which resembles finding a coding region.

There are currently two approaches developed for finding genes. One is based on similarity search, known as the extrinsic approach, BLAST (Altschul *et al*., 1997), the other one is related to statistical analyses of the sequence under consideration and is based on an intrinsic approach which include GLIMMER (Steven *et al*., 1997), GeneMark web software (Besemer and Borodovsky, 2005) and ZCURVE_V (Guo and Zhang, 2006). Gene annotation in viruses often relies upon similarity search methods. The specificity of these methods is high but the sensitivity is relatively low since they may miss either genes that are unique to a particular genome or those highly divergent from known homologs (Mils *et al*., 2003). Unfortunately, viral gene-finding tools based on statistics, currently are very few, except GeneMark gene-

finding family (Guo and Zhang, 2006).

One statistical model for analyzing the genome is the Markov Chain (Borodovsky and McIninch, 1993) which is used as a basis to construct many other algorithms for prediction of genes in genome sequences. This method is based on probability prediction of a sequence under a number of assumptions, by using calculations derived from the structure of nucleotides of the genome sequence. In this paper we have evaluated the sensitivity and specificity of Markov Chain Modeling to find genes in viral double stranded DNA genomes.

MATERIALS AND METHODS

In this study, 63 complete double stranded DNA samples belonging to viral genomes were randomly selected from the list of existing viral ds DNA complete genomes with no RNA stage, available at the national center for biotechnology information (NCBI) database (Pruitt *et al*., 2003), and downloaded from GenBANK (Benson *et al*., 2004) for statistical analysis. The sample size was determined based on operation characteristic (OC) curves for the fixed effects model, provided that the power of the tests be greater than 0.95 (Montgomery, 2001), DNA sequence consists of a string of nucleotides {A, T, C, and G}. Each base of the sequence strand was assumed as a random variable and defined as X = 1, 2, 3, 4, respectively.

Therefore each sequence with n+ 1 base was defined as a chain consisting of n+1 dependent random variables and the probability for the sequence was defined as below:

$$P(X_n, X_{n-1}, \ldots, X_0) = P(X_n|X_{n-1}, \ldots, X_0)\, P(X_{n-1}, X_{n-2}, \ldots, X_0)$$

So, with the first order Markov assumption we have:

$$P(X_n|X_{n-1}, \ldots, X_0) = P(X_n|X_{n-1})$$

Then:

$$P(X_n, X_{n-1}, X_{n-2}, \ldots, X_0) =$$

$$P(X_n \mid X_{n-1})\, P(X_{n-1} \mid X_{n-2}) \ldots P(X_1 \mid X_0)\, P(X_0) =$$

$$P(X_0) \prod_{i=1}^{n} P(X_i \mid X_{i-1})$$

To estimate the probability of a given DNA sequence we must derive the following probabilities:

$P(X_i|X_{i-1})$ & $P(X_0)$,

where $P(X_0)$ is the probability of observation at the first base of a sequence (initial probability) and $P(X_i|X_{i-1})$ is the transition probability which can be presented in a 4×4 matrix called a transition matrix.

These probabilities are estimated by the maximum likelihood approach which is based on observations.

In order to estimate the initial probability, we applied two approaches, based on the selected regions of the genome sequences. The first approach involved finding the ORFs of each reading frame on both strands and selecting those ORFs which were longer than the smaller known gene of a given genome. Therefore, the nucleotide adenine could be found as the first base of all sequences so that:

$$P(X_0 = 1) = 1 \qquad \& \qquad P(X_0 \neq 1) = 0,$$

where, X_0= 1 indicates presence of adenine as the first base.

The second approach was to find all stop codons of the reading frames on both strands and to select regions between two successive stop codons, which were longer than the smaller known gene of a given genome. Since there is a stop codon at the first base of each region, therefore:

$$P(X_0 = 2) = 1 \qquad \& \qquad P(X_0 \neq 2) = 0,$$

where, X_0= 2 indicates presence of thymine as the first base.

Elements of the transition matrix were derived by calculating frequencies of dinucleotides in a set of sequences from coding regions in order to estimate the transition matrix for these regions and the whole genome sequence, as well as having a comparable transition matrix.

In this research two types of estimations were utilized for the transition matrix. The first one was to take all gene sequences in both strands to estimate the transition matrix for the coding regions and the second one was to estimate two different matrices for each strand of the DNA sequence using their gene sequences.

The second order Markov assumption produces the following probabilities:

$$P(X_n|X_{n-1}, \ldots, X_0) = P(X_n|X_{n-1}, X_{n-2}),$$

then

$$P(X_0)\, P(X_1 \mid X_0) \prod_{i=2}^{n} P(X_i \mid X_{i-1}, X_{i-2})$$

$$P(X_{n-1} \mid X_{n-2}, X_{n-3}) \ldots P(X_1 \mid X_0)\, P(X_0) =$$

$$P(X_0) P(X_1 \mid X_0) \prod_{i=2}^{n} P(X_i \mid X_{i-1}, X_{i-2})$$

This assumption results in similar estimations as explained above. However there are some differences

in the results such as the conditional probability P $(X_1|X_0)$ that can be easily estimated using our proposed methods of selection.

After estimation of the transition matrices for the whole genome and the gene sequences for a given region of the sequence, it is possible to calculate probabilities under the two assumptions using each transition matrix and hence derive the following equations (Durbin *et al*., 1998).

For the first order Markov chain:

$$\frac{P_G(X_n, X_{n-1}, X_{n-2}, ..., X_0)}{P_W(X_n, X_{n-1}, X_{n-2}, ..., X_0)} = \frac{P_G(X_0)\prod_{i=1}^{n} P_G(X_i \mid X_{i-1})}{P_W(X_0)\prod_{i=1}^{n} P_W(X_i \mid X_{i-1})}$$

For the second order Markov chain:

$$\frac{P_G(X_n, X_{n-1}, X_{n-2}, ..., X_0)}{P_W(X_n, X_{n-1}, X_{n-2}, ..., X_0)} = \frac{P_G(X_0)P_G(X_1 \mid X_0)\prod_{i=2}^{n} P_G(X_i \mid X_{i-1}, X_{i-2})}{P_W(X_0)P_W(X_1 \mid X_0)\prod_{i=2}^{n} P_W(X_i \mid X_{i-1}, X_{i-2})}$$

where P_G is the probability under the assumed gene region and P_W is the probability for the whole genome. If the above proportion results in a value grater than 1, it can be concluded that the sequence is a gene.

The two Markov models, two selections of sequences and two estimations of transition matrices lead us to have eight types of algorithms:

- FO-ORF-1TM: this algorithm is based on the first order Markov chain, selection of ORFs and estimation of one transition matrix for both strands.
- FO-ORF-2TM: this algorithm is based on the first order Markov chain, selection of ORFs and estimation of two separate transition matrices for each strand.
- FO-SC-1TM: this algorithm is based on the first order Markov chain, selection of regions between stop codons and estimation of one transition matrix for both strands.
- FO-SC-2TM: this algorithm is based on the first order Markov chain, selection of regions between stop codons and estimation of the two separate transition matrices for each strand.
- SO-ORF-1TM: this algorithm is based on the second order Markov chain, selection of ORFs and estimation of one transition matrix for both strands.
- SO-ORF-2TM: this algorithm is based on the second order Markov chain, selection of ORFs and estimation of two separate transition matrices for each strand.
- SO-SC-1TM: this algorithm is based on the second order Markov chain, selection of regions between stop codons and estimation of one transition matrix for both strands.
- SO-SC-2TM: this algorithm is based on the second order Markov chain, selection of regions between stop codons and estimation of two separate transition matrices for each strand.

By selection of the open reading frames, it can be concluded that a gene is predicted correctly if one ORF is exactly representing one gene, but for the selection of regions between stop codons the prediction is correct only if one gene exists in the region. So the algorithms based on ORF are precise but the others are not.

In order to compare the algorithms, we calculated their sensitivities and specificities for all analyzed genomes. Sensitivities and specificities of these eight algorithms were defined as below (Yin and Yao, 2007):

Sensitivity = $TP/(TP+FN)$

Specificity = $TN/(TN+FP)$,

where *TP* is the number of regions which have been annotated in the NCBI and detected with algorithm both as genes, *FN* is the number of annotated regions in NCBI as gene, which were not detected by the algorithm. *TN* is the number of analyzed regions that have not been annotated in the NCBI and not detected by the algorithm as genes and *FP* is the number of analyzed regions that have not been annotated in the NCBI as genes but have been detected with the algorithm.

Since there are the same groups of cases (virus genomes) and that these algorithms are used on the groups which are not independent, a repeated measure method was utilized with three factors, i.e. “type of Markov model”, “type of region selection”, and “type of estimation of transition matrix” each with two levels as follows:

Type of Markov model: first order or second order;
Type of region selection: ORFs or between stop codons;

Type of estimation of transition matrix: one transition matrix or two transition matrices.

For this purpose, normal groups were applied to test the normality of sensitivities and specificities resulting from each algorithm, by using the Kolmogorov-

Smirnov test, hence the normalities were accepted ($p > 0.05$). On the other hand, the differences among sensitivity and specificity averages of groups which are related to genome characteristics must be considered. For this purpose we defined certain known genome characteristics as random variables and entered them one by one into a repeated measure model, as covariates. These variables were as follows:

1. Number of annotated genes
2. Number of base-pairs which are in genes per genome length
3. Average of gene lengths
4. Number of annotated genes per genome length
5. Frequency of adenine or thymine on both strands of the DNA sequence that is equal.
6. Frequency of adenine or thymine on both strands of DNA sequence that is equal.

For each genome, information was gathered as measurements regarding each characteristic.

RESULTS

Sensitivity: As shown in Table 1, derived p-values from the Kolmogorov-Smirnov test regarding the normality of the sensitivities of the eight algorithms for finding genes are more than (0.05). Therefore, it can be assumed that they are normally distributed with the mean and standard deviation noted in Table1.

Results of the repeated measurements analysis in the first column (non-covariate) of Table 2 show that the effects of all three factors; the order of Markov chain

Table 1. Kolmogorov-Smirnov test results for normality of the eight algorithms' sensitivities.

	K-S test p-values	Mean	Std. Deviation
FO-ORF-1TM	0.268	0.651750	0.1891459
FO-ORF-2TM	0.596	0.616273	0.1842035
FO-SC-1TM	0.383	0.624352	0.1104067
FO-SC-2TM	0.258	0.613716	0.1075789
SO-ORF-1TM	0.614	0.685098	0.2041213
SO-ORF-2TM	0.904	0.610121	0.1706034
SO-SC-1TM	0.484	0.820204	0.1131639
SO-SC-2TM	0.912	0.773628	0.1198575

"FO" is first order, "SO" is second order, "ORF" is open reading frame, "SC" on stop codons, "1TM" on one transition matrix for whole genome and "2TM" is one transition matrix for each strand of DNA genome.

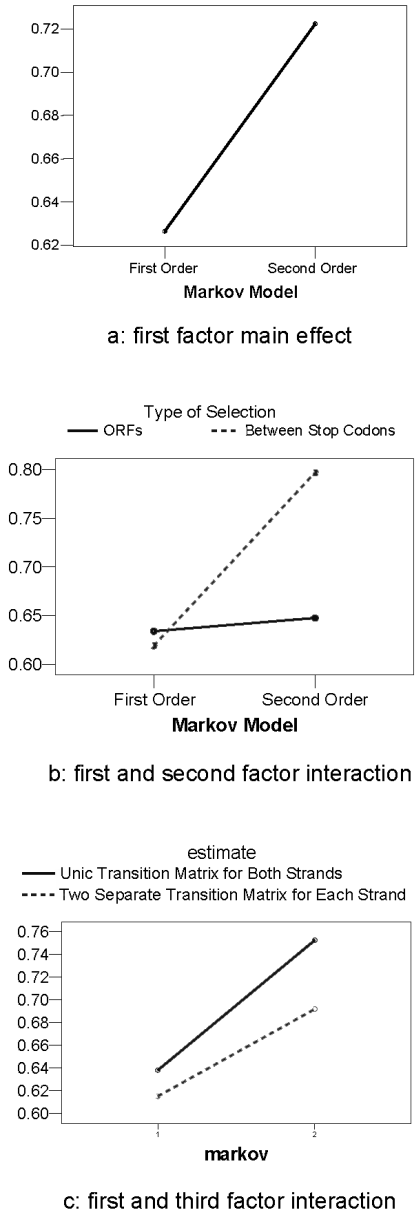


Figure 1. Estimated sensitivity means derived from repeated measurement analysis.

($p < 0.001$), region selection ($p=0.002$) and type of estimation ($p < 0.001$), are statistically significant. This means that there was significant difference of sensitivity means between levels of these factors. This also suggests that the mutual interactions between the Markov model and the other two factors have a significant effect as well ($p < 0.001$), and that at each level of these two factors the difference between Markov chains is significant.

In Figure 1, means of sensitivities at each level of repeated measurement model factors are illustrated. As it shows, The algorithms based on second order

Table 2. Repeated measure analysis results for sensitivities with three factors, Markov model, region selection and type of estimate transition matrix and results of this model with mentioned covariates.

	Covariates						
Factors	None	Number of annotated genes	Gene BPs / genome length	Gene length average	Number of annotated genes/ length of genome	A & T content	C & G content
Markov Model	0.000	0.000	0.010	0.000	0.407	0.071	0.120
Type of Selection	0.002	0.133	0.804	0.011	0..925	0.452	0.523
Type of Estimation	0.000	0.000	0.935	0.000	0.635	0.368	0.292
Markov Model* Type of Selection	0.000	0.000	0.119	0.005	0.000	0.382	0.012
Markov Model* Type of Estimation	0.000	0.002	0.862	0.000	0.486	0.844	0.095
Type of Selection* Type of Estimation	0.003	0.000	0.170	0.890	0.073	0.096	0.746
Markov Model* Type of Selection* Type of Estimation	0.819	0.100	0.556	0.219	0.355	0.881	0.808

"gene BPs/Genome length" is number of base-pairs which are in genes per genome length, "gene length average" is average of gene lengths, "number of annotated genes/ length of genome" is Number of annotated genes per genome length, "A & T content" is Frequency of nucleotide Adenine or Thymine on both strands of DNA sequence that are equal and "C & G content" is Frequency of nucleotide Adenine or Thymine on both strands of DNA sequence that are equal.

(SO) Markov chain have greater sensitivity average to detect genes than those based on first order (FO) Markov chain. The sensitivity of algorithms based on the region between stop codons is more than the ORF based models that we expected. This is as a result of the sensitivity calculations in these algorithms, which are not exact and precise (if an annotated gene in the selected region exists, it means that the gene is found). For interactions between the first factor and the second factor, selection of the region according to Figure 1 shows an obvious difference between the effects of the Markov models at the different levels of region selection. So measurement analyses were separately repeated for each type of region selection and it revealed an insignificant difference between the Markov chains with respect to the algorithms based on the ORF ($p= 0.443$) and significant difference between the Markov chains for the algorithms based on the region between stop codons ($p< 0.001$). However, for both levels of estimation of transition matrix, Markov chain models were significantly different ($p < 0.001$).

Adding covariates, i.e. "number of annotated genes", "gene base pairs per genome length", "gene length average", "number of annotated genes per length of genome", "A & T content" and "C & G content" to the repeated measurement model resulted in the p-values as indicated in Table 2. The last three covariates have produced insignificant differences between the first and second order Markov models. Therefore, it can be concluded that Markov properties are dependent on these three variables.

For interaction between the first factor, "Markov model" and the second factor, "region selection" entering covariates, "gene base pairs per genome length" and "A & T content of the genome" have produced insignificant effects ($p= 0.119$ and $p= 0.382$, respectively).

Specificity: Similar to normality of sensitivities, the Kolmogorov-Smirnov test showed that specificities of the eight algorithms are normally distributed (Table 3). Results of the repeated measurements analysis of specificities are provided in Table 4. From these data (non-covariate), it can be concluded that the effect of all three factors, the order of the Markov chain, region selection and type of estimation on specificity are significant ($p< 0.001$), which means that there is a significant difference between specificity means through the different levels of these factors. Meanwhile, it can be

Table 3. Kolmogorov-Smirnov test results for normality of the eight algorithms' specificities.

	K-S test p-values	Mean	Std. Deviation
FO-ORF-1TM	0.690	0.401511	0.1646015
FO-ORF-2TM	0.707	0.429727	0.1714144
FO-SC-1TM	0.822	0.449951	0.0864561
FO-SC-2TM	0.623	0.442779	0.0658232
SO-ORF-1TM	0.949	0.405440	0.1996034
SO-ORF-2TM	0.293	0.473767	0.1645868
SO-SC-1TM	0.848	0.781396	0.0933406
SO-SC-2TM	0.973	0.791996	0.0917712

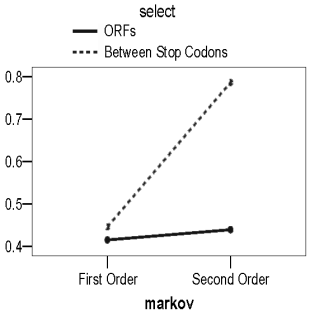
"FO" is first order, "SO" is second order, "ORF" is open reading frame, "SC" on stop codons, "1TM" on one transition matrix for whole genome and "2TM" is one transition matrix for each strand of DNA genome.

Table 4. Repeated measure analysis results for specificities with three factors, Markov model, Region selection and type of estimate transition matrix and results of this model with covariates.

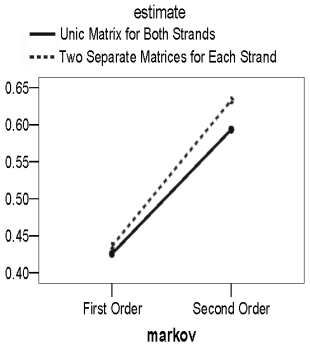
	Covariates						
Factors	None	Number of annotated genes	Gene BPs / Genome length	Gene length average	Number of annotated genes/ length of genome	A & T content	C & G content
Markov Model	0.000	0.000	0.054	0.000	0.000	0.289	0.000
Type of Selection	0.000	0.000	0.256	0.283	0.000	0.011	0.000
Type of Estimation	0.000	0.000	0.629	0.001	0.870	0.610	0.114
Markov Model* Type of Selection	0.000	0.000	0.070	0.000	0.035	0.252	0.000
Markov Model* Type of Estimation	0.000	0.000	0.905	0.022	0.886	0.045	0.770
Type of Selection* Type of Estimation	0.000	0.000	0.432	0.055	0.153	0.118	0.602
Markov Model* Type of Selection* Type of Estimation	0.092	0.114	0.199	0.002	0.046	0.296	0.083

"gene BPs / Genome length" is number of base-pairs which are in genes per genome length, "gene length average" is average of gene lengths, "number of annotated genes/ length of genome" is Number of annotated genes per genome length, "A & T content" is Frequency of nucleotide Adenine or Thymine on both strands of DNA sequence that are equal and "C & G content" is Frequency of nucleotide Adenine or Thymine on both strands of DNA sequence that are equal.

a: first factor main effect



b: first and second factor interaction



c: first and third factor interaction

Figure 2. Estimated sensitivity means derived from repeated measurement analysis.

revealed that mutual interactions between the Markov model and the other two factors also have a significant effect ($p < 0.001$). This suggests that at each level of these two factors, the difference between the Markov chains is significant.

Similar to the case of the sensitivities, Figure 2 shows that means of specificities at the SO Markov chain and algorithms based on the region between stop codons are greater than the FO Markov chain and ORF based models, respectively. Evaluation of interactions between "Markov model" and "region selection", (Figure 2) shows obvious difference between effects of Markov models in each levels of region selection. So, measurement analyses were repeated for each type of region selection that shows an insignificant difference again for specificities ($p=0.221$) between the Markov chains for algorithms based on the ORF and a significant difference between the Markov chains for algorithms based on the region between stop codons ($p < 0.001$).

DISCUSSION

In non-exact situations, sensitivity and specificity of the second Markov models for prediction of genes is better than the first Markov models, because in the algorithms based on the region between the stop codons this difference was statistically significant. But for algorithms based on ORFs which are exact methods for finding genes, the Markov models were the same, because, the first and second order Marcov models have no difference in discrimination of ORFs with respect to the cod-

Table 5. Paired t-test results for comparison of eight algorithms' sensitivity with GeneMark in 63 complete viral genomes.

Groups	Mean	p-value
FO-ORF-1TM-GeneMark	-0.1533656	0.000
FO-ORF-2TM-GeneMark	-0.1888419	0.000
FO-SC-1TM-GeneMark	-0.1807629	0.000
FO-SC-2TM-GeneMark	-0.1913987	0.000
SO-ORF-1TM-GeneMark	-0.1200170	0.001
SO-ORF-2TM-GeneMark	-0.1949939	0.000
SO-SC-1TM-GeneMark	0.0150889	0.499
SO-SC-2TM-GeneMark	-0.0314869	0.148

ing and non-coding regions. These results reveal the need for statistical research on differences between two the methods for the selection of regions.

The addition of a number of covariates to the repeated measure model showed a non-significant difference between the sensitivities of the two Markov models. This means that there are relations between the Markov models and the covariates which can assist in modeling the genome by using such information and higher order Markov Models.

Our proposed methods for application of the Markov models to prediction of genes on genome sequences basically differs from the currently in use methods such as GeneMARK because our method for prediction of initial probability is based on two types of region selection that poses some limits on analyses and produces differences between our calculations of sensitivities and specificities and other algorithms. Whilst conventional methods substantially calculate these criteria based on nucleotides which may or may not exist in a gene region, our methods is based on complete gene sequences. It should be noted that our proposed methods can only be compared with other methods with respect to sensitivities. For Comparison of sensitivity averages derived from our algorithms we have derived sensitivities from GeneMARK, that is the most popular software for gene prediction in viruses from GeneMark VIOLIN database and used paired t.test between sensitivity averages which lead to (Table 5) interesting results, in the sense that the difference between the two suggested algorithms based on "second order Markov model" and "selected region between stop codons. This suggests that our methods for prediction of initial probability and transition matrix for the Markov model is well-matched to the viral genome sequence for detecting genes.

The results of this study also show that better and simpler algorithms for solving the problem of gene prediction can be developed by adding further information to the proposed models.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*. 25: 3389-3402.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004). GenBank: update. *Nucleic Acids Res*. 32: D23-D26.

Besemer J, Borodovsky M (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. 33: W451-W454.

Borodovsky M, McIninch (1993). GENMARK: Parallel gene recognition for both DNA strands. *Computers Chem*. 17: 123-133.

Borodovsky M, McIninch (1993). Recognition of genes in DNA sequence with ambiguities. *BioSystems* 30: 161-171.

Borodovsky M, Sprizhitsky YA, Golovanov EI, Alexandrov AA (1986). Structural patterns in primary structures of the functional regions of the genome in *Escherichia coli* I. Frequency characteristics. *Mol Biol*. 20: 826-832.

Borodovsky M, Sprizhitsky YA, Golonvanov EI, Alexandrov AA (1986). Structural patterns in primary structures of the functional regions of the genome in *Escherichia coli* II. Nonuniform markov models. *Mol Biol*. 20: 833-840.

Borodovsky M, Sprizhitsky YA, Golovanov EI, Alexandrov AA. (1986) Structural patterns in primary structures of the functional regions of the genome in *Escherichia coli* III. Computer recognition of coding regions. *Mol Biol*. 20: 1144-1150.

Claverie JM (1997). Computational methods for identification of genes in vertebrate genomic sequences. *Hum Mol Genet*. 6: 1735-1744.

Durbin R, Eddy SR, Krogh A, Mitchison G (1998). *Biological sequence analysis probabilistic models of proteins and nucleic acids*. Cambridge University Press, United Kingdom. PP. 46-51.

GeneMark VIOLIN [*http://opal.biology.gatech.edu/GeneMark/VIOLIN/*].

Guo FB, Zhang CT (2006). ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics* P.

Mills R, Rozanov M, Lomsadze A, Tatusova T, Borodovsky M (2003). Improving gene annotation of complete viral genomes. *Nucleic Acid Res*. 31: 7041-7055.

Montgomery DC (2001). *Design and analysis of experiment*. John Wiley & sons, Inc, New York. PP. 189.

Pruitt KD, Tatusova T, Maglott DR (2003). NCBI Reference Sequence project: update and current status. *Nucleic Acids Res*. 31: 34-37.

Salzberg SL, Delcher AL, Kasif S, White O (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. 26: 544-548.

Yin C, Yau S (2007). Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J Theor Biol*. 247: 687-694.