

Research Article

Zebra Finch Glucokinase Containing Two Homologous Halves Is an *In Silico* Chimera

Khrustalev Vladislav Victorovich,¹ Lelevich Sergey Vladimirovich,²
and Barkovsky Eugene Victorovich¹

¹ Department of General Chemistry, Belarusian State Medical University, Dzerzinskogo 83, 220116 Minsk, Belarus

² Department of Clinical Laboratory Diagnostics, Allergy and Immunology, Grodno State Medical University, Gorkogo 80, 230009 Grodno, Belarus

Correspondence should be addressed to Khrustalev Vladislav Victorovich; vykhrustalev@mail.ru

Received 9 September 2013; Accepted 29 September 2013

Academic Editors: P. Durrens and A. Fedorov

Copyright © 2013 Khrustalev Vladislav Victorovich et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chimerical nature of the gene annotated as Zebra finch (*Taeniopygia guttata*) glucokinase (hexokinase IV) has been proved in this study. N-half of the protein encoded by that gene shows similarity with glucokinase from other vertebrates, while its C-half shows similarity with C-halves of hexokinases II. We mapped 7 new exons coding for N-half of hexokinase II and 4 new exons coding for glucokinase of Zebra finch. Finally, we reconstructed normal genes coding for Zebra finch glucokinase and hexokinase II which are situated in "head-to-tail" orientation on the chromosome 22. Because of the error in gene annotation, exons encoding N-half of normal glucokinase have been fused with exons encoding C-half of normal hexokinase II, even though they are separated from each other by the sequence 98066 nucleotides in length.

1. Introduction

Methods of phylogenetic analysis are usually used for reconstruction of the relations between distinct species or between families of homologous proteins. Nucleotide sequences of homologous genes are used as a material for fundamental works in computational biology and phylogenetic. In this study, the situation is quite different. Methods of phylogenetic analysis and methods of computational biology helped to find an error in gene annotation.

The volume of nucleotide sequences including those of complete prokaryotic and eukaryotic genomes is increasing in geometric progression in the last decade. There are many different automatic gene finding algorithms developed to annotate those sequences. Even though most of the annotations are correct, there are still some mistakes which may lead to wrong conclusions. The material from public databases should not be taken as something absolutely correct. In case if something is wrong with phylogenetic trees one should carefully recheck all the nucleotide sequences used.

There are five types of hexokinase encoded by five different genes in genomes of vertebrates: hexokinase I (HKI); hexokinase II (HKII), hexokinase III (HKIII), hexokinase domain containing protein I (HKDCI), and glucokinase (GK) [1]. HKI, HKII, HKIII, and HKDCI consist of two homologous halves. GK, which is often referred to as hexokinase IV, contains only a single "half" of hexokinase. It was shown that N-halves of HKI and HKIII are not catalytically active, unlike their C-halves [2]. In contrast, both halves of HKII are able to catalyze phosphorylation of hexoses [3].

Phylogenetic relations between glucokinase and two halves of hexokinase have been studied previously with the aim to reconstruct evolutionary history of the "standard set" (HKI, HKII, HKIII, HKDCI, and GK) of hexokinase formation [1, 4]. According to one of the hypotheses, divergence between common predecessor of all hexokinase and glucokinase happened before the duplication and fusion of the common predecessor of all hexokinase [4]. According to the more recent hypothesis [1], the common predecessor of all hexokinase and glucokinase has undergone the duplication

and fusion event. Then N-half of the glucokinase has been lost. According to the latest hypothesis, glucokinase had existed as a protein containing two homologous halves during the certain period of its evolution. The major evidence of that hypothesis would be the existence of “double” glucokinase in some species. However, such protein has not been found yet.

In this work we showed that the gene encoding glucokinase from bird Zebra finch (*Taeniopygia guttata*) genome contains two homologous halves. The initial aim of our study was to determine the time of the divergence between two homologous halves of the *Taeniopygia guttata* glucokinase. However, results of the phylogenetic part of the study showed that N-half of that “double” glucokinase is similar to “single” glucokinase, while C-half is similar to C-half of hexokinase II. Thus, the final aim of the study was to reconstruct previously unknown hexokinase II gene of *Taeniopygia guttata*.

2. Material and Methods

Predicted gene (ENSTGUG00000003490) from Ensembl database (<http://www.ensembl.org/>) encoding a product annotated as “glucokinase (hexokinase 4)” of Zebra finch (*Taeniopygia guttata*) has the following borders: 130,183–236,222. There are 18 exons predicted for this gene. The corresponding predicted protein can be found in UniProt database (<http://www.uniprot.org/>) as well (H0YZB2).

We used a nucleotide sequence (forward strand: 129,163–236,222) of chromosome 22 from *Taeniopygia guttata* genome (taeGut3.2.4) to show the distribution of exons along this region of DNA. This sequence has been cut down into windows 540 nucleotides in length. Each step of the window was equal to 60 nucleotides. GC-content has been calculated in each of those windows. We also performed calculation of GC-content in three codon positions (1GC; 2GC and 3GC) for all 18 annotated exons by “VVK Protective Buffer” algorithm [5] (<http://www.barkovsky.hotmail.ru/>).

We used Ensembl tool entitled “BLAST/BLAT” exclusively for *Taeniopygia guttata* genome (http://www.ensembl.org/Taeniopygia_guttata/blastview/) to search for nucleotide sequences similar to Chicken (*Gallus gallus*) hexokinase II and glucokinase genes. Seven additional exons encoding N-half of hexokinase II have been found by us in the region of chromosome 22 upstream to eleven exons encoding its C-half. Moreover, we found three additional exons encoding N-terminus of glucokinase and a single exon encoding its C-terminus. Levels of 1GC; 2GC and 3GC have been calculated by “VVK Protective Buffer” [5] in those newly described exons too. Splicing sites for “new” exons have been predicted by “FSPLICE” algorithm from SoftBerry server (<http://linux1.softberry.com/>). Predictions have been made for canonical splicing sites (“AG” for donor and “GT” for acceptor sites) using the data set of *Gallus gallus*.

To perform initial phylogenetic analysis for N- and C-halves of *Taeniopygia guttata* glucokinase we collected GK genes from vertebrates and hexokinase genes from *Homo sapiens* genome. To perform more thorough phylogenetic analysis with reconstructed GK and HK II of *Taeniopygia guttata* we collected all the available genes encoding hexokinase I, II, and III, as well as hexokinase domain

containing protein I from genomes of vertebrates which can be found in the Ensembl (<http://www.ensembl.org/>) data base [6]. There should be four types of hexokinase and a single glucokinase in each genome [1]. In case if some of the enzymes from the abovementioned “standard set” were absent in the Ensembl data base, we searched for them in GenBank data base using NCBI BLAST algorithm (<http://www.ncbi.nlm.nih.gov/>). Complete list of nucleotide sequences with identifiers can be found in Supplementary Material file. We avoided the usage of those sequences which are partial, as well as those (mostly from Ensembl data base) which include nucleotides designated by letter “N” (we used to call those defects of sequences “PolyN tracts”).

N-halves of hexokinase have been separated from C-halves according to the results of the “REPRO” algorithm which is able to find repeated regions in protein sequences (<http://www.ibi.vu.nl/programs/reprowww/>) [7]. N-half of Zebra finch glucokinase has been separated from its homologous C-half with the help of the same algorithm.

All the sequences were aligned by MUSCLE algorithm integrated into MEGA 5 program [8]. JTT (Jones, Taylor and Thornton) [9] amino acid evolutionary distances between all the sequences have been calculated with the help of MEGA 5 program [8]. Complete deletion mode has been chosen. Phylogenetic trees have been built by ME (Minimum Evolution) method [10] with the help of MEGA 5 program [8].

3. Results and Discussion

In Figure 1, one can see the ME phylogenetic tree for glucokinase from vertebrates and human hexokinase which have been cut into two halves. N-half of the predicted Zebra finch GK can be found on the branch with other glucokinase. It is situated in the same clade with *Gallus gallus* and *Anolis carolinensis* glucokinase, while two later proteins show more similarity with each other than with N-half of *Taeniopygia guttata* glucokinase (see Figure 1). C-half of Zebra finch GK groups together with C-half of *Homo sapiens* HK II (see Figure 1). So, *predicted glucokinase from Zebra finch is a chimera which is composed of glucokinase itself and C-half of hexokinase II*. Here we should highlight that gene coding for HK II has not been found in Zebra finch genome yet. The nature of the chimerical glucokinase/hexokinase II is not absolutely clear. Theoretically, one cannot expect that such a chimeric protein may exist *in vivo*, while, in our opinion, this chimera is the consequence of incorrect gene annotation.

The gene coding for Zebra finch glucokinase occupies 106039 nucleotides in the chromosome 22. Interestingly, first 7 exons (coding for glucokinase itself) are separated from the last 11 exons (coding for C-half of hexokinase II) by a rather long intron 98066 nucleotides in length containing several relatively GC-poor fragments (see Figure 2). According to the results of Ensembl BLAT/BLAST analysis, 7 “new” exons coding for N-half of hexokinase II exist in the 3'-end of this long “pseudointron”. Their coordinates in chromosome 22 (in nucleotides) are from 223,162 to 223,331; from 226,647 to 226,694; from 227,122 to 227,192; from 228,944 to 229,043; from 229,153 to 229,258; from 229,519 to 229,698; from

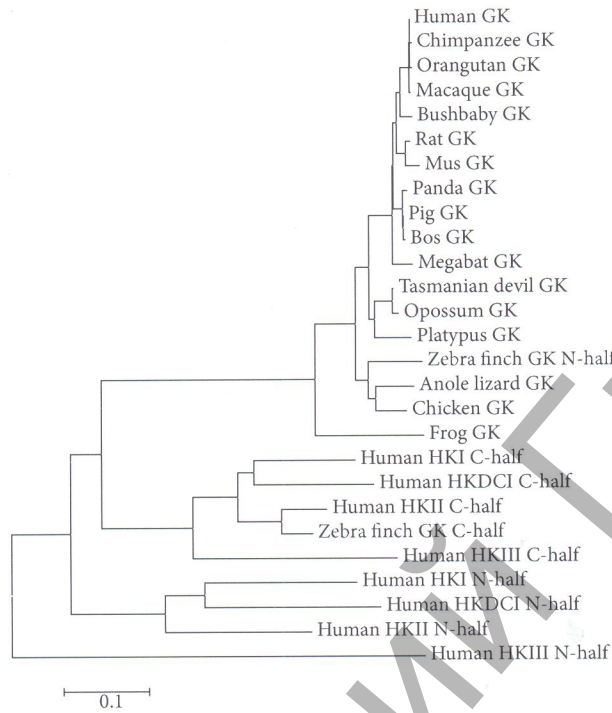


FIGURE 1: Minimum evolution (ME) phylogenetic tree built on the basis of JTT evolutionary distances (complete deletion) between sequences of N- and C-halves of human hexokinase, glucokinase from vertebrates, and N- and C-halves of *Taeniopygia guttata* glucokinase aligned by MUSCLE algorithm.

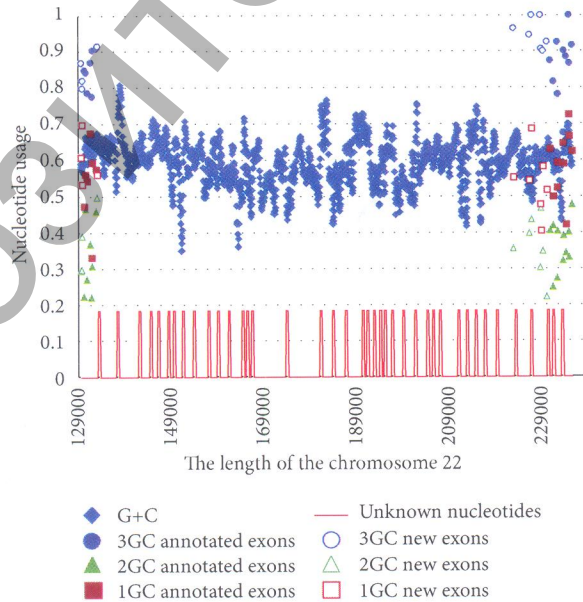


FIGURE 2: Distribution of GC content along the length of the region of chromosome 22 containing exons annotated in Ensembl and newly annotated exons coding for glucokinase and hexokinase II of *Taeniopygia guttata*. GC-content distribution between three codon positions (1GC; 2GC; 3GC) is given for each exon. Defects of sequence (unknown nucleotides, i.e., “polyN” tracts) are also shown.

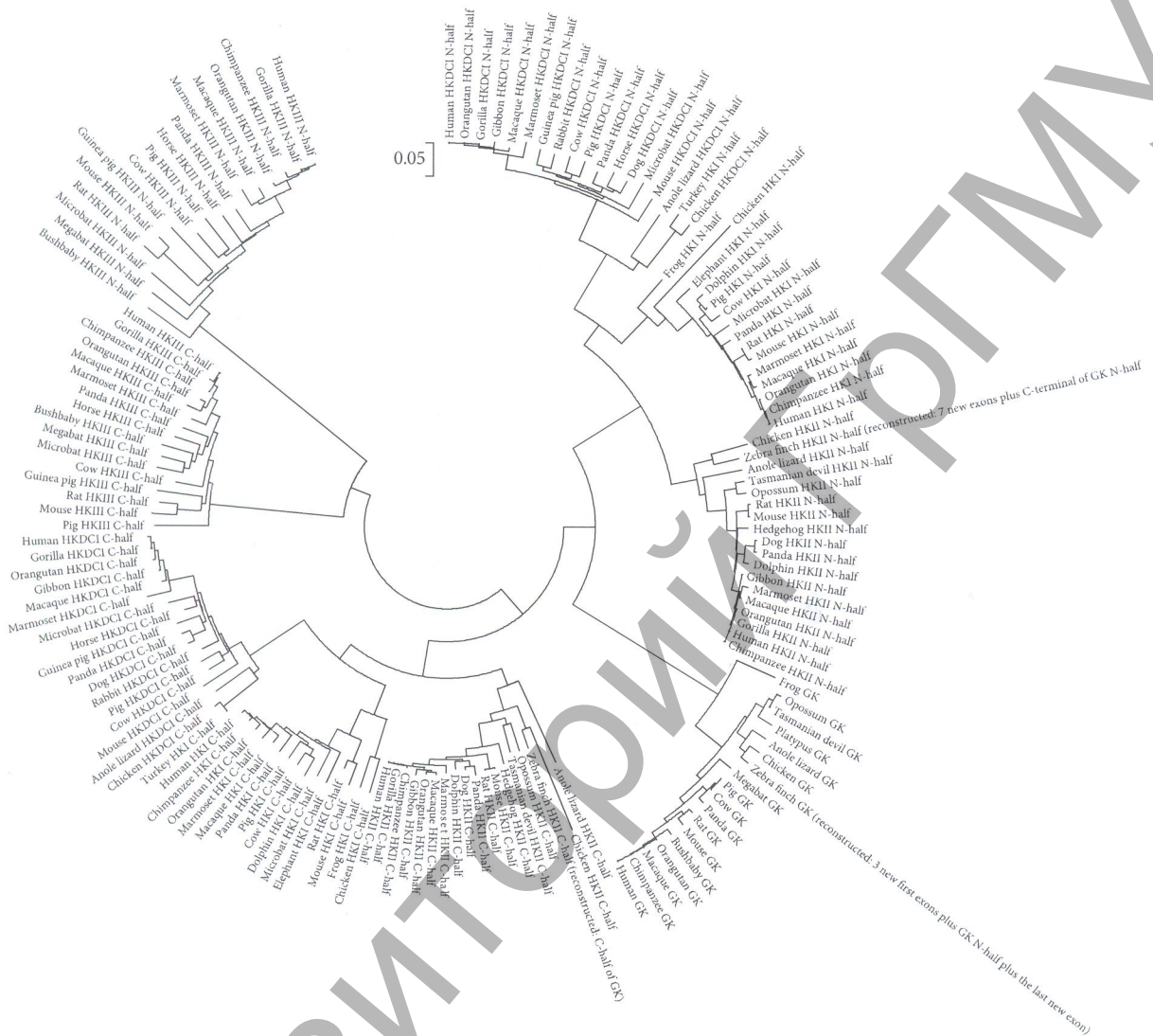


FIGURE 3: Minimum evolution (ME) phylogenetic tree built on the basis of JTT evolutionary distances (complete deletion) between sequences of N- and C-halves of hexokinase and glucokinase from vertebrates together with reconstructed N- and C-halves of *Taeniopygia guttata* hexokinase II and glucokinase aligned by MUSCLE algorithm.

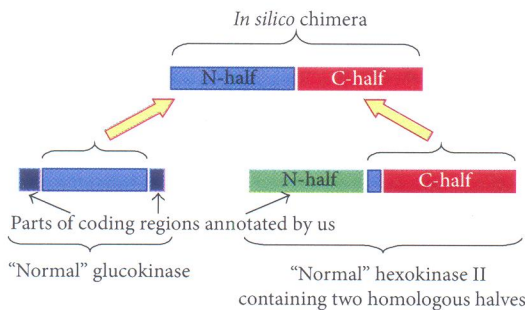


FIGURE 4: The scheme of the gene annotation defect which led to the *in silico* chimera formation.

230,423 to 230,499 (see Figure 2). Acceptor splicing sites have been found by “FSPLICE” algorithm for the first, fourth, fifth and sixth exons. Donor splicing sites have been found for the second, third, fourth, and fifth exons.

Three additional exons encoding N-terminus of the Zebra finch glucokinase occupy the following locations in chromosome 22 (in nucleotides): from 129,426 to 129,589; from 129,709 to 129,767; from 129,782 to 129,889. Acceptor splicing sites have been found for the first and second exons; donor splicing sites have been found for the first and third ones. There is a single nucleotide deletion in the second exon, relative to sequences of other glucokinase. That kind of deletion (which is, probably, a sequencing defect) resulted in frameshifting.

We found out that 40 amino acid residues from the C-terminus of Zebra finch glucokinase N-half actually belong to the C-terminus of hexokinase II N-half. “Original” Zebra finch glucokinase C-terminus has been reconstructed by us. It is encoded by newly mapped exon. That “new” exon has been found in chromosome 22: from the nucleotide 132,945 to the nucleotide 133,088 (see Figure 2). There is a donor splicing site in its 5'-end and a terminal codon in its 3'-end.

There were three gaps in the reconstructed amino acid sequence of hexokinase II N-half. It is likely that those gaps occurred due to defects in nucleotide sequence determination: there are areas with “PolyN” tracts situated between the first and second, between the second and third and between the sixth and seventh exons (see Figure 2).

In general, there is specific pattern of GC-content distribution between three codon positions in both previously mapped and newly mapped exons: 3GC > 1GC > 2GC. This pattern is characteristic for genes which are under the influence of mutational GC-pressure [11–13]. Exons coding for both glucokinase and hexokinase II seem to be situated in GC-rich isochores [14] of Zebra finch chromosome 22. 3GC levels for three of the exons even reached 100% (see Figure 2), while average 3GC level for both “new” and already annotated exons encoding hexokinase II is equal to $87.91 \pm 4.97\%$. In case if exons encoding N-half of hexokinase II were inactive, 2GC and 1GC levels would grow to the higher level under the influence of GC-pressure [13]. However, average 2GC level for 7 “new” hexokinase II exons is not significantly higher than that for 11 already annotated exons ($36.38 \pm 7.29\%$ versus $36.72 \pm 4.02\%$; $P = 0.94$). The same situation is characteristic for their average 1GC levels ($53.66 \pm 7.65\%$ versus $58.43 \pm 5.23\%$; $P = 0.34$).

N-half of Zebra finch hexokinase II reconstructed by us occupies correct branch (it groups with N-half of *Gallus gallus* hexokinase II) in ME tree from Figure 3, as well as reconstructed glucokinase and C-half of hexokinase II do. So, in our opinion, Zebra finch possesses normal functional glucokinase gene and normal functional hexokinase II gene which are situated on the same chromosome 22 near each other.

Interestingly, in Figure 3 both N- and C-halves of Turkey hexokinase I can be found on branches with N- and C-halves of HKDCI (together with Chicken homologues). This fact is the evidence that the gene from Turkey genome annotated as that coding for hexokinase I is

actually coding for hexokinase domain containing protein I (HKDCI).

Similar error has already been described by us for pyruvate kinases. A protein annotated as Lamprey liver pyruvate kinase shows more similarity with muscle pyruvate kinases of vertebrates than with liver pyruvate kinases [15].

Wrong determination of isoenzyme type is relatively common mistake in gene annotation, while the error described in the present work is a weird one.

One of the benefits of automatic gene annotation is the absence of mistakes caused by so-called “human factor”. However, in certain cases, such as in the case described in this work, corrections of automatic annotations should be introduced “manually” with the help of computational biology methods.

4. Conclusions

Gene annotation algorithm “fused” most of the exons coding for glucokinase of Zebra finch with exons coding for C-terminus (C-terminus of N-half and the whole C-half) of its hexokinase II situated downstream on the same chromosome 22 (as it is schematically represented in Figure 4) and predicted a relatively long “pseudointron” in the place of intergenic spacer.

References

- [1] D. M. Irwin and H. Tan, “Molecular evolution of the vertebrate hexokinase gene family: identification of a conserved fifth vertebrate hexokinase gene,” *Comparative Biochemistry and Physiology D*, vol. 3, no. 1, pp. 96–107, 2008.
- [2] H. J. Tsai, “Functional organization and evolution of mammalian hexokinases: mutations that caused the loss of catalytic activity in N-terminal halves of type I and type III isozymes,” *Archives of Biochemistry and Biophysics*, vol. 369, no. 1, pp. 149–156, 1999.
- [3] K. J. Ahn, J. Kim, M. Yun, J. H. Park, and J. D. Lee, “Enzymatic properties of the N- and C-terminal halves of human hexokinase II,” *BMB Reports*, vol. 42, no. 6, pp. 350–355, 2009.
- [4] M. L. Cárdenas, A. Cornish-Bowden, and T. Ureta, “Evolution and regulatory role of the hexokinases,” *Biochimica et Biophysica Acta*, vol. 1401, no. 3, pp. 242–264, 1998.
- [5] V. V. Khrustalev, M. Arjomandzadegan, E. V. Barkovsky, and L. P. Titov, “Low rates of synonymous mutations in sequences of mycobacterium tuberculosis GyrA and KatG genes,” *Tuberculosis*, vol. 92, no. 4, pp. 333–344, 2012.
- [6] P. Flicek, M. R. Amode, and K. Beal, “Ensembl 2012,” *Nucleic Acids Research*, vol. 40, pp. D84–D90, 2012.
- [7] R. A. George and J. Heringa, “The REPRO server: finding protein internal sequence repeats through the web,” *Trends in Biochemical Sciences*, vol. 25, no. 10, pp. 515–517, 2000.
- [8] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, “MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods,” *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.
- [9] D. T. Jones, W. R. Taylor, and J. M. Thornton, “The rapid generation of mutation data matrices from protein sequences,”

- Computer Applications in the Biosciences*, vol. 8, no. 3, pp. 275–282, 1992.
- [10] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.
- [11] N. Sueoka, “Directional mutation pressure and neutral molecular evolution,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2653–2657, 1988.
- [12] O. K. Clay and G. Bernardi, “GC3 of genes can be used as a proxy for isochore base composition: a reply to Elhaik et al,” *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 21–23, 2011.
- [13] V. V. Khrustalev and E. V. Barkovsky, “An *in-silico* study of alphaherpesviruses ICP0 genes: positive selection or strong mutational GC-pressure?” *IUBMB Life*, vol. 60, no. 7, pp. 456–460, 2008.
- [14] M. Costantini and G. Bernardi, “Replication timing, chromosomal bands, and isochores,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3433–3437, 2008.
- [15] S. V. Lelevich, V. V. Khrustalev, E. V. Barkovsky, and T. A. Shedogubova, “The influence of ethanol on pyruvate kinases activity *in vivo*, *in vitro*, *in silico*,” *American Journal of Medical and Biological Research*, vol. 1, pp. 6–15, 2013.

РЕПОЗИТОРИЙ ГРГМУ