BIOINFORMATICS

Bioinformatic analysis of inverted repeats of coronaviruses genome

O. Yu. Limanskaya[1, 2]

[1]Mechnikov Institute of Microbiology and Immunology AMS of Ukraine
14, Pushkinska Str., Kharkiv 61057 Ukraine

[2]National Scientific Centre "Institute of Experimental and Clinical Veterinary Medicine", Ukrainian Academy of Agrarian Sciences
83, Pushkinska Str., Kharkiv 61023 Ukraine

olga.limanskaya@mail.ru

Aim. *To design the maps of matched and mismatched potential hairpin structures in the genomes of human and animal coronaviruses.* ***Methods.*** *Bioinformatic analysis of coronaviruses nucleotide sequences, atomic force microscopy.* ***Results.*** *Thermodynamically stable matched and mismatched inverted repeats forming hairpin structures that can appear in genomic RNA of the human and animal coronaviruses (severe acute respiratory syndrome virus, murine hepatitis virus, porcine epidemic diarrhea virus, transmissible gastroenteritis virus, and bovine coronavirus) are determined. The maps of hairpin localization (which are a part of the genome signaling mechanisms) are obtained for the genome of coronaviruses.* ***Conclusions.*** *The genes encoding replicase and spike glycoproteins of coronaviruses are the main sites of the localization of potential conservative structural motives. The hairpins are shown to be conservative structural elements inside the set of coronavirus isolates of one species.*

Keywords: severe acute respiratory syndrome virus, coronavirus, hairpin structure, inverted repeat

Introduction. Similar to other non-canonical formations, hairpin-loop structures, which can be formed in nucleic acids by inverted repeats, are significant genome elements, playing a specific biological role. It is believed that they are involved in the regulation of DNA replication and transcription [1, 2].

The method of fluorescent flow cytometry allowed determining ~10^5 hairpin-loop structures in the nucleus [3].

Besides a specific role of matched and mismatched inverted repeats in mutagenesis, they are associated with a series of human genetic diseases (hereditary angioneurotic edema, antithrombin deficiency, and deficiency of human serum cholinesterase) [4].

Regardless of rather intense study on palindrome localization in genome of different organisms, the role and distribution of hairpin structures in genome of viruses and bacteria are still to be determined. Therefore, we have searched for potential hairpin structures in coronaviruses genome. The family of coronaviruses

(CoV) includes porcine epidemic diarrhea virus, infectious bronchitis virus, murine hepatitis virus, and transmissible gastroenteritis virus. There are also coronaviruses of humans, cattle, horses, and cats. Sequencing proved that severe acute respiratory syndrome virus (SARS-CoV) may also be related to *Nidovirales* order, *Coronaviridae* family, *Coronavirus* genus.

Coronaviruses are divided into three serogroups, each having cross serological reactivity and similar genome organization. All known human coronaviruses belong to groups I and II. SARS-CoV forms a new group IV, since [the] performed genetic and antigenic researches demonstrated its being distant from all the known groups of coronaviruses [5].

[The] Current work presents distribution of matched and mismatched inverted repeats in the genome of certain coronaviruses, highly dangerous for humans, [the] severe acute respiratory syndrome virus, in particular. An analysis of the maps of potential hairpin structures showed that the distribution of inverted repeats is the same within the set of coronavirus isolates of one species.

Therefore, the comparison of the distribution of hairpin structures may serve as another instrument (along with philogenetic analysis) in the research of evolutionary relations and genome organization of both coronaviruses and representatives of other species.

Materials and Methods. Complete sequences of isolates of severe acute respiratory syndrome virus (SARS-CoV) (number AY27848, AY279354, AY268070 of GenBank database), bovine coronavirus (AF220295), transmissible gastroenteritis virus (NC_002306), infectious bronchitis virus (AY251817, AY251816, AF391157, AF391156, AF391154, AY237817, AY223860, AF470630, AF4, 70629, AF470628, AF467921), porcine epidemic diarrhea virus (NC_003436), murine coronavirus (NC_0018460), feline coronavirus (AY204523, AY204524, AY204525) as well as *pGEMEX* plasmids (X65317) were used in the work.

Oligo (version 3.4) [6] and RNA2 of GeneBee package [7] were used to search for matched and mismatched inverted repeats and to determine their thermodynamic parameters.

Atomic force microscope (AFM) Nanoscope III with D-scanner (*Veeco Instruments Inc.*, USA) was used. AFM images of the sample of supercoiled *pUC8* plasmid DNA (2665 bp) after application on standard amino amica were captured in the air in “height” mode using tapping variant of AFM and unsharpened probes of *KTEK International company* (Russian Federation) with resonance frequency of 300–360 kHz. The sample was prepared according to the method, previously described in [8].

Results and Discussion. It has been revealed that hairpin-loop structures may be a part of promoters and transcription termination sites as the presence of cruciforms is a signal for the stop of RNA-polymerase and termination of synthesis of RNA-transcripts with subsequent dissociation of the complex, formed by RNA-polymerase and DNA-RNA-transcripts. One of the mentioned transcription terminators for T7 RNA-polymerase is the site of transcription termination of *pGEMEX* plasmid which is an internal transcription terminator, ~ 90 bp long, with the efficiency of 70–80% [9]. The analysis of the site of transcription termination of *pGEMEX* DNA for the presence of thermodynamically stable inverted repeats allowed us to find a mismatched inverted repeat of 28 bp long, with the free energy – G = 11.2 kcal/mol. The termination of T7 RNA-polymerase transcription with elongation of transcription on *pGEMEX* DNA matrix, containing this inverted repeat in the site of the terminator, has been previously demonstrated *in vitro* [10]. Therefore, taking into consideration the parameters of *pGEMEX* DNA hairpin and literature data concerning the parameters of hairpins in hairpin-loop structures, observed in the course of *in vivo* [11] and *in vitro* [12] experiments, hairpins with the loop length of 5 nucleotides and minimal energy – G ~ 9 kcal/mol were selected for further analysis.

The diagram of their distribution on the physical map of SARS genome (Fig.1) was built on the basis of determined potential (i.e. thermodynamically stable) hairpins in SARS virus genome (Table). It is noteworthy that the hairpins determined are conservative structural motives for SARS virus. The comparison of their localization on genome of several SARS isolates showed that their location is the same for the majority of hairpins. In our opinion, it may serve as evidence to a

Matched and mismatched thermodynamically stable hairpin-like structures, which may possibly be formed by inverted repeats, in genomic RNA of severe acute respiratory syndrome virus (number AY291451) for GenBank database

#	Stem length, bp	Loop length, n.	-G, kcal/mol	Location on genome	Protein	Type of repeat
1	15	4	16,9	2570–2603	Replicase A	Mismatched-2
2	14	4	12,6	2959–2990	Replicase A	Matched-1
3	17	3	14,5	2894–2929	Replicase A	Mismatched-1
4	15	4	11,5	3261–3294	Replicase A	Mismatched-1
5	10	3	12,7	3481–3503	Replicase A	Matched-1
6	16	4	12,3	4051–4086	Replicase A	Matched-1
7	15	4	14,6	5668–5701	Replicase A	Mismatched-1
8	12	6	15,1	5338–5367	Replicase A	Mismatched-2
9	20	5	17,2	6983–7028	Replicase A	Mismatched-2
10	11	5	11,5	7140–7166	Replicase A	Matched-1
11	11	4	10,3	8634–8659	Replicase A	Matched-1
12	14	4	12,3	10921–10952	Replicase A	Matched-1
13	12	7	10,3	12314–12344	Replicase A	Mismatched-1
14	11	3	10,7	13363–13387	Replicase A	Mismatched-1
15	13	5	13,8	13404–13434	Replicase A	Mismatched-1
16	**14**	**4**	**16,3**	**13888–13919**	**Replicase B**	**Mismatched-2**
17	**16**	**3**	**15,7**	**13945–13979**	**Replicase B**	**Mismatched-2**
18	10	3	12,7	16460–16482	Replicase B	Mismatched-1
19	11	4	11,4	17166–17191	Replicase B	Mismatched-1
20	17	4	16,1	17521–17558	Replicase B	Mismatched-2
21	8	3	11,5	20230–20248	Replicase B	Matched-1
22	15	6	11,3	21376–21411	Replicase B	Mismatched-1
23	11	4	11,8	23447–23472	Glycoprotein S	Mismatched-1
24	17	5	16,9	23636–23674	Glycoprotein S	Mismatched-2
25	19	4	16,6	24075–24116	Glycoprotein S	Mismatched-2
26	10	6	10,4	26102–26127	Gene E	Matched-1

Note. Types of repeats mismatched-1 (matched-1), mismatched-2 correspond to hairpins with the energy value (– G) of over 10 and 15 kcal/mol, respectively. Locations of hairpins, sequences and secondary structures of which are presented in Fig.2, are in bold.

specific role of hairpin structures in the chain of signalling mechanisms of SARS virus functioning.

All the repeats analyzed were divided into two types – matched and mismatched ones (the stem of latter contains non-complementary nucleotides or deletions of nucleotides in one of the chains of hairpin stem). Besides, the repeats were differentiated into three groups according to their energy level. The first,

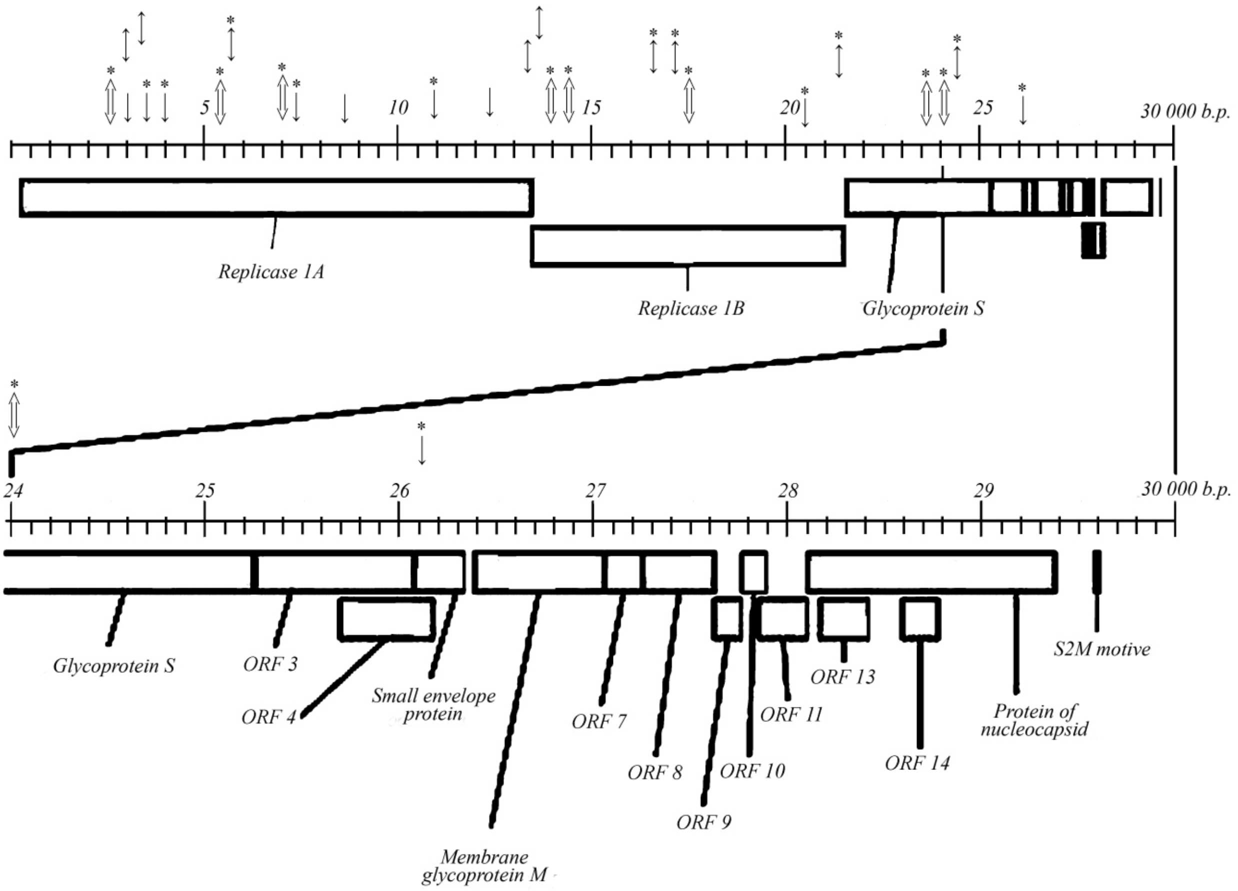

Fig.1. Physical map of severe acute respiratory syndrome virus (number AY291451) with indicated locations of known genes. Arrows indicate the location of determined thermodynamically stable matched and mismatched hairpin structures; asterisks indicate hairpin structures, the location of which coincides with that of similar structures of another isolate of SARS virus (number AY278488); ↕, ⇕ - mismatched hairpins with the energy - G over 10 and 15 kcal/mol respectively; - matched hairpins with the energy over 10 kcal/mol.

second, and third groups consisted of repeats with the energy (– G) of 10–15, 15–20, and 20 kcal/mol, respectively.

The sequence and secondary structure of two typical mismatched inverted repeats, the energy of which exceeded –15 kcal/mol, are presented in Fig.2. We used the same method to obtain the diagrams of distribution and parameters of hairpin-like structures for bovine coronavirus (Fig.3), murine coronavirus (Fig.4, *a*), porcine epidemic diarrhea virus (Fig.4, *b*), and transmissible gastroenteritis virus (Fig.4, *c*).

It should be mentioned that approximately two thirds of coronavirus genome is a matrix for the synthesis of replicases 1A and 1B, one third of the genome encodes structural proteins (nucleoprotein, spike glicoproteins S, M, and E) as well as a series of non-structural proteins (Fig.1).

Among analyzed sequences of animal coronavirus isolates and SARS virus, the highest susceptibility to forming hairpin-like structures was revealed for SARS virus, and a possibility of forming up to 26 structures for one isolate was demonstrated (Table). The majority of hairpin-like structures (20 out of 26) can be formed in the genes, encoding replicase (Fig.1).

The authors of [13] used computer modelling (the program, predicting secondary RNA structure) to investigate the secondary and tertiary structures of 3'-untranslated region (UTR) of genome RNA of SARS-CoV, structural elements of which play a significant role in replication of viruses, [as] compared to

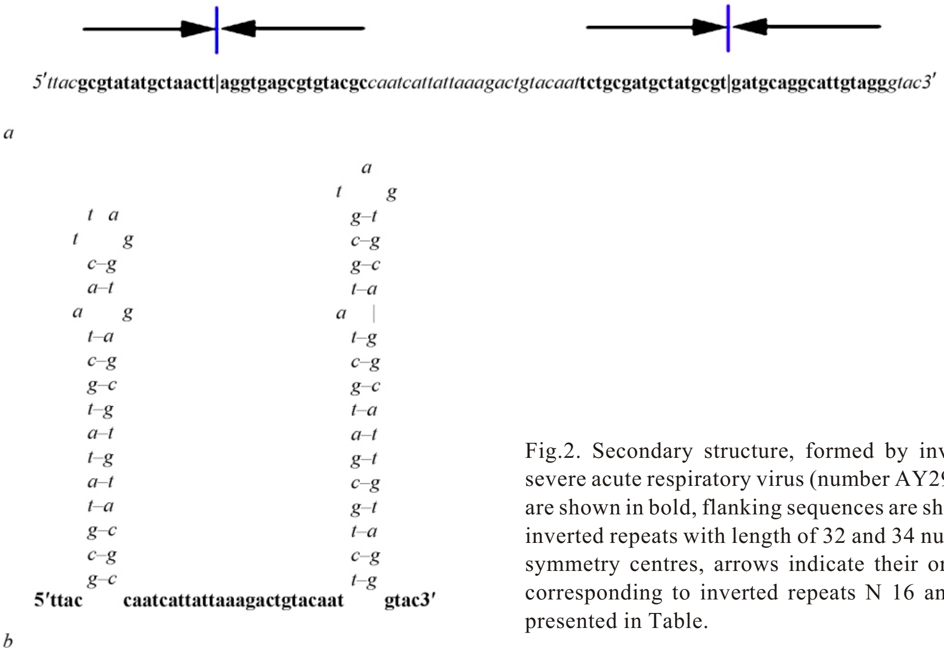
5′ttacgcgtatatgctaactt|aggtgagcgtgtacgccaatcattattaaagactgtacaattctgcgatgctatgcgt|gatgcaggcattgtagggtac3′
a
5′ttac caatcattattaaagactgtacaat gtac3′
b

Fig.2. Secondary structure, formed by inverted repeats in the fragment of severe acute respiratory virus (number AY291451) (complementary sequences are shown in bold, flanking sequences are shown in italics): *a* - two mismatched inverted repeats with length of 32 and 34 nucleotides (vertical lines show their symmetry centres, arrows indicate their orientation); *b* - hairpin structures, corresponding to inverted repeats N 16 and N 17, parameters of which are presented in Table.

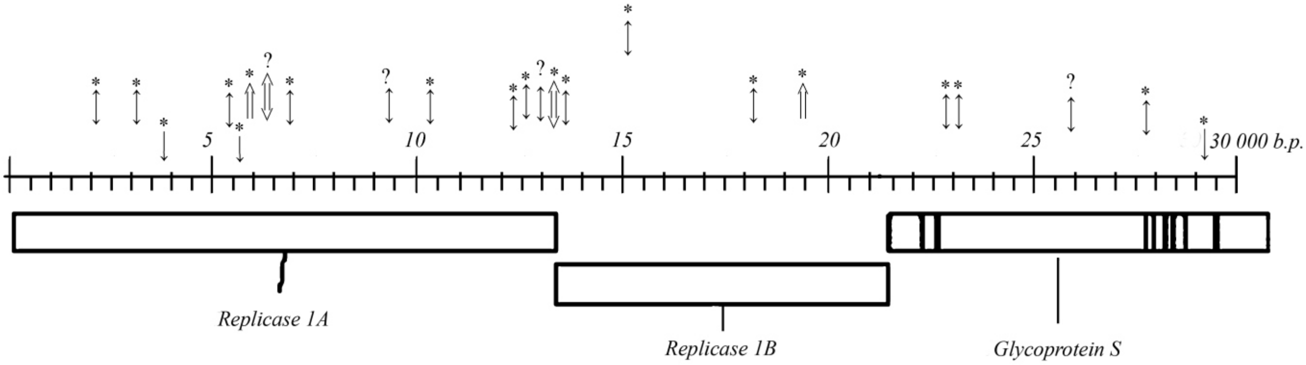

Fig.3. Physical map of bovine coronavirus (number NC_003405) with indicated locations of known genes. Arrows indicate the location of hairpin structures; asterisks indicate hairpin structures, the location of which coincides with that of similar structures of another isolate of bovine coronavirus (number AF220295); question mark corresponds to hairpin structures, the location of which does not coincide for two mentioned isolates of bovine coronavirus; ↕, ⇕ - mismatched hairpins with the energy (- *G*) over 10, 15, and 20 kcal/mol respectively; - matched hairpins with the energy over 10 kcal/mol.

structures of the same regions of previously characterized coronaviruses. Three conservative structural motives were shown in 3'UTR – hairpin-like structures and a single local secondary hairpin-like structure, formed in the course of folding 3'-end RNA fragments of SARS-CoV.

It should be mentioned that the structure of a hairpin, determined by computer analysis, depends on both the search algorithm used and the hairpin parameters. Therefore, contrary to the authors of [13], we concentrated our efforts on the search for thermodynamically stable matched and mismatched repeats, i.e. the ones, actually observed in previous experiments. We did not consider mismatched repeats with more than two pairs of non-paired nucleotides and with the size of a loop exceeding six nucleotides. The example of such thermodynamically stable hairpin is a hairpin-loop structure, formed by two hairpin structures in supercoil *pUC8* plasmid (Fig.5). *pUC8* plasmid contains several inverted repeats which can form hairpin-loop structures. Free energy G of the most stable structure (indicated with arrows in Fig.5) is –17.8 kcal/mol, 11 bp form the

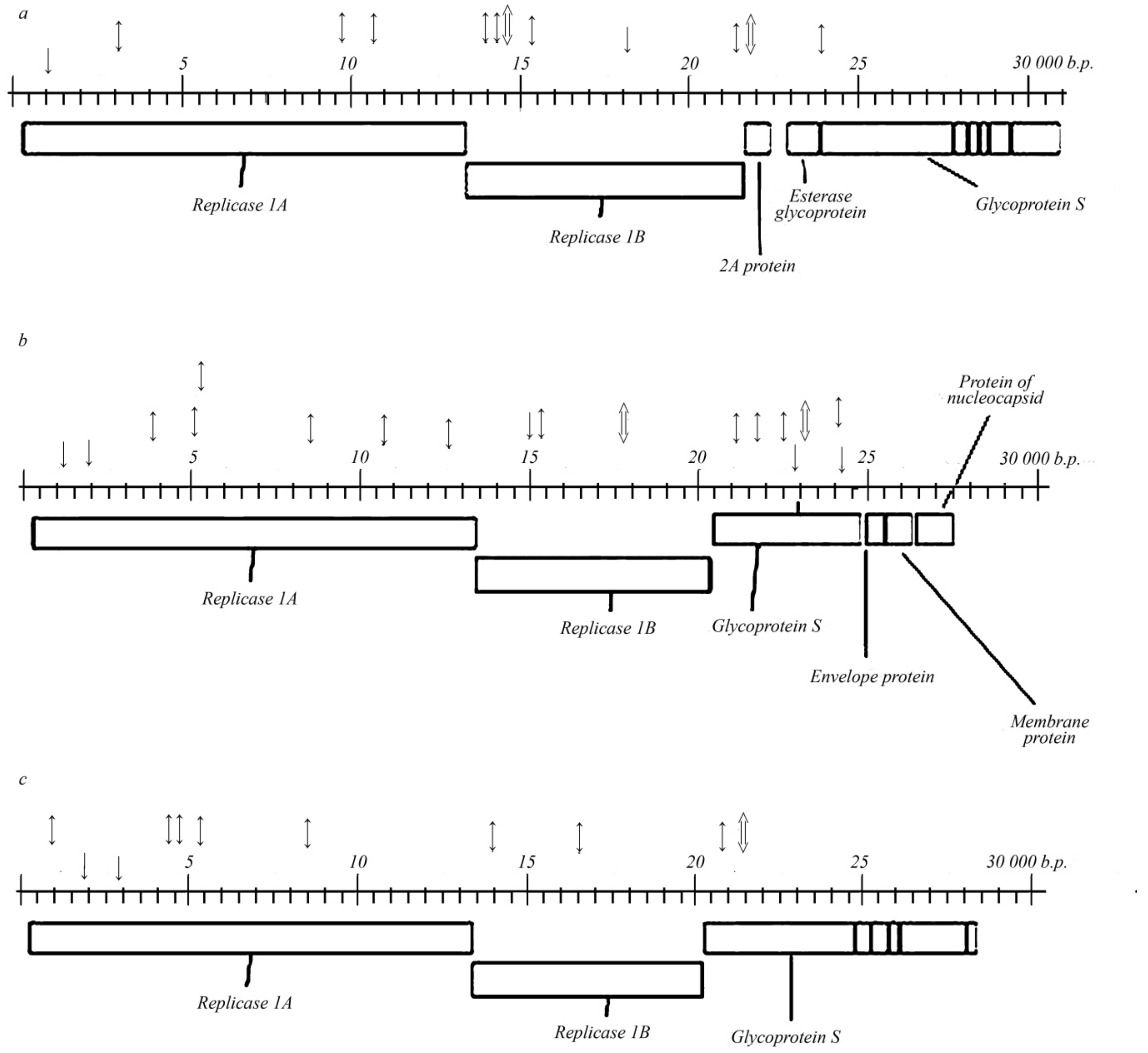

Fig.4. Physical maps of murine hepatitis virus (NC_001846) (*a*), porcine epidemic diarrhea virus (number NC_003436) (*b*), and transmissible gastroenteritis (number NC_002306) (*c*) with indicated location of known genes. Arrows indicate the location of hairpin structures; ↕, ⇕ - mismatched hairpins with the energy over 10 and 15 kcal/mol, respectively; - matched hairpins with the energy over 10 kcal/mol.

stem of the hairpin, and the loop contains four nucleotides. However, G–T pair is not considered as non-complementary one in RNA2 program of GeneBee software, used by us to predict the secondary structure of coronavirus RNA. The formation of G–T Watson-Crick pair is possible due to the formation of rare tautomer enol and iminoforms of nucleotides [14]. [The] Presented AFM image of hairpin-loop structure of *pUC8* DNA demonstrates that among several palindromes, which are a part of *pUC8* DNA, the only one and the most stable thermodynamically, hairpin-loop structure is formed *in vitro*.

The possibility of formation of 23 thermodynamically stable conservative structural motives was shown for genomic RNA of bovine coronavirus (Fig.3). The location of the majority of these motives

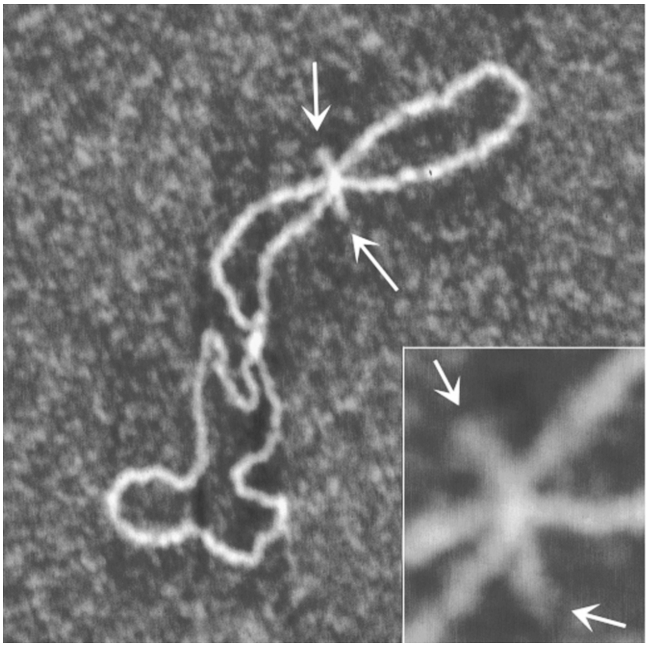

Fig.5. AFM image of supercoiled *pUC8* plasmid in the air. Image size - 372 nm x 372 nm. Arrows indicate two hairpins, forming hairpin-loop structure. The insert size with zoomed image of hairpins is 55 nm x 62 nm.

coincides with the location of hairpins for another isolate of bovine coronavirus (indicated in Fig.3) similar to isolates of SARS virus.

The analysis of genomic RNA of murine hepatitis virus (Fig.4, *a*) allowed revealing 12 hairpin-like structures in this sequence. The investigation of genomic RNA of porcine epidemic diarrhea virus proved the possibility of forming 18 hairpins (Fig.4, *b*), 10 of which are located in the site of the gene, encoding replicase.

The investigation of a complete sequence of genomic RNA of another porcine coronavirus – transmissible gastroenteritis virus – testifies to the possibility of existence of 11 hairpin-like structures, nine of which are also located in the gene, encoding replicase (Fig.4, *c*).

The composition of localization maps of inverted repeats brings up several questions. Firstly, two hairpins at 5'- and 3'-ends of DNA matrix chain are sufficient for initiation and termination of transcription. At the same time, the sequence of gene, encoding replicase A of bovine coronavirus, contains 14 hairpins. Therefore, a biological function of the majority of hairpins revealed is yet to be defined. Secondly, the absence of hairpins at 5'-end of the gene of replicase A of SARS virus (Fig.1), bovine coronavirus (Fig.3) testifies to the fact that similar[ly] to hairpins, other non-canonical DNA structures (triplexes, in particular) may serve as signals for enzyme binding. Thirdly, SARS virus differs from other coronaviruses both in qualitative character of distribution of hairpins and in their quantitative parameters. For instance, the number of highly stable mismatched inverted repeats with the energy – G over 15 kcal/mol for SARS virus is seven, while only two repeats with the free energy – G over 20 kcal/mol and one repeat with the energy over 15 kcal/mol were found for bovine coronavirus. The abovementioned testifies to the possibility of using the distribution of thermodynamically stable inverted repeats for the purpose of structural differentiation of viruses.

Therefore, the computer analysis of isolates of different animal coronaviruses and SARS allowed [of] determining the main localization sites of potential conservative structural motives to be genes of replicase and spike glycoproteins of coronaviruses. We believe that the comparison of the distribution of hairpin structures may serve as another instrument (along with philogenetic analysis) in the research of evolutionary relations and genome organization of both coronaviruses and representatives of other species. Besides, the maps of hairpin distribution obtained may be important for further investigation on molecular mechanisms and regulatory role of such alternative structures as hairpin and hairpin-loop structures, which are often discussed in literature [15, 16].

The work was partially supported by grants of AMS 47/2002 and AMS 72/2007 of the Academy of Medical Sciences of Ukraine.

О. Ю. Лиманська

Біоінформатичний аналіз інвертованих повторів геному коронавірусів

Резюме

Мета. *Створення карт локалізації досконалих і недосконалих потенційних шпилькових структур у геномі коронавірусів людини і тварин.* ***Методи****. Біоінформатичний аналіз нуклеотидних послідовностей коронавірусів, атомно-силова мікроскопія.* ***Результати****. Визначено термодинамічно стабільні досконалі* *та недосконалі інвертовані повтори, які утворюють шпилькові структури, що можуть виникати у геномній РНК коронавірусів людини і тварин – вірусів тяжкого гострого респіраторного синдрому, гепатиту миші, епідемічної діареї свині, трансмісивного гастроентериту та бичачого коронавірусу. Створено карти локалізації шпильок (які є одним із ланцюгів сигнальних механізмів функціонування геному) на геномі коронавірусів.* ***Висновки****. Основними сайтами локалізації потенційно можливих консервативних структурних мотивів є гени реплікази та гліко- протеїнів шипів коронавірусів. Шпилькові структури є консервативними елементами всередині набору ізолятів одного виду коронавірусів.*

Ключові слова: вірус тяжкого гострого респіраторного синдрому, коронавірус, шпилькова структура, інвертований повтор.

О. Ю. Лиманская

Биоинформатический анализ инвертированных повторов генома коронавирусов

Резюме

Цель*. Создание карт локализации совершенных и несовершенных потенциальных шпилечных структур в геноме коронавирусов человека и животных.* ***Методы****. Биоинформатический анализ нуклеотидных последовательностей коронавирусов, атомно-силовая микроскопия.* ***Результаты****. Определены термодинамически стабильные совершенные и несовершенные инвертированные повторы, образующие шпилечные структуры, которые могут возникать в геномной РНК коронавирусов человека и животных – вирусов тяжелого острого респираторного синдрома, гепатита мыши, эпидемической диареи свиньи, трансмиссивного гастроэнтерита и бычьего коронавируса. Созданы карты локализации шпилек (являющихся одной из цепей сигнальных механизмов функционирования генома) на геноме коронавирусов.* ***Выводы****. Основными сайтами локализации потенциально возможных консервативных структурных мотивов служат гены репликазы и гликопротеинов шипов коронавирусов. Шпилечные структуры являются консервативными элементами внутри набора изолятов одного вида коронавирусов.*

Ключевые слова: вирус тяжелого острого респираторного синдрома, коронавирус, шпилечная структура, инвертированный повтор.

REFERENCES

1. *McClellan J., Boublikova P., Palecek E., Lilley D.* Superhelical torsion in cellular DNA response directly to environmental and genetic factors // Proc. Nat. Acad. Sci. USA.–1990.–**87**, N 21.–P. 8373–8377.
2. *Bagga R., Ramesh N., Brahmachari S.* Supercoil-induced unusual DNA structures as transcriptional block // Nucl. Acids Res.–1990.–**18**, N 11.–P. 3363–3369.
3. *Ward G., McKenzie R., Zannis-Hadjopoulos M., Price G.* The dynamic distribution and quantification of DNA cruciforms in eukaryotic nuclei // Exp. Cell Res.–1990.–**188**, N 2.–P. 235–246.
4. *Bessler J.* DNA inverted repeats and human disease // Frontiers in Biosci.–1998.–N 3.–P. d408–d418.
5. *Poon L., Guan Y., Nicholls J., Yuen K., Peiris J.* The aetiolody, origins, and diagnosis of severe acute respiratory syndrome // Lancet Infect. Dis.–2004.–**4**, N 11.–P. 663–671.
6. *Rychlik W., Spencer W., Rhoads R.* Optimization of the annealing temperature for DNA amplification *in vitro* // Nucl. Acids Res.–1990.–**18**, N 21.–P. 6409–6412.
7. *Brodsky L., Drachev A., Tatuzov R., Chumakov K.* The package of programs for biopolymer sequence analysis: GeneBee // Biopolymers and Cell.–1991.–**7**, N 1.–P. 10–14.
8. *Limanskii A. P.* Study of cruciform structure in supercoiled *pUC8* plasmid DNA by atomic force microscopy and computer modelling // Biopolymers and Cell.–2002.–**18**, N 5.–P. 401–405.
9. *pGEMEX-1* and *pGEMEX-2* vectors // Techn. Bull., Promega.–2000.–N 253.–P. 1–13.
10. *Limanskii A. P.* Visualization of DNA–T7 RNA polymerase complex by atomic force microscopy // Biopolymers and Cell.–2007.–**23**, N 1.–P. 3–13.
11. *Panayotatos N., Fontaine A.* A native cruciform DNA structure probed in bacteria by recombinant T7 endonuclease // J. Biol. Chem.–1987.–**262**, N 23.–P. 11364–11368.
12. *Panyutin I., Klishko V., Lyamichev V.* Kinetics of cruciform formation and stability of cruciform structure in superhelical DNA // J. Biomol. Struct. and Dyn.–1984.–**1**, N 4.–P. 1311–1324.
13. *Zarudnaya M. I., Potyahaylo A. L., Hovorun D. M.* Conservative structural motifs in the 3' untranslated region of SARS coronavirus // Biopolymers and Cell.–2003.–**19**, N 3.–P. 298–303.
14. *Saenger W.* Principles of nucleic acids structure.–New York etc.: Springer, 1984.–556 p.
15. *Odynets K. A., Kornelyuk A. I.* Molecular aspects of organization and expression of SARS-CoV coronavirus genome // Biopolymers and Cell.–2003.–**19**, N 5.–P. 414–431.
16. *Voineagu I., Narayanan V., Lobachev K., Mirkin S.* Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins // Proc. Nat. Acad. Sci. USA.–2008.–**105**, N 29.–P. 9936–9941.

UDC 577.2:577.32
Received 15.10.08