UDC 23775.24

I.S. Shcherbatenko

Zabolotny Institute of Microbiology and Virology, National Academy of Sciences of Ukraine;
154 Acad. Zabolotny St., Kyiv, MSP, D03680, Ukraine

GRAPHICAL VISUALIZATION OF THE BIOLOGICALLY SIGNIFICANT SEGMENTS IN THE SEQUENCE SETS OF THE RELATIVE PLANT VIRUSES

The author's and collaborators' computational investigations of the conserved biologically significant segments within viral nucleotide and amino acid sequences are considered in the article.

The results obtained suggest that the interactive graphical visualization of the short identical or similar sites in the sequence sets of relative viruses allows to reveal various specific elements such as right, inverted, tandem, opposite and regular repeats; deletion/insertion; GC/AT-rich sites; contexts of translation initiation and termination codons; transcription initiation signals; spontaneous nucleotide substitutions; codon usage bias etc.

To reveal and investigate different biologically significant sequences very short and simple computer programs, based on common sequence scanning algorithm, may be employed. Various graphic objects, which appeared during visualization of similar sites, may be computationally converted into corresponding nucleotide or amino acid sequences followed by writing within a text file. The change of some scanning parameters or slight modification of certain program modules allows to enlarge the program potentialities.

A set of little and simplified computer programs obtained by successive modifications of the initial program is a suitable tool for quick revealing and investigating various biologically significant sequence sites.

Key words: plant viruses, computational analysis of viral sequences, biologically significant sequence sites, graphical visualization of the conserved sequence segments.

It is known that genomic sequences encode extensive overlapping information – the code for protein production and numerous codes for regulatory sequences: promoters, enhancers, codon contexts, binding sites, species-specific combinations of nucleotides, biologically significant motifs, and others, which influence gene transcription, translation, posttranscriptional processes etc. [3]. To coexist simultaneously within the same sequence, all codes have to be degenerated [2]. A well-known example of such degeneration is the redundancy of the genetic code – differences in the number of synonymous codons between the canonical amino acids. An evidence of degeneration of regulatory codes may be a large variety of conservative nucleotide block features in bacterial promoters, including a distance between blocks and their localization relative to the transcription start codon as well as nucleotide identity in the same block of different bacteria or bacterial genes [12].

So far as different biologically significant sequence segments are degenerated to various extents and contain highly conserved functional motifs, there is a possibility for their revelation by graphical visualization as the clusters of the short identical or very similar sites in the sequence sets of relative viruses. The realizations of this possibility in our investigations are summarized here.

Virus sequences and software tools. Complete genomic sequences and coding sequence annotations of the (+)ssRNA plant viruses from 17 genera used in the work were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/). Model random sequences were generated using a random number generator. The CDS translations were used as amino acid sequences.

To reveal and investigate the conserved sequence segments and putative regulatory sequences in viral genomes a set of simplified computer programs (utilities), based on common sequence scanning algorithm, have been used. The algorithm includes: 1) putting of working parameters; 2) reading a short site (reading site) from the definite sequence (reading sequence) with the defined step (reading step); 3) searching the similar site (searching site) in the same or another sequence (searching sequence) with the definite searching depth (the distance between the reading and similar (searching) sites positions); 4) graphic visualization of the reading and searching sites if their similarity and/or contents in the sequence set are equal or more than that of the defined ones; 5) changing of working parameters.

According to the above-mentioned scanning algorithm there are five program modules: put, read, search, show and change. The put and change modules define the basic scanning parameters: virus

species and sequence regions, reading and searching sequences, reading step, site length and similarity, window width, graphic size and dimension, color of dots and lines, etc. Each of software tools allows an interactive changing of many scanning parameters to search different conservative sites, but some investigations require development of a new program, which may be easily destined by a slight modification of one or more program modules.

Depending on the used scanning parameters, various graphic objects may appear on the display. The marking of certain graphical objects with a graphic cursor results in the object conversion into corresponding nucleotide or amino acid sequence followed by its output to display and to the text file.

Graphic objects of viral sequences. Testing the scanning program Scangene on different nucleotide and amino acid sequences showed its fitness for quick revealing of various sequence features such as right, inverted, tandem and regular repeats, deletion/insertion, GC/AT-rich sites, etc. [15]. The graphical object showed in Fig. 1 (vertical lines and dots) was obtained using: the genomic sequence of the TMV-L strain as reading and searching sequences; 4-nucleotide reading site; 1-nucleotide reading step; 100% similarity of the reading and searching sites. The graphic size was squeezed to window width (600 pixels). The regular repeat caacaacaacaacaaacaacaacaaca producing the specific graphical object is responsible for the translation enhancement [4]. The sequence and genomic position of this repeat (from 20 to 46) were revealed by the program conversion followed by marking the object with the graphic cursor.

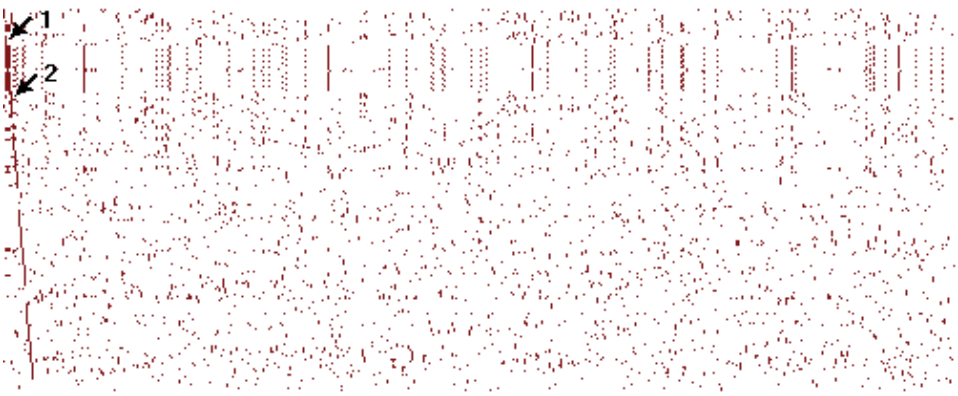

Fig. 1. Regular nucleotide repeats in the genomic sequence of the tobacco mosaic virus strain TMV-L (accession no. X02144). 1, 2 – genomic positions (from 20 to 46) of the caacaacaacaacaaacaacaacaaca site containing short regular repeats. Four-nucleotide parts of this site: caac, aaca, acaa, caaa and aaac read with one-nucleotide step form vertical lines and point clusters in a consequence of site positions plotting

Another graphical object forms a long nucleotide repeat in the genomic sequence of the potato virus X strain PVX-X1 (Fig. 2). The repeat is discovered in the graphic window by scanning the sequence with 15-nucleotide reading site and the same reading step. The conversion of the graphical object into the corresponding nucleotide sequence reveals the 1133-nucleotide region located in 5303 to 6435 genomic positions and repeated on the 3' end of the viral RNA in genomic positions from 6436 to 7568. The repeat contains 15 nucleotide substitutions and includes the 3' end part of the TGB2 gene as well as TGB3 and CP genes and 3' untranslation region (3'UTR) of PVX-X1 strain (Fig. 3). The PVX-X1 genome, in fact, is the PVX-S1 genome with addition of the 1133-nucleotide region, which differs from such region of PVX-S1 genome by 15 nucleotides. It remains so far unclear whether the long nucleotide repeat in the genomic sequence of the potato virus X strain PVX-X is caused by sequencing errors or not.

The short opposite repeats in movement proteins TGBp1 and TGBp3 of the potato virus X strain PVX-CP4 are shown in Fig. 4. Amino acid sites AGA, NAI, IDSE and LSLE located in 3, 8, 48 and 62 positions of the TGBp3 protein are also located in TGBp1 protein but in the inverse order: in 224, 211, 163 and 107 positions. Similar scattered localization is observed in GAU, SIT and LSF sites in TGBp3 and TGB2 proteins as well as in AGA, AAA, LDA and TRG sites in CP and TGBp1 proteins. How these opposite repeats arise and what are their role in virus protein functions is still unclear.

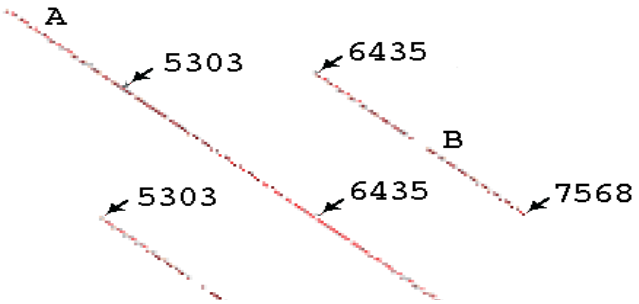

Fig. 2. The long nucleotide repeat in the genomic sequence of the potato virus X strain PVX-X (accession no. M72416). A – genomic position dots of 15-nucleotide sites read with 15-nucleotide step. B - position dots of nucleotide sites, witch are similar to read ones. 5303-6435 – the 1133-nucleotide region, which repeats on the 3'-end of the viral RNA in genomic positions from 6436 to 7568

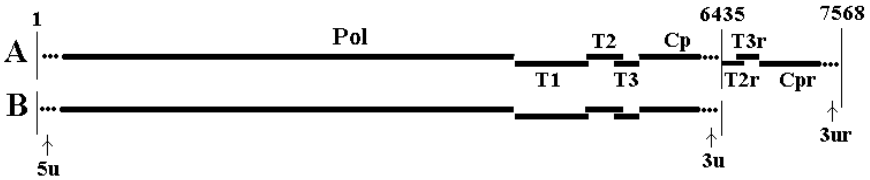

Fig. 3. The genome structure of two PVX strains. A – PVX-X1 (accession no. M72416), B – PVX-S1 (accession no. X05198). 1...6435 – genome length of PVX-S1 and identical sequence of both strains. 1...7568 – genome length of PVX-X1. Pol, T1, T2, T3, Cp – PVX genes: polymerase, TGB1, TGB2, TGB3 and coat protein, accordingly. 5u – 5'UTR (untranslation region), 3u – 3'UTR. T2r, T3r, Cpr, 3ur – repeats of the TGBp2, TGBp3, Cp and 3'UTR on the 3'-end of the PVX-X1 genome

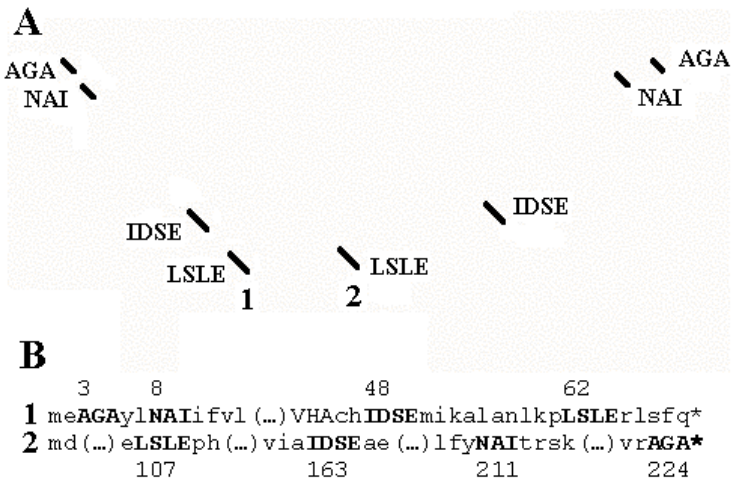

Fig. 4. Scattered localization of short identical amino acid sites in movement proteins TGBp1 and TGBp3 of the potato virus X strain PVX-CP4 (accession no. AF172259). A – scattered sites visualization. B – sites positions (3...224). 1 – TGBp3 protein. 2 – TGBp1 protein. * – stop codons. (...) – cutting sequence fragments

The searching for the similar nucleotide sites in viral genomes showed some similarity between certain viruses from different genera and families. Some of them can have clearly expressed clusters of relatively long nucleotide sites, numerous short sites scattered along the viral genome, diffuse clusters of short sites located in several genomic positions as well as little clusters or single nucleotide sites located in a certain genome region [1].

Thus, short and simple computer programs, based on common sequence scanning algorithm, can be used to reveal various specific elements in nucleotide and amino acid sequences. To enlarge the program potentialities some modifications of the scanning algorithm were used: two-symbolic nucleotide alphabet instead of four-symbolic one; visualization of position-linked clusters of short nucleotide sites instead of the searching sites; the input of searching sites instead of reading them from a sequence, etc. Various program modifications allowed investigating some transcription and translation signals, spontaneous nucleotide substitutions and codon usage bias in viral genomes.

Transcription and translation signals. The contexts of translation initiation codon (a nucleotide sequences surrounding the AUG) have been analyzed by computation in 15 tobamoviruses (45 genes) and 22 potexviruses (110 genes). The results obtained indicate both a few key similarities and some differences between Kozak's eukaryotic [8] and viral AUG contexts. A distinctive feature of viral translation initiation contexts is a high variation of context elements between various viruses and/or genes [6].

Similar variations were found in contexts of the leaky stop codons in the replicase and coat protein genes of 118 plant virus strains, belonging to 68 species of 13 genera [7]. Though most of the leaky terminal codons are followed by caauua, cgguuu, gggugc, ggaggc or guagac hexanucleotides, there are vast varieties of other identical nucleotides located at different distances from the terminal codon. The results obtained support a hypothesis that the efficiency of translation termination as well as stop codon read-through may be determined at genome or gene level rather than by a short nucleotide context surrounding the stop codon.

Using two-symbolic alphabet (R – purines: A or G and Y – pyrimidines: C or T), the sequences responsible for the transcription termination: a pyrimidine triplets surrounded with purine hexaplets [12] were found at the 3' end of the TMV genomic RNA [15].

Transcription initiation signals in the subgenomic promoters of fifteen tobamoviruses were searched by visualization of position-linked clusters of short nucleotide sites [10]. It was discovered that 22 of 30 promoters tested contain the five-nucleotide site tcgtt in two sequence regions upstream of the translation start codon (Fig. 5). The seven-nucleotide site gattcgt, which contains the tcgtt was found in eight promoters. After aligning the sequences of the movement and coat protein subgenomic RNAs by the tcgtt positions, the nine-nucleotide motive gg/attcgttt, flanged with length-variable gc- and at-containing sequences were discovered in most tested tobamovirus subgenomic promoters [5].

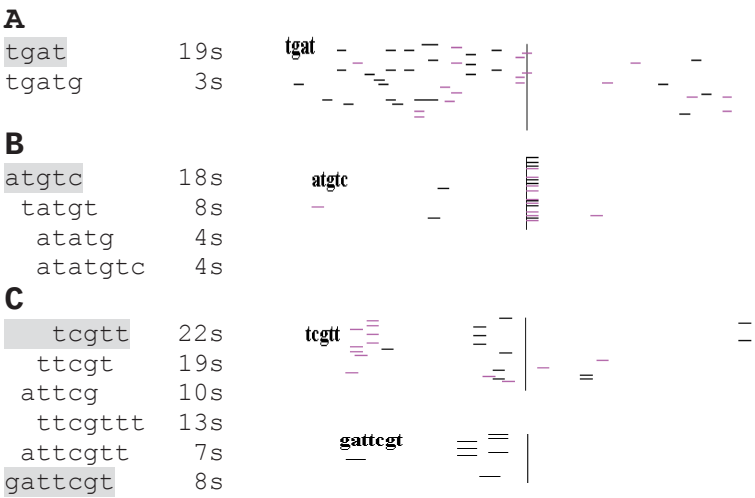

Fig. 5. Localization of position-linked short nucleotide sites in 30 subgenomic promoter regions of tobamoviral genes. Tgat...gattcgt - nucleotide sites. 19s...8s – the number of promoter regions containing marked sites. Vertical lines – the ATG position (start codon of translation). Horizontal lines – positions of nucleotide sites relative to the coat protein gene ATG (dark lines) or the movement protein gene ATG (light lines). A – scattered localization of the tgat site in 19 promoter regions. B – presence of the atgtc site mainly at the beginning of genes. C – localization of the tcgtt and gattcgt sites in two sequence regions upstream of the translation start codon

The gg/attcgttt motive is similar to the ICR2 consensus (ggttcgantcc) of pol III promoters in tRNA genes [13] and to the stemloop gggattcgaattccc, found in the domain D1 of the TMV-L strain. This domain is located on the 3' end of virus genome, and is important for promotion of the minus-strand RNA synthesis [14]. The gg/attcgttt and gggattcgaattccc are maintained in the promoters of many TMV strains and other tobamoviruses (Fig. 6). The analogous sites with some nucleotide substitutions occur in most tobamovirus promoters. However, no gggattcgaattccc-similar sequences were found either on the 3' end of the minus-strand or on the 5' end of the plus-strand genomic RNAs.

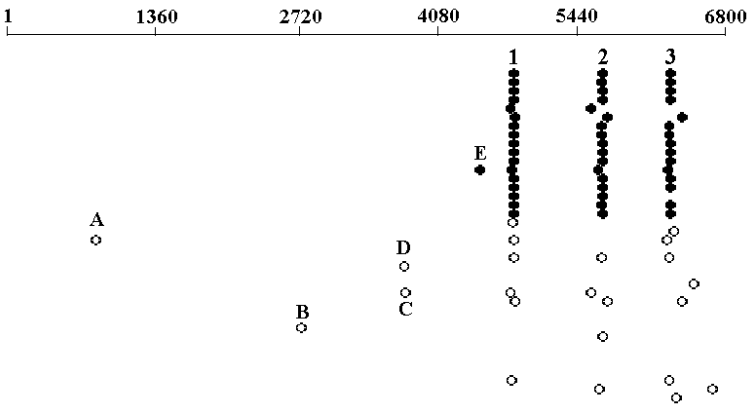

Fig. 6. Conservative sites of tobamoviral promoters. 1 – the GGTTCGTTT site of the movement protein subgenomic RNA promoter (4799-4836 positions). 2 – GATTCGTTT and GATTCGTAT sites of the coat protein subgenomic RNA promoter (5579-5737 positions). 3 – GGGATTCGAATTCCC and GGGgTTCGAgTTCCC sites of the minus strand RNA genomic promoter (6347-6582 positions). A - E – additional positions of conservative sites: GATTCGTAT – 781 (A), 3769 (D), 3782 (C); GATTCGTTT – 2774 (B); GGGATgCGAgTTCCC – 4905 (E). Light circles – tobamoviruses, black circles – TMV strains

The results obtained suggest similar transcription initiation signals in the tobamovirus promoters of the subgenomic and minus-strand genomic RNA, but propose quite another promotion of the plus-strand RNA synthesis. These differences may appear due to replication of various viral RNAs by different RdRp complexes, containing different viral and host components.

Spontaneous nucleotide substitutions and codon usage bias. A comparative computational analysis of codon usage bias and spontaneous nucleotide substitutions was performed in genomic sequences of the soybean dwarf virus and potato virus X. It was shown that frequency of nucleotide substitutions depends on nucleotide and strain pairs, length and localization of gene regions as well as on the substitution types and their codon positions [11]. The most relation of substitution frequency is shown by different nucleotide pairs (u→c/c→g, 1240), codon positions (third/second, 17.3) and substitution types (transitions/transversions, 5.8), the least – forward and back substitutions in the same nucleotide pairs (1.03-1.6).

The synonymous codon usage in virus genes varies widely depending on the gene, amino acid and codon, gene overlapping, two first codon's nucleotides, mononucleotide context, located upstream and downstream of the codon, GC-content in virus-encoded genes and in the third codon positions [9]. The dependence of the codon usage bias on amino acid, codon and virus gene is shown in Fig. 7. It can be seen that the replicase gene of the tobacco mosaic virus strain TMV-V has 16 amino acids, containing only one preferable synonymous codon. The coat and 126k proteins have 14 such amino acids, and the movement protein gene – 13. Six amino acids give the greatest contribution to overall codon bias having only one preferable codon in each of four viral genes: Arg – aga, Val – gtt, Gly – gga, Tyr – tac, Cys – tgt and Asn – aat.

The least contribution to codon bias in the coat protein gene is made by valine and alanine having the same frequency of three synonymous codons : Val – 4 gtt, 4 gta, 4 gtg and 2 gtc; Ala – 4 gcc, 4 gca, 4 gcg and 2 gct. More detailed obvious analysis of the codon bias may be performed by visualization of each of viral genes separately with all junctions of synonymous codons.

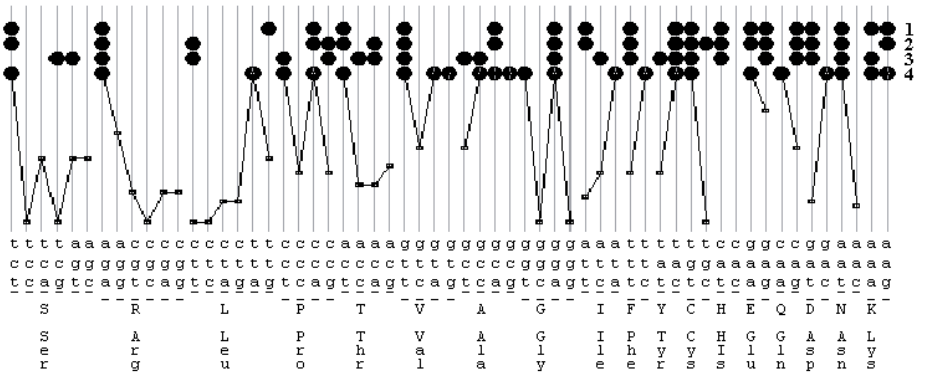

Fig. 7. Codon usage bias in four genes of the tobacco mosaic virus strain TMV-V (accession no. J02415). Small rectangles – synonymous codons, dark circles – preferable synonymous codons, broke lines – a junction of codons. 1...4 - preferable codons of the virus genes: 1 – replicase (Re, 183k protein), 2 – Mt-He (126k protein), 3 – movement protein, 4 – coat protein gene. S Ser...K Lys – amino acids, tct...aag – codons

Concluding remarks. It is clear that, searching, identifying and investigating the coding, the control and regulatory signals for gene expression are very important for understanding the origin, evolution and existence of life. The wide-spread approach to these investigations is a comparative computational analysis of simple viral sequences. For searching and analyzing the biologically significant conserved segments within nucleotide and amino acid sequences a lot of computation methods and programs have been developed over the past several decades. Some of the software tools are designed for a wide multifunctional analysis, others – for the narrow specific tasks.

Because of continuous growing of sequence databases, it has become increasingly important to develop a simple-to-use computer programs for comparative sequence analysis by interactive graphical visualization and visual investigation of various conservative sequence elements. So far as the biologically significant sequences contain highly conserved functional motifs, some of them may be revealed by graphical visualization as the clusters of the short identical or similar sites in the sequence sets of relative viruses. The possibility of such investigation using a set of simplified computer programs based on comon sequence scanning algorithm was shown in our experiments.

It was demonstrated that the interactive changing of some program parameters or writing of a new program by a slight modification of the initial one allow to reveal various specific sequence elements such as right, inverted, tandem, opposite and regular repeats; deletion/insertion; GC/AT-rich sites; contexts of translation initiation and termination codons; transcription initiation signals in the subgenomic and the minus-strand RNA promoters; spontaneous nucleotide substitutions and codon usage bias in viral genomes; some similarity between certain viruses from different genera and families, etc.

Thus, the interactive graphical visualization of the short identical or similar sites in the sequence sets of relative viruses can be used as a simple, fast and obvious way for searching, identifying and analyzing the coding, control and regulatory signals for gene expression.

И.С. Щербатенко

Институт микробиологии и вирусологии им. Д.К. Заболотного НАН Украины;
ул. Академика Заболотного, 154, г. Киев, ГСП, Д03680, Украина

ГРАФИЧЕСКАЯ ВИЗУАЛИЗАЦИЯ БИОЛОГИЧЕСКИ ВАЖНЫХ СЕГМЕНТОВ В НАБОРАХ СИКВЕНСОВ РОДСТВЕННЫХ ВИРУСОВ РАСТЕНИЙ

Р е з ю м е

В статье рассматриваются исследования биологически важных консервативных сегментов в нуклеотидных и аминокислотных сиквенсах вирусов растений, проведенные автором с сотрудниками методом компьютерного анализа.

Полученные результаты показывают, что интерактивная графическая визуализация коротких идентичных или подобных сайтов в наборах сиквенсов родственных вирусов дает возможность выявить разнообразные специфические элементы такие как прямые, инвертированные, тандемные, противоположные и регулярные повторы; делеции/вставки; GC/AT-богатые участки; контексты кодонов, участвующих в инициации и терминации трансляции; сигналы инициации транскрипции; спонтанные замены нуклеотидов; избирательность использования кодонов и пр.

Для выявления и исследования различных биологически важных участков сиквенсов могут использоваться очень короткие и простые компьютерные программы, основанные на общем алгоритме сканирования. Разнообразные графические объекты, появляющиеся в процессе визуализации подобных сайтов, могут быть программно конвертированы в соответствующие нуклеотидные или аминокислотные сиквенсы и записаны в текстовые файлы. Изменения некоторых параметров сканирования или незначительная модификация определенных программных модулей расширяют возможности компьютерных программ.

Наборы маленьких и простых компьютерных программ, полученные путем последовательных модификаций некой исходной программы, представляют собой удобный инструмент для быстрого выявления и исследования различных биологически важных сегментов сиквенсов.

Ключевые слова: вирусы растений, компьютерный анализ вирусных сиквенсов, биологически важные участки сиквенсов, графическая визуализация консервативных сегментов сиквенсов.

І.С. Щербатенко

Інститут мікробіології і вірусології ім. Д.К. Заболотного НАН України;
вул. Академіка Заболотного, 154, м. Київ МСП, Д03680, Україна

ГРАФІЧНА ВІЗУАЛІЗАЦІЯ БІОЛОГІЧНО ВАЖЛИВИХ СЕГМЕНТІВ У НАБОРАХ СИКВЕНСІВ СПОРІДНЕНИХ ВІРУСІВ РОСЛИН

Р е з ю м е

В статті розглядаються дослідження біологічно важливих консервативних сегментів у нуклеотидних та амінокислотних сиквенсах вірусів рослин, проведені автором з співробітниками методом комп'ютерного аналізу.

Отримані результати показують, що інтерактивна графічна візуалізація коротких ідентичних або подібних сайтів у наборах сиквенсів споріднених вірусів дає можливість виявити різнорідні специфічні елементи такі як прямі, інвертовані, тандемні, протилежні та регулярні повтори; делеції/вставки; GC/AT-багаті ділянки; контексти кодонів, причетних до ініціації та термінації трансляції; сигнали ініціації транскрипції; спонтанні заміни нуклеотидів; вибірковість використання кодонів тощо.

Для виявлення і дослідження різних біологічно важливих ділянок сиквенсів можуть використовуватися дуже короткі і прості комп'ютерні програми, засновані на загальному алгоритмі сканування. Різноманітні графічні об'єкти, що з'являються в процесі візуалізації подібних сайтів, можуть бути програмно конвертовані у відповідні нуклеотидні чи амінокислотні сиквенси і записані у текстові файли. Зміни деяких параметрів сканування або незначна модифікація певних програмних модулів суттєво розширюють можливості комп'ютерних програм.

Набори маленьких і простих комп'ютерних програм, отримані шляхом послідовних модифікацій якоїсь вихідної програми, являють собою зручний інструмент для швидкого виявлення і дослідження різних біологічно важливих сегментів сиквенсів.

Ключові слова: віруси рослин, комп'ютерний аналіз вірусних сиквенсів, біологічно важливі ділянки сиквенсів, графічна візуалізація консервативних сегментів сиквенсів

1. *Gordejchyk O.I., Shcherbatenko I.S.* Searching for similar nucleotide sites in genomic sequences of phytopathogenic viruses // Microbiol. Zh. – 2009. – **71**, N 4. – P. 63–70.
2. *Cohanim A.B., Haran T.E.* The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes // Nucleic Acids Res. – 2009. – **37**, N 19. – P. 6466–6476.
3. *Forman J.J., Coller H.A.* The code within the code: MicroRNAs target coding regions. // Cell Cycle. – 2010. – **9**, N 8. – P. 15–33, 18–41.
4. *Gallie D.R.* The 5'-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F // Nucleic Acids Res. – 2002. – **30**, N 15. – P. 3401–3411.

5. *Gordejchyk O.I., Oleshchenko L. T., Shcherbatenko I.S.* Similar nucleotide blocks in tobamoviral subgenomic promoters // Microbiol. Zh. – 2007. – N 1. – P. 42–51.
6. Gordejchyk O.I., Shcherbatenko I.S. Context sequences of translation initiation codon in tobamo- and potexvirus genes // Microbiol. Zh. – 2010. –**72**, N 1. – P. 39–46.
7. *Gordejchyk O.I., Shcherbatenko I.S.* Contexts of the suppressive termination codons of translation in genes of positive-sense ssRNA plant viruses // Microbiol. Zh. – 2010. – 72, N 6. – P. 58–65.
8. *Kozak M*. Structural features in eukaryotic mRNAs that modulate the initiation of translation. // J. Biol. Chem. – 1991. – **266**, N 30. – P. 19867–19870.
9. *Kyrychenko A.N., Gordejchyk O.I., Shcherbatenko I.S.* Codon bias and nucleotid substitutions in soybean dwarf virus // Microbiol. Zh. – 2012. – **74**, N 3. – P. 90–97.
10. *Kyrychenko A.N., Shcherbatenko I.S.* Concervative nucleotide sites in tobamoviral subgenomic RNA promoters // Microbiol. Zh. – 2006. – **68**, N 3. – P. 63–71.
11. *Kyrychenko A.N., Shcherbatenko I.S.* Spontaneous nucleotide substitutions in potato virus X genes // Microbiol. Zh. – 2000 (in press).
12. *Lewin B.* Genes. – New York: John Wiley and Sons, 1987. – 544 p.
13. *Marsh L.E., Pogue G.P., Hall T.C*. Similarities among plant virus (+) and (-) RNA termini imply a common ancestry with promoters of eukaryotic tRNAs // Virology. – 1989. – **172**, N 2. – P. 415–427.
14. *Osman T.A., Hemenway C.L., Buck K.W*. Role of the 3' tRNA-like structure in tobacco mosaic virus minus-strand RNA synthesis by the viral RNA-dependent RNA Polymerase *in vitro*. //. J. Virology. – 2000. – **74,** N 24. – P. 11671–1168.
15. *Shcherbatenko I.S.* Computer analysis of viral genomes structure organization by sequenses scanning // Microbiol. Zh. – 2003. – **65**, N 1-2. – P. 217–228.

Отримано 04.09.2011