Volume 14 Number 1 1986 Nucleic Acids Research

Relation between genomic and capsid structures in RNA viruses

K.Yamamoto* and H.Yoshikura

Department of Bacteriology, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Received 6 June 1985

ABSTRACT

We described a new computer program for calculation of RNA secondary structure. Calculation of 20 viral RNAs with this program showed that genomes of the icosahedral capsid viruses had higher folding probabilities than those of the helical capsid viruses. As this explains virus assembly quite well, the information of capsid structure must be imprinted not only in the capsid protein structures but also in the base sequence of the whole genome.

We compared folding probability of the original sequence with that of the random sequence in which base composition was the same as the original. All the actual genomes of RNA viruses were more folded than the corresponding random sequences, even though most transcripts of chromosomal genes tended to be less folded. The data can be related to encapsidation of viral genomes. It was thus suggested that there exists a relation between actual sequences and random sequences with the same base ratios, and that the base ratio itself has some evolutional meaning.

INTRODUCTION

Architecture of viruses with RNA genomes are grouped into two types based on the arrangement of morphologic subunits: those with helical symmetry and those with icosahedral symmetry. In helical capsids, the RNA is located in a helical groove between the capsids, while, in icosahedral capsids , the RNA is considered to be tightly packed in a central core. The capsid structures must primarily be determined by the capsid protein structures, but the viral genome itself must have to take such a secondary structure as is appropriate for the encapsidation. Therefore, the information for capsid structure must be imprinted not only in the genes coding for capsids but also in the base sequence of the entire genome which determines its folding structure. In order to test this hypothesis, we calculated the folding probabilities of 20 viral RNAs with our program which is outlined below. We

[1]Source program will be provided upon receipt of a self-addressed mailing label and a blank tape. A small charge will be requested to cover mailing and processing.

also compared folding probabilities of existing sequences with those of random sequences with the same base ratios. An interesting relation existed.

METHOD

Computer programs to find the most probable hydrogen-bonding pattern of a given nucleotide sequence are now available (1-6). They are using the algorithm which was originally developed by Nussinov et al. (5) and is probably the best one for this purpose. However, they cannot be applied to long sequences such as viral entire genomes which often exceed several kilo bases; in addition, it takes long calculation time, and consequently it is expensive. Another problem with these programs is that, even if more than two biologically significant structures are possible with comparable but slightly different stabilities (7), the program detects only the most stable one and neglects the others. To circumvent these problems, we developed a computer program which calculated RNA folding structures probabilistically (8). Recently, we modified the program, with some loss of precision so that it can be applied to sequences of any length we want to analyze (9). Using this program, we devised a map in which the sum total of information values (defined in Ref. 8 a parameter of folding probability) in every 50 bases was plotted along the entire stretch of the sequence. The map is called "information georgaphy" (IG) (10). In IG of tobacco mosaic virus genome (Fig. 2), a high folding probability region is located at the 3' end of the genome, which exactly corresponds to the encapsidation signal (11). Therfore, IG is valid. Information mass volume (IMV) is the sum total of information values of the whole genome normalized for 200 bases.

The random sequences were obtained by the random number generating program in VAX11/780 RUNTIME LIBRARY. All the calculations (batch job) were terminated within 10 hours.

The outline of the system is shown in Fig. 1. The algolism described has been programed in Fortran. It is implemented on the VAX11/780 operating system. All the figures were drawn by using a NWX-235 colour graphic display terminal with a laser printer hard copy unit LBP-10. These units were supported by the graphic software package GRAPAC II.

RESULTS

[1] Application of the program to the analysis of viral genomes: Fig. 2 shows the information geographies (IG) of 20 viral RNA sequences. The IGs from 1 to 7 are those of viruses with icosahedral capsids (murine leukemia virus, poliovirus, foot and mouth disease virus, semliki forest virus, bacteriophage MS2, sindbis virus, turnip yellow mosaic virus), and IGs from 8 to 13 are those of viruses with helical capsids (two regions of vesicular stomatitis virus, 7 different segments of influenza viruses, tobacco mosaic virus, rabies virus, snowshoe hare bunyavirus, corona virus). RNAs of the former group tended to be more folded than the latter. IMVs of the former group exceeded 35 and occasionally 100, while IMVs of the latter group were less than 30 (Fig.4-B). This is in complete agreement with prior observations on virus assembly. Namely, for a long thread of RNA to be

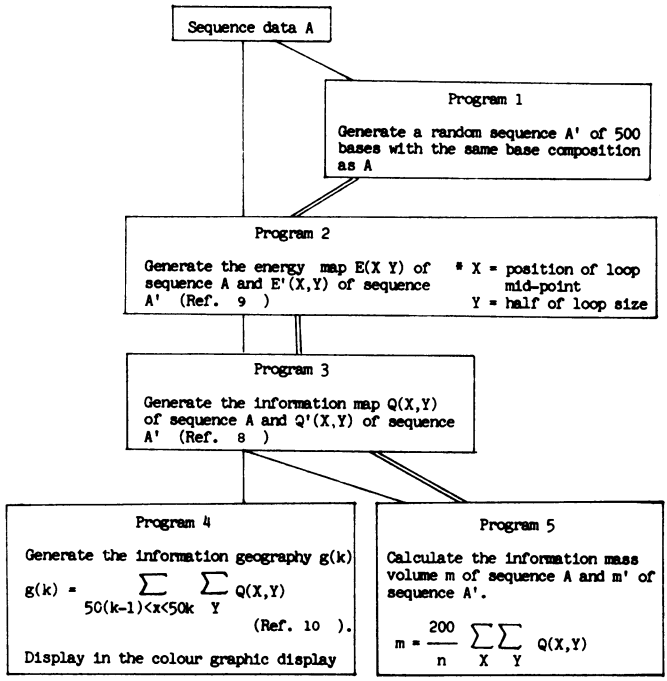

Fig. 1. Outline of the system.

enclosed in a small capsid with icosahedral symmetry, the thread itself has to be crumpled into a small ball, while, for the long molecule to be coiled helically along the inner surface of a long tube (helical capsids), the thread itself should not be folded, but rather, extended so as to facilitate the coiling. These data indicate that the RNA base sequence of the whole viral genome itself contains the information for capsid structure.

[2] Order vs chaos Genetic sequences now present are considered to have arisen from random sequences after a long period of mutation and selection processes. The random sequences had the follwing properties. (1) IMVs of random sequences with the same base ratio obtained by independent shufflings converged to a fixed value as the sequence became longer (data not shown). (2) IMVs of various base ratios were calculated with only one shuffled sequence. Here, the base ratios giving rise to high IMVs are localized in one closed area (Fig.3-A for G=25% and variable A, T and C; Fig.3-B for A=T and variable G and C), and if we could calculate IMVs for all the possible base ratios, the region would occupy a closed space shaped like a pear cut into quaters in a 3-dimensional space determined by A, T and G axes

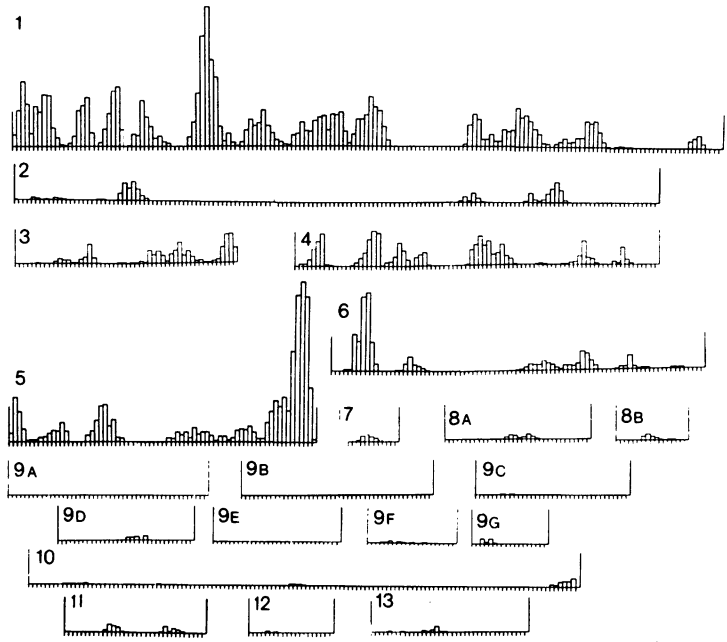

Fig. 2. Information geographies (IG) of RNA viruses. Ordinate: information value, g(k). Abscissa: base sequence (graduated every 50 bases) The 5' end is located on the left side of each IG. 1:Moloney murine leukemia virus entire genome 8332 bp. 2:Poliovirus type 1 (Mahoney strain) entire genome 7433 bp. 3:Foot and mouse disease virus capsid polypeptides 2802 bp. 4:Semliki forest disease virus membrane capsid and glycoproteins mRNA 4214 bp. 5:bacteriophage MS2 complete genome 3569 bp. 6:Sindbis virus (hr strain) 26S mRNA for structural protein 4350 bp. 7:turnip yellow mosaic virus coat protein 695 bp. 8A:Vesicular stomatitis virus glycoprotein (M) mRNA 1665 bp. 8B:Vesicular stomatitis virus matrix protein (M) mRNA 831 bp. 9A:Influenza A/PR/8/34 (H1N1) polymerase 3 (Seg. 1) RNA 2341 bp. 9B:Influenza A/PR/8/34 (H1N1) polymerase 2 (Seg. 3) RNA 2233 bp. 9C:Influenza A/PR/8/34 (H1N1) hemagglutinin (Seg. 4) RNA 1178 bp. 9D:Influenza A/NT/60/68 (H3N2) nucleoprotein (Seg. 5) 1565 bp. 9E:Influenza A/NT/60/68 (H3N2) neuraminidase (Seg. 6) RNA 1467 bp. 9F:Influenza A/PR/8/34 (H1N1) matrix protein (Seg. 7) RNA 1027 bp. 9G:Influenza S/UDORN/72 (H3N2) nonstructural protein (Seg. 8) 890 bp. 10:Tobacco mosaic virus (strain vulgare) complete genome 6395 bp. In the long stretch of tobacco mosaic virus genome, only the 3' end which contains initiation region for the assembly of the virus particle (Butler et al, 1977) had high folding probabilities. 11:Rabies virus ERA strain glycoprotein 1650 bp. 12:Snowshoe hare bunyavirus small viral RNA 982 bp. 13:Mouse hepatitis virus strain A59, nucleocapsid protein gene 1840 bp.

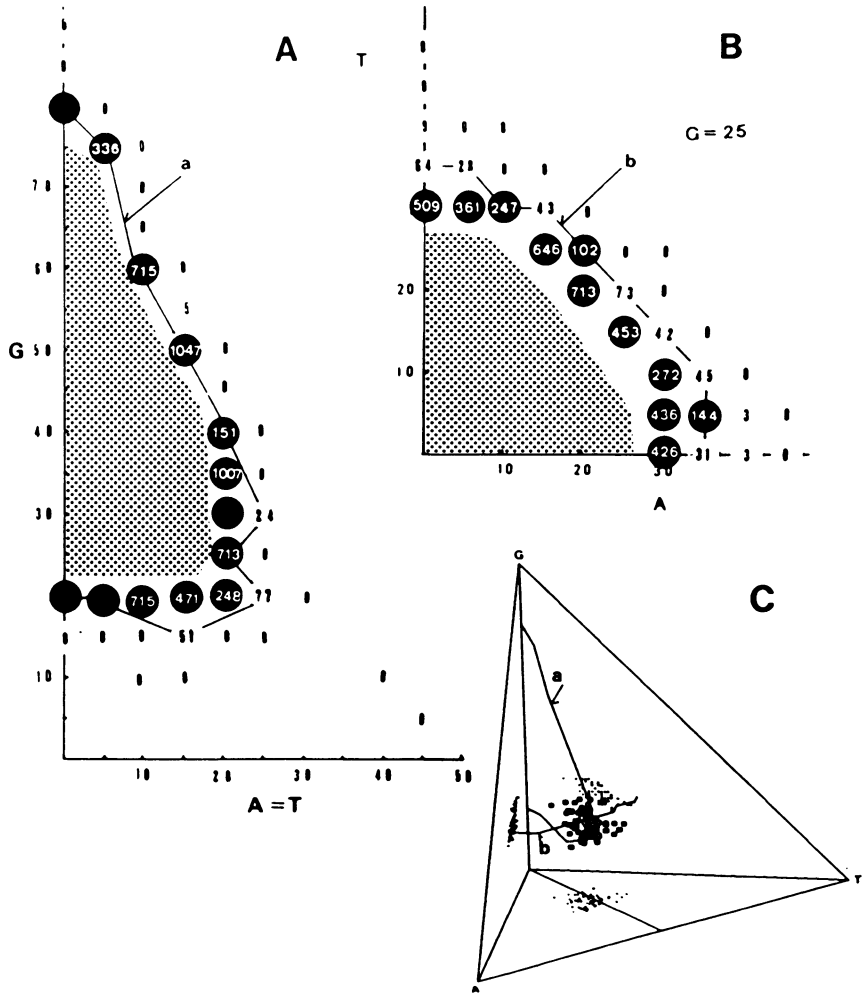

Fig. 3. IMVs of random sequences (500bp. long). The IMV values are plotted in the positions determined by the base ratio. Fig.A: G=25% and variable C,T,and A. Fig.B: A=T and variable G and C. The inner shaded part of the close area had IMVs too high to be calculated within 10 hr. Fig.C: data in Figs.A and B are plotted in the 3-dimensional space AT, TG, GA. The base ratios of existing sequences analyzed in this paper are plotted by squares and their projections to AT, TG and GA planes are indicated by dotts in Fig.C.

(Fig.3-C). Both these data indicate that the IMV of a random sequence was determined by the base ratio alone, almost independently of shuffling (however this applies only if the sequence is not too short). In Fig. 3-C,

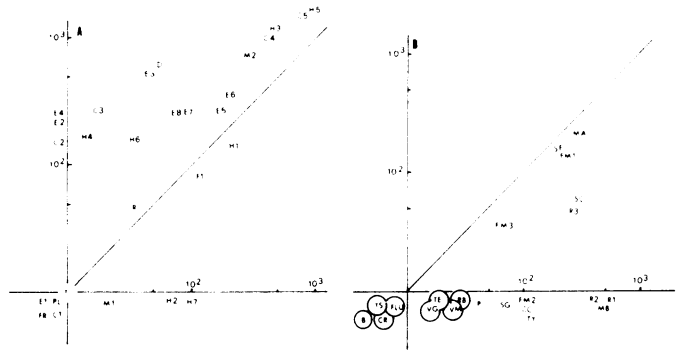

Fig. 4. IMVs of actual sequences are plotted against those of random sequences (500bp. long) with the same base ratio. Ordinate: IMV of random sequences. Abscissa: IMV of actual sequences.

A: Chromosomal genes of prokaryotes and eukaryotes. H1:human enkephalin precursor mRNA 1351 bp. H2:human hla-dr antigen alpha-chain mRNA ivs fragments 1199 bp. H3:human hla-dr antigen-like beta chain mRNA 1034 bp. H4:human hla-dr antigen heavy chain gene encoding aa3-229 1806 bp. H5:human humos gene homologus to transforming gene of MSV. 1303 bp. H6:human ig Kappa 1-chain constant region gene (inv 3 allele) 1209 bp. H7:human leukocyte interferon a precursor mRNA 961 bp. M1:mouse pancreatic alpha-amylase mRNA 1557 bp. M2:mouse transplantation antigen h-2Kb mRNA, partial, 1235 bp. R:rabbit beta 1-globin with 2 ivs, type1 allele, 1827 bp. C1:chiken mRNA for ovalbumin 1873 bp. C2:chiken myosin alKali 1-chain a1 mRNA 3'end 848 bp. C3:messenger fragment for chiken pro-alpha-2(I) collagen. C4:chiken histon h2b gene: complete sequence flanks 842 bp. C5:chiken brain tubulin beta chain mRNA 1652 bp. FR:frog (rana temporaria) gamma-i-crystallin mRNA 464 bp. F1:angler fish preproinsulin mRNA 655 bp. D:D. melanogaster heat shock protein hsp 70 gene 1 at 87 c1. 2832 bp. PL:soybean actin gene 1620 bp. E1:E. coli lac operon: lac2 (3'end), lacy (complete), laca (5'end) 1500 bp. E5 E. coli Rec-A gene and flanks 1390 bp. E6: E. coli threonine operon thr-a gene 2463 bp. E7 E8 E2,E3,E4:E. coli tryptophan operon: entire DNA sequence 7335 bp. (E7:trp E, E8 trp D, E2:trp C, E3:trp B, E4:trp A). T:T. brucei variant surface glycoprotein mRNA 1626 bp.

B:RNA virus genomes. Viruses with the capsids which had helical symmetry are circled.
TY: turnip yellow mosaic virus coat protain. P:Polio virus type 1 (Mahoney). R1:Moloney murine leukemia virus 0.2-3.2kb. R2:Moloney murine leukemia virus 2.6-5.4kb. R3 Moloney murine leukemia virus 5.2-8.2kb. MA:bacteriophage MS2 gene A. MB:bacteriophage MS2 gene B. VG:vesicular stomahtis virus glycoprotain. VM vesicular stomahtis virus matrix protain. SC:sindbis virus capsid protain. SG:sindbis virus glyco protain. SF:semliki forest virus. FLV:various segments of influenza viruses. FM1:foot and mouth disease virus VP1. FM2:foot and mouth disease virus VP2. FM3:foot and mouth disease virus VP3. RB:rabies virus ERA strain glycoprotain. TS:TMV 0-2kb. TE:TMV 44-64kb. CC:cumcumbar mosaic virus RNA segment 4. CR:mouse hepatitis virus strain A59, nucleocapsid protain gene 1840 bp.

it is observed that base ratios of existing sequences are not distributed randomly in the A-T-G space, but clustered in the bulged out part of the pear-shaped region

Now, we compared the IMVs of the existing sequences with those of random sequences whose base ratios were the same as the existing ones. As shown in Fig.4-B, in RNA viruses, the IMVs of the actual sequences were higher than those of the random sequences, though in genes of prokaryotic and eukaryotic organisms, most of the actual sequences tended to have rather lower IMVs than the corresponding random sequences (Fig.4-A). The result is highly significant, because it suggests that the RNA genomes of icosahedral viruses are more folded than expected from the corresponding random sequences, even although the transcripts of most chromosomal genes are less folded; also the data can be readily related to encapsidation of the RNA virus genomes.

The data suggests that a certain relation exists between the actual sequences and random sequences with the same base ratios, and that the base ratio itself is meaningful.

ACKNOWLEDGEMENT

The sequence data used in the analysis were derived from EMBL Nucleotide Sequence Data Library and GENBANK Nucleotide Sequence Data Library. As the sequences used in our analysis were too numerous to be cited in the references, only the codes of the sequences in these libraries are indicated, for example, tabacco mosaic virus [TMVV/GENBANK: Guilley et al. Nucl. Acids Res. 6, 1287 (1979)], chiken mRNA for albumin [GGALB2/EMBL] etc. (references are not shown for chromosomal genes). Otherwise the GENBANK and EMBL Nucleotide sequence data libraries used are release 16.0 and release 3.0, respectively. The contributions of the researchers who determined the base sequences are acknowledged. The work was supported in part by grants from the Japanese Ministry of Education and Welfare. We thank Mr. S. Ishii and Mr. Y Sakurada of the Data Processing Center in the University of Tokyo for their advice.

*To whom correspondence should be addressed

REFERENCES

1. Comay, E., Nussinov, R. and Comay,O (1984). Nucleic Acids Res. 12 53-66
2. Hogeweg, P. and Hesper,B (1984) Nucleic Acids Res. 12 67-74
3. Jacobson, A.B. Good, L. Simonetti, J., and Zuker M (1984) Nucleic Acids Res. 12, 45-52

4. Papanicolaou, C. Gouy M. and Ninio, J. (1984) Nucleic Acids Res. 12, 31-44
5. Nussinov, R. Piecezenik, G., Griggs, J.R. and Kleitman, D J. (1978) SIAM J. Appl. Math, 35, 68-82
6. Zucker, M. and Stiegler P. (1980) Nucleic acids Res 9, 133-148.
7. Thom, R. (1975) "Structural stability and morphogenesis". Chapter 7, 124-150
8. Yamamoto K. Kitamura, Y. and Yoshikura, H. (1984) Nucleic Acids Res. 12, 335-346
9. Yamamoto, K , and Yoshikura, H. (1985) CABIOS in press.
10. Yamamoto, K. and Yoshikura, H. (1984) Jap. J. Exp. Med. 54, 241-247.