MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **56** (2006) 249-259

ISSN 0340 - 6253

Similarity analysis of RNA secondary structures based on 4D representation

Jianwei Gao*, Xingping Zhang
School of Business Administration North China Electric Power University, Zhuxin Road, Beijing, 102206,China

(Received March 20, 2006)

Abstract. In this paper, we propose a 4-D representation of RNA secondary structure. Based on this representation, we outline an approach to compute the similarities of RNA secondary structures by constructing a 3-component vector whose components are the leading eigenvalues of the L/L matrices. The examinations of similarities/dissimilarities among the secondary structures at the 3′-terminus of different viruses illustrate the utility of the approach.

1 Introduction

Mathematical analysis of large volume genomic biological sequence data is one of the challenges for bio-scientists. In recent years several authors outlined different approaches to compute the similarities of DNA sequences based on 2-D or 3-D graphical representations[2-10]. Graphical representation of DNA sequence provides a simple way of viewing, sorting and comparing various gene structures.

Ribonucleic acid(RNA) is an important molecule which performs a wide range of functions in the biological system. In particular, it is RNA(not DNA) that contains genetic information of virus such as HIV and therefore regulates the functions of such virus. RNA has recently become the center of much attention because of its catalytic properties, leading to an increased interest in obtaining structural information[11-15]. Similar with the graphical representations of DNA sequences, we can also outline several graphical representations of RNA primary sequences based on 2-D and 3-D to compute the similarities of RNA structures(either primary structures or secondary structures). Recently, B.Liao el present graphical method to compute the similarities of RNA secondary structures.[12-15]

In this paper, we shall propose a 4-D representation, which avoids the limitation associated with non-crossing, and make a comparison for the secondary structures. Based on the order of free bases A, G, U , C and base pairs A-U, C-G, we shall reduce a RNA secondary structure into three matrices. We construct a 3-component vector consisting of the leading eigenvalues or normalized leading eigenvalues of the L/L matrices. The similarities are computed by calculating the Euclidean distance between the end points of the vectors or by calculating the correlation angle of two vectors. In Figure 1, the secondary structures at

*Corresponding author E-mail: *mount_cn*@163.*com*

the 3′-terminus belonging to nine different viruses are listed, which were reported by John F.Bol[1]. The examinations of similarities/dissimilarities among the secondary structures at the 3′-terminus of different viruses illustrate the utility of our approach.

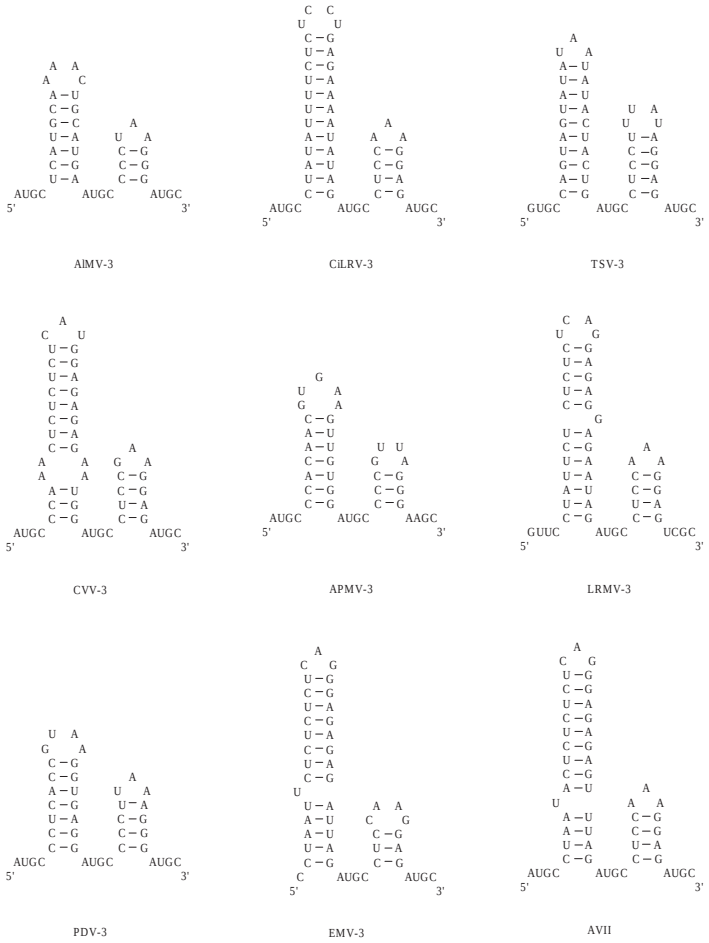

Figure 1: Secondary structure at the 3'-terminus of RNA 3 of alfalfa mosaic virus(AlMV-3 [16]), citrus leaf rugose virus(CiLRV-3 [17]), tobacco streak virus (TSV-3 [18,19]), citrus variegation virus(CVV-3 [17]), apple mosaic virus (APMV-3 [20]), prune dwarf ilarvirus(PDV-3 [21]),lilac ring mottle virus(LRMV-3 [22]), elm mottle virus(EMV-3 [23]) and asparagus virus II(AVII[24]). Numbering of nucleotides is from the 3'end of RNA 3.

2 4-D representation of RNA secondary structures

The secondary structure of an RNA is a set of free bases and base pairs forming hydrogen bonds between A-U and G-C. Let A', U', G', C' denote A, U, G, C in the base pair A-U and G-C, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure. For example, the substructure of AlMV-3 corresponds the characteristic sequence $G'G'G'AAUC'C'C'$, pseudoknot B corresponds the characteristic sequence $C'U'G'G'C'G'AUUGCG'A'G'A'C'C'A'UGUC'G'C'$ $C'A'G'CUCU'G'G'U'C'U'C'CA$ (from $3'$ to $5'$)(see Figure 2).

Figure 2: A pseudoknot and substructure of AlMV-3

We will illustrate the four-dimensional characterization of RNA secondary structure. We construct a map between the bases of characteristic sequences and plots in 4-D space, then we will obtain a 4-D representation of the corresponding RNA secondary structure. In 4-D space points, vectors and directions have four components, and we will assign the following basic elementary directions to the four free bases and two base pairs.

free base A $(-1,0,0,0)$; U $(1,0,0,0)$
C $(0,-1,0,0)$; G $(0,1,0,0)$

base pair $A'(-1,0,1,0)$; $U'(1,0,1,0)$
$C'(0,-1,-1,0)$; $G'(0,1,-1,0)$

We will reduce a RNA secondary structure into a series of nodes $P_0, P_1, P_2, \ldots, P_n$, whose coordinates $x_i, y_i, z_i, s_i (i = 0, 1, 2, \ldots, n$, where n is the length of the RNA secondary structure

being studied)satisfy

$$\begin{cases} x_i=-1, y_i=0, z_i=0, s_i=i & \text{if } g_i=A \\ x_i=1, y_i=0, z_i=0, s_i=i & \text{if } g_i=U \\ x_i=0, y_i=-1, z_i=0, s_i=i & \text{if } g_i=C \\ x_i=0, y_i=1, z_i=0, s_i=i & \text{if } g_i=G \\ x_i=-1, y_i=0, z_i=1, s_i=i & \text{if } g_i=A' \\ x_i=1, y_i=0, z_i=1, s_i=i & \text{if } g_i=U' \\ x_i=0, y_i=-1, z_i=-1, s_i=i & \text{if } g_i=C' \\ x_i=0, y_i=1, z_i=-1, s_i=i & \text{if } g_i=G' \end{cases}$$

In other words, let $G=g_1g_2\cdots$ be an arbitrary characteristic sequence of RNA secondary structure, we have a map ϕ, which maps G into a plot set. Explicitly, $\phi(G)=\phi(g_1)\phi(g_2)\cdots$, where

$$\phi(g_i)=\begin{cases} (-1,0,0,i) & \text{if } g_i=A \\ (1,0,0,i) & \text{if } g_i=U \\ (0,-1,0,i) & \text{if } g_i=C \\ (0,1,0,i) & \text{if } g_i=G \\ (-1,0,1,i) & \text{if } g_i=A' \\ (1,0,1,i) & \text{if } g_i=U' \\ (0,-1,-1,i) & \text{if } g_i=C' \\ (0,1,-1,i) & \text{if } g_i=G' \end{cases}$$

For example, the corresponding plot set of the substructure of AlMV-3 (Figure 2)is
$\{(0,1,-1,1),(0,1,-1,2),(0,1,-1,3),(-1,0,0,4),(-1,0,0,5),(1,0,0,6),$
$(0,-1,-1,7),(0,-1,-1,8),(0,-1,-1,9)\}$

We call the corresponding plot set characteristic plot set. The curve connecting all plots of the characteristic plot set in turn is called characteristic curve.

Bases of RNA can be classed into groups: purine(A,G)/pyrimidine(C,U), amino(A,C)/keto(G,U) and week-H bond(A,U)/strong-H band(G,C). We can obtain only three representations corresponding to the three classifications. The map ϕ and the following maps ϕ', ϕ'' correspond the three classifications. We call them pattern $AGUC, AUCG, AUGC$.

$$\phi'(g_i)=\begin{cases} (-1,0,0,i) & \text{if } g_i=A \\ (1,0,0,i) & \text{if } g_i=C \\ (0,-1,0,i) & \text{if } g_i=U \\ (0,1,0,i) & \text{if } g_i=G \\ (-1,0,1,i) & \text{if } g_i=A' \\ (0,-1,1,i) & \text{if } g_i=U' \\ (1,0,-1,i) & \text{if } g_i=C' \\ (0,1,-1,i) & \text{if } g_i=G' \end{cases}$$

$$\phi''(g_i) = \begin{cases} (-1,0,0,i) & \text{if } g_i = A \\ (1,0,0,i) & \text{if } g_i = G \\ (0,-1,0,i) & \text{if } g_i = U \\ (0,1,0,i) & \text{if } g_i = C \\ (-1,0,1,i) & \text{if } g_i = A' \\ (0,-1,1,i) & \text{if } g_i = U' \\ (0,1,-1,i) & \text{if } g_i = C' \\ (1,0,-1,i) & \text{if } g_i = G' \end{cases}$$

3 Similarities/dissimilarities among the RNA secondary structures of nine virus

Since we have no graphical representation to be associated with a random walk in 4-D space, in order to find some of the invariants sensitive to the RNA secondary structure, we will transform the 4D representation of the RNA secondary structure into another mathematical object, a matrix. Once we have a matrix representing a RNA secondary structure, we can use some of matrix invariants as descriptors of the sequence. One of the matrices is the L/L matrix whose elements are defined as $l_{i,j} = \frac{d_{i,j}}{\sum_{k=i}^{j-1} d_{k,k+1}}, i, j = 1, 2, \ldots, n$, where $d_{i,j}$ is the Euclidean distance between a pair of vertices, N is the length of RNA secondary structure. Its eigenvalues $\lambda_t, t = 1, 2, \ldots, n$, and in particular its leading eigenvalue can be used as descriptors of a RNA secondary structure. λ_t satisfies equation $(L/L)\beta = \lambda_t\beta$, where β is a vector corresponding to eigenvalue λ_t. Leading eigenvalue is the largest eigenvalue, we assume it is λ_1.

We will characterize the coding sequences of the RNA secondary structure of 9 species, shown in Figure 1, by means of the leading eigenvalue of the L/L matrices. Obviously, any RNA secondary structure corresponds three leading eigenvalue $\lambda_1^1, \lambda_1^2, \lambda_1^3$, which belongs to three L/L matrix respectively. In Table 1 we list the leading eigenvalues of the L/L matrices associated with three essentially different patterns of the characteristic curves representing each of the coding sequences.

Table 1: The leading eigenvalues of the L/L matrices associated with three essentially different patterns of the characteristic curves for the coding sequences of Figure 1

Patterns	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AUGC	22.1290	28.8686	25.1672	26.5094	23.2095	26.7110	23.5671	24.9052	27.0960
AUCG	21.7990	27.7213	23.7221	26.3873	23.6588	26.3700	23.3180	24.5620	26.9276
AGUC	22.3609	27.4667	24.9228	25.2421	23.9160	24.9357	23.4157	23.2203	25.6693

Next, we will illustrate the use of the 4-D quantitative characterization of RNA secondary structure with the examination of similarities/dissimilarties among the 9 coding sequences. We construct 3-component vectors$(\lambda_1^1, \lambda_1^2, \lambda_1^3)$ and $(\lambda_1^1/n, \lambda_1^2/n, \lambda_1^3/n)$, where $\lambda_1^1, \lambda_1^2, \lambda_1^3$ are the

leading eigenvalues of L/L matrices, n is the number of bases making up the corresponding RNA secondary structures.

The similarities among such vectors can be computed in two ways: (1) to calculate the Euclidean distance between the end points of the vectors; (2)to calculate the correlation angle of two vectors. The smaller Euclidean distance between the end points of two vectors, the more similar the RNA secondary structure. And, the smaller correlation angle between two vectors, the more similar the RNA secondary structure[7]. For example, in Figure 3, vector $\overrightarrow{oa}$ and vector $\overrightarrow{ob}$ correspond the coding sequence of AlMV-3 and CiLRV-3, respectively. $\|\overrightarrow{ba}\|$ is the Euclidean distance between the end points of the vectors $\overrightarrow{oa}$ and $\overrightarrow{ob}$. α is the the correlation angle of two vectors $\overrightarrow{oa}$ and $\overrightarrow{ob}$.

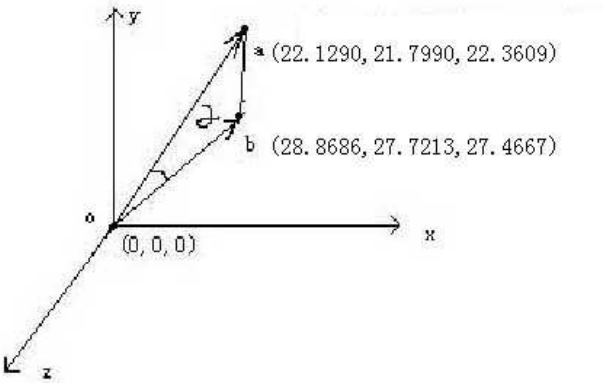

Figure 3:The correlation angle of two vectors

Obviously, if $\alpha = 0$ and $\|\overrightarrow{ba}\| = 0$, then vector $\overrightarrow{oa}$ and vector $\overrightarrow{ob}$ correspond the same structure, if $\alpha \to 90^o$ or $\|\overrightarrow{ba}\| \to \infty$, then the corresponding structure of $\overrightarrow{oa}$ and the corresponding structure of vector $\overrightarrow{ob}$ have little similarity.

In Table 2, we give the similarities and dissimilarities for the coding sequences of Figure 1 based on the Euclidean distances between the end points of the 3-component vectors $(\lambda_1^1, \lambda_1^2, \lambda_1^3)$. The most similar are LRMV-3 and CVV-3 with the lowest value 0.36718. The more similar are PDV-3 and APMV-3 with a value of o.7018, AVII and CVV-3 with a value of 0.90472, AVII and LRPMV-3 with a value of 0.99866.

In Table 3, the similarities and dissimilarities for 9 coding sequences that based on the correlation angle between two vectors. Observing Table 3, we find the more similar species pairs are EMV-3-LRMV-3, AVII-CVV-3, AVII-LRMV-3, AVII-EMV-3 and LRMV-3-CVV-3. The similar results can be found in Table 5.

In Table 4, the similarities and dissimilarities for 9 coding sequences that based on the Euclidean distances between the end points of the 3-component vectors$(\lambda_1^1/n, \lambda_1^2/n, \lambda_1^3/n)$. The more similar species pairs are (AVII, CVV-3), (AVII, LRMV-3), (LRMV-3, CVV-3) and (EMV-3, LRMV-3). For example, AVII and LRMV-3 have the same structure properties:

Table 2: The similarity/dissimilarity matrix for the coding sequences of Figure 1 based on the Euclidean distances between the end points of the 3-component vectors of the leading eigenvalues of the L/L matrices

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	10.32303	4.41501	6.96719	2.65418	6.96551	2.34348	4.00999	7.86887
CiLRV-3		0	6.01377	3.50631	7.819	3.58988	7.99318	6.61223	2.64627
TSV-3			0	3.00112	2.20233	3.0651	2.23373	1.91640	3.81481
CVV-3				0	4.48248	0.36718	4.62676	3.16115	0.90472
APMV-3					0	4.54433	0.7018	2.04332	5.37252
LRMV-3						0	4.63721	3.07772	0.99866
PDV-3							0	1.83764	5.52747
EMV-3								0	4.04886
AVII									0

Table 3: The similarity/dissimilarity matrix for the coding sequences of Figure 1 based on the angle between the end points of the 3-component vectors of the leading eigenvalues of the L/L matrices

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.024788	0.017889	0.030583	0.013919	0.037556	0.009142	0.037902	0.032476
CiLRV-3		0	0.024204	0.016774	0.033981	0.018935	0.018261	0.018968	0.017241
TSV-3			0	0.038036	0.031749	0.042532	0.021882	0.042664	0.039304
CVV-3				0	0.032912	0.008027	0.021462	0.008533	0.002102
APMV-3					0	0.040873	0.015791	0.041345	0.035014
LRMV-3						0	0.028596	0.000555	0.005944
PDV-3							0	0.028966	0.023381
EMV-3								0	0.006459
AVII									0

Table 4: The similarity/dissimilarity matrix for the coding sequences of Figure 1 based on the Euclidean distances between the end points of the 3-component vectors of the normalized leading eigenvalues of the L/L matrices

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.038071	0.112615	0.117063	0.020698	0.119905	0.012460	0.129305	0.116917
CiLRV-3		0	0.084720	0.085393	0.055897	0.086924	0.042107	0.096491	0.084880
TSV-3			0	0.033121	0.130251	0.037083	0.12707	0.039110	0.034198
CVV-3				0	0.132646	0.007061	0.123824	0.013182	0.001921
APMV-3					0	0.135677	0.017117	0.145123	0.13253
LRMV-3						0	0.126211	0.009764	0.005457
PDV-3							0	0.135799	0.123532
EMV-3								0	0.012813
AVII									0

Table 5: The similarity/dissimilarity matrix for the coding sequences of Figure 1 based on the cosine of the angle between the end points of the 3-component vectors of the leading eigenvalues of the L/L matrices

Species	AlMV-3	CiLRV-3	TSV-3	CVV-3	APMV-3	LRMV-3	PDV-3	EMV-3	AVII
AlMV-3	0	0.999693	0.999840	0.999532	0.999903	0.999295	0.999958	0.999282	0.999473
CiLRV-3		0	0.999707	0.999859	0.999423	0.999821	0.999833	0.999820	0.999851
TSV-3			0	0.999277	0.999496	0.999096	0.999761	0.999090	0.999228
CVV-3				0	0.999458	0.999968	0.99977	0.999964	0.999998
APMV-3					0	0.999165	0.999875	0.999145	0.999387
LRMV-3						0	0.999591	0.999999	0.999982
PDV-3							0	0.999581	0.999727
EMV-3								0	0.999979
AVII									0

a single bugle loop, two hairpin loops, a free 4-base leader, a free 4-base trailer and a free base 4-base separator between the two hairpin structures. The similar results can be found in references[12-15].

The Euclidean distance between end points of vectors and the correlation angle between vectors are different measures of the similarity of RNA secondary structures. Observing Table 2 and Table 3 we find that there exists an overall qualitative agreement among similarities despite some difference. In general, the correlation angle is the best measure for the similarities. Obviously, the correlation angle and the cosine of the correlation angle are equivalent for measuring the similarity.

4 Conclusion

We have proposed a 4D representation based on the classifications of bases and base pairs, and presented a similarity measure between RNA secondary structures. A simple 4D representation substitutes the complicated molecular structure. The advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different RNA structures, and avoids the limitation of non-crossing. In our approach, the insertion, deletion, and substitution of plots correspond to the insertion, deletion, and substitution of letters in the compared structures, respectively. One difference from the alignments of RNA secondary structures is that our approach considers not only sequence structures but also chemical structures of RNA secondary structures.

5 Acknowledgment

This work is supported by the Natural Science Foundation of China Grant NO.70501010.

References

[1] C. B. E. M. Reusken, F. B. John, Structural elements of the 3'-terminal coat protein binding site in alfalfa mosaic virus RNAs, Nucleic Acids Res. 14 (1996) 2660-2665.

[2] M. Randic, M. Vracko, N. Lers, D. Plavsic, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202-207.

[3] M. Randic, M. Vracko, On the similarity of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40 (2000) 599-606.

[4] M. Randic, Condensed representation of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40 (2000) 50-56.

[5] C. X. Yuan, B. Liao, T. M. Wang, New 3-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 379 (2003) 412-417.

[6] B. Liao, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chem. Phys. Lett. 388 (2004) 195-200.

[7] B. Liao, T. M. Wang, Analysis of similarity of DNA sequences based on triplets, J. Chem. Inf. Comput. Sci. 44 (2004) 1666-1670.

[8] B. Liao, K. Q. Ding, A graphical approach to analyzing DNA sequences, J. Comput. Chem. 26 (2005) 1519-1523.

[9] B. Liao, Y. S Liu, R. F. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, Chem. Phys. Lett. 421 (2006) 313318.

[10] B. Liao, M. S. Tan, K. Q. Ding, Application of 2D graphical representation of DNA sequence, Chem. Phys. Lett. 414 (2005) 296-300.

[11] B. Liao, T. M. Wang, General combinatorics of RNA hairpins and cloverleaves, J. Chem. Inf. Comput. Sci. 43 (2003) 1138-1142.

[12] B. Liao, T. M. Wang, A 3D Graphical representation of RNA secondary structures, J. Biomol. Struct. Dynamics 21 (2004) 827-832.

[13] B. Liao, K. Q. Ding, T. M. Wang, On a six-dimensional representation of RNA secondary structures, J. Biomol. Struct. Dynamics 22 (2005) 455-464.

[14] B. Liao, T. M. Wang, K. Q. Ding, On a seven-dimensional representation of RNA secondary structures, Mol. Simulat. 14 (2005) 1-9.

[15] W. Zhu, B. Liao, K. Q. Ding, A condensed 3D graphical representation of RNA secondary structures, J. Mol. Struct. (Theochem) 757 (2005) 193-198.

[16] E. C. Koper-Zwarthoff, F. T. Brederode, P. Walstra, J. F. Bol, Nucleotide sequence of the 3'noncoding region of alfalfa mosaic virus RNA4 and its homology with the genomic RNAs, Nucleic Acids Res. 7 (1979) 1887-1900.

[17] S. W. Scott, X. Ge, The complete nucleotide sequence of RNA 3 of citrus leaf rugose and citrus variegation ilarviruses, J. Gen. Virol. 76 (1995) 957-963.

[18] E. C. Koper-Zwarthoff, F. T. Brederode, P. Walstra, J. F. Bol, Nucleotide sequence of the putative recognition site for coat protein in the RNAs of alfalfa mosaic virus and tobacco streak virus, Nucleic Acids Res. 8 (1980) 3307-3318.

[19] B. J. C. Cornelissen, H. Janssen, D. Zuidema, J. F. Bol, Complete nucleotide sequence of tobacco streak virus RNA 3, Nucleic Acids Res. 12 (1984) 2427-2437.

[20] R. H. Alrefai, P. J. Shicl, L. L. Domier, C. J. D'Arcy, P. H. Berger, S. S. Korban, The nucleotide sequence of apple mosaic virus coat protein gene has no similarity with other Bromoviridae coat protein genes, J. Gen. Virol. 75 (1994) 2847-2850.

[21] S. W. Scott, X. Ge, The complete nucleotide sequence of the RNA 3 of lilac ring mottle ilarvirus, J. Gen. Virol. 76 (1995) 1801-1806.

[22] E. J. Bachman, S. W. Scott, G. Xin, V. V. Bowman, Complete nucleotide sequence of prune dwarf viral RNA 3: Implications for coat protein activation of genome replication in ilarviruses, Virology 201 (1994) 127-131.

[23] F. Houser-Scott, M. L. Baer, K. F. Liem, J. M. Cai, L. Gehrke, Nucleotide sequence and structural determinants of specific binding of coat protein or coat protein peptides to the 3' untranslated region of alfalfa mosaic virus RNA 4, J. Virol. 68 (1994) 2194-2205.

[24] EMBL/GenBank/DDBJ databases. Accession no. X86352.