Virus Classifications Based on the Haar Wavelet Transform of Signal Representation of DNA Sequences

MOHAMED EL-ZANATY [1], MAGDY SAEB [1], A. BAITH MOHAMED[1],
SHAWKAT K. GUIRGUIS[2], EMAN EL-ABD [3]

1. School of Engineering, Computer Department,
Arab Academy for Science, Technology and Maritime Transport
Alexandria, EGYPT
2. Information Technology Department,
Institute of Graduate Studies & Research,
Alexandria University, EGYPT
3. Molecular Biology Group, Medical Technology Center,
Medical Research Institute,
Alexandria University, EGYPT

***Abstract*:** Various ways of classifying plant viruses are tried as virus classification may reveal possible evolutionary relationship and enable prediction of the properties of unexampled viruses. Viral molecular taxonomy represents an important achievement to reach an unbiased objective classification. In order to simplify and accelerate the viral molecular taxonomy, a digital signal representing of viral genomes is presented. The approach applies known signal processing techniques for the analysis of genomic information. A classification method is proposed by applying the Haar Wavelet Technique on the resulting genome signals. Based on multi-level Haar transform, the search starts at the nth trend and follow the various levels upward until a match is found. The Quick Response (QR) representation of genome sequences may prove to bear practical applications in sequence alignment and phylogenetic analysis.

***Key words*:** DNA, Plant viruses, viral molecular taxonomy, Haar Wavelet, Multi-resolution Analysis, Quick Response.

1. Introduction

Plant viruses were first classified according to the properties of the viral particles based on host, symptoms, transmission route, vectors, morphology, chemistry, serology, cross protection, and the disease they cause [1]. Viruses lie in the shaded area between living and non-living matters, therefore when they are present in the living form they will be selected for their ability to sequester the host metabolic systems. However, when present in the non-living form they will be selected for stability and efficient dispersion. Hence, other classification systems are necessitated where genetic information of viral particles and host-virus interactions are present in order to achieve unbiased classification [1, 2].

The classical taxonomy named after Linnaeus, the Linnaean system, grouped species into a hierarchy of increasing inclusive categories [3]. In the first grouping level, species that appear to be closely related are grouped in the same genus. Then related genera are placed in the same family, families into orders, orders into classes, classes into phyla, phyla into kingdoms, and more recently, kingdoms into domains. The classification levels become more specific towards the bottom. Many organisms belong to the same kingdom; fewer belong to the same phylum, and so forth with species being the most specific classification. The problem with this taxonomic system is that it tries to perform incompatible tasks such as naming, classifying, and searching for similarity and phylogeny all at the same time.

Computer classification was initially based on Adanson’s method [4] where all possible characters given the same weight then defining it using groups of correlated characters. Other computer classification programs relying on qualitative and quantitative or multistate properties have been developed [1]. Additional programs, likewise, identify the most used character in defining the clusters of classification [5]. Quite recently, we have developed a model

Received 28/12/2011 Reviewed 27/1/2012

where DNA nucleotides were represented in a compact form similar to Quick Response (QR) representation to reduce the storage requirements in the database [6]. To speed up the search process the Haar Wavelet technique was applied to the resulting DNA signals. Based on multi-level Haar transform, the search starts at the n^{th} trend and follow the various levels upward until a match is found. Thus applying Haar wavelet transform will divide the database into clusters with the same trends. These clusters can be regarded as classification groups that enable the researcher to identify the group of the viral query sequence.

2. Methodology

The Haar wavelet is the simplest type of wavelet, used for compressing signals and for removing noise. The Haar transform decomposes a discrete signal into two sub-signals; each is about half the original signal length. One sub-signal is a running average or "the trend" and the other sub-signal is a running difference or "the fluctuations". We will concentrate on the trend sub-signal. The first trend sub-signal, $a^1 = (a_1, a_2,\ldots,a_{N/2})$, for the signal **f** is computed by taking a running average in the following way. Its first value, a_1, is computed by taking the average of the first pair of values of **f**: $(f_1 + f_2) /2$, and then multiplying it by $\sqrt{2}$, That is:

$$a_1 = (f_1+f_2) / \sqrt{2}$$
$$a_2 = (f_3+f_4) / \sqrt{2}$$

And so on…

For example the binary signal:

$(0011110101101100)_2$

After applying the first round of Haar wavelet it takes the form

$(0, \sqrt{2},\sqrt{2},1/\sqrt{2}, 1/\sqrt{2}, 1/\sqrt{2},\sqrt{2},0)$

The second round of the Haar wavelet on the trend signal only it will give

$(1, 2, \sqrt{2}, 1)$

In this work we use two DNA signal representation methods, the first one is the single dimension binary representation and the second one is a binary matrix representation (QR). In this section we will apply the first one and in the next section the second will be applied.

Mapping the DNA sequences to binary representation is a simple straight forward procedure. For most tasks, a flat encoding of 2 bits/nucleotide, assigned in alphabetical order would be a sufficient starting point.

$A = (00)_2$	or	$A = 0_Q$
$C = (01)_2$	or	$C = 1_Q$
$G = (10)_2$	or	$G = 2_Q$
$T = (11)_2$	or	$T = 3_Q$

For example the DNA sequence

ATTCCGGAGCTAG

It will be represented in binary format as:

$(00111101011010001001110010)_2$

Therefore we can apply the Haar wavelet transform on the resulting binary signal [6]. In this work, the Haar wavelet transform is applied to the single dimension binary representation of 17 sequences (AB000472, AB000473, AB000474, AB027007, AB027008, AB027009, AB027010, AB004456, AB004457, AB004544, AB098341, AB098342, AB098343, AB098344, HQ588899, HQ588900, HQ588901) of 5 types of plant viruses including four types of ssRNA positive strand and one type of ssDNA viruses. Table 1 and Table 2 demonstrate the last three trends of each sequence in the group. The search process goes in a bottom-up fashion. It starts to compare the smallest trend signals and if it matches, the search goes upward. This process is repeated until reaching an exact match. The database is serving similar to a large tree and the search process advances through this tree branch. Figure 1 depicts the search process in the database starting from the tree root or the larger trend. This tree is actually a B-tree; therefore searching a B-tree is $O(t\log_t(n))$ where t is the block size. As observed $O(t\log_t(n)) < O(n)$;

therefore this method provides faster search time as compared to the traditional string matching methods [10, 11, 12, 13,14].

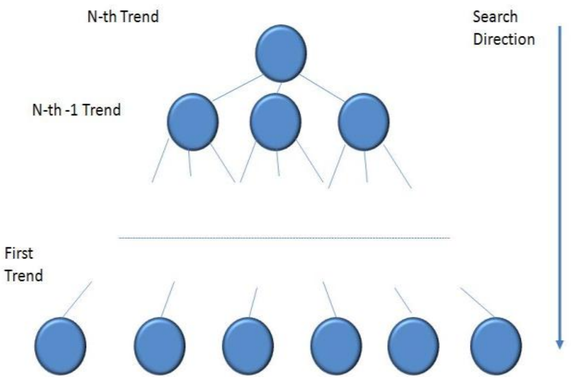
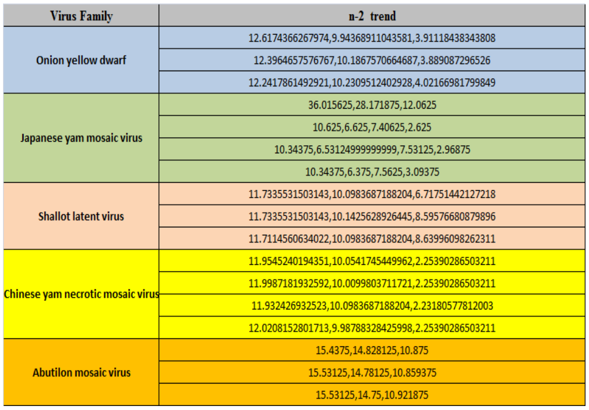


Fig.1: The search process in the DNA database.

Based on the Haar wavelet transform, the method provides a numeric value that identifies evolutionary groups. The fourth trend can be thought of as the species family and the various levels as the sub-species and so on until reaching the individual using the original DNA signal. As a matter of fact, this type of classification is the numeric representations of taxonomic species groups. Studying the visual representation of the DNA provides an insight into the evolutionary process of a specific species [6].

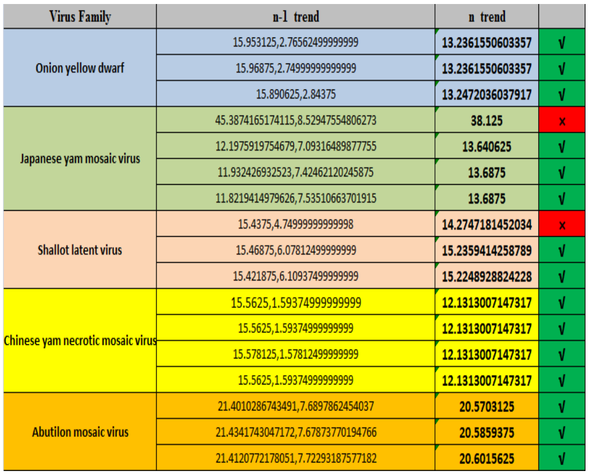
Table 1
The n-2th trend for five plant viruses' sequences



Virus Family	n-2 trend
Onion yellow dwarf	12.6174366267974,9.94368911043581,3.91118438343808
	12.3964657576767,10.1867570664687,3.889087296526
	12.2417861492921,10.2309512402928,4.02166981799849
Japanese yam mosaic virus	36.015625,28.171875,12.0625
	10.625,6.625,7.40625,2.625
	10.34375,6.53124999999999,7.53125,2.96875
	10.34375,6.375,7.5625,3.09375
Shallot latent virus	11.7335531503143,10.0983687188204,6.71751442127218
	11.7335531503143,10.1425628926445,8.59576680879896
	11.7114560634022,10.0983687188204,8.63996098262311
Chinese yam necrotic mosaic virus	11.9545240194351,10.0541745449962,2.25390286503211
	11.9987181932592,10.0099803711721,2.25390286503211
	11.932426932523,10.0983687188204,2.23180577812003
	12.0208152801713,9.98788328425998,2.25390286503211
Abutilon mosaic virus	15.4375,14.828125,10.875
	15.53125,14.78125,10.859375
	15.53125,14.75,10.921875

Table 1 presents the n-2th trends for each group sequences and separates each group with different color. As observed there a small difference between the trends' numbers.

Table 2
The n-1th and nth trend for five plant viruses' sequences



Virus Family	n-1 trend	n trend	
Onion yellow dwarf	15.953125,2.76562499999999	13.2361550603357	√
	15.96875,2.74999999999999	13.2361550603357	√
	15.890625,2.84375	13.2472036037917	√
Japanese yam mosaic virus	45.3874165174115,8.52947554806273	38.125	x
	12.1975919754679,7.09316489877755	13.640625	√
	11.932426932523,7.42462120245875	13.6875	√
	11.8219414979626,7.53510663701915	13.6875	√
Shallot latent virus	15.4375,4.74999999999998	14.2747181452034	x
	15.46875,6.07812499999999	15.2359414258789	√
	15.421875,6.10937499999999	15.2248928824228	√
Chinese yam necrotic mosaic virus	15.5625,1.59374999999999	12.1313007147317	√
	15.5625,1.59374999999999	12.1313007147317	√
	15.578125,1.57812499999999	12.1313007147317	√
	15.5625,1.59374999999999	12.1313007147317	√
Abutilon mosaic virus	21.4010286743491,7.6897862454037	20.5703125	√
	21.4341743047172,7.67873770194766	20.5859375	√
	21.4120772178051,7.72293187577182	20.6015625	√

Table 2 extends Table 1 and presents the n-1th and nth trends. As observed that the difference between trends' numbers decreases from n-2th to nth.

3. Results

As observed from Table 1 and Table 2, each group of sequences was classified according to its trend. The numbers of nth trend are very close to each other but there is a slight difference between numbers of the n-1th trend. This difference increases in the n-2th trend and so forth. This difference is contingent on the length of the sequence and hence the nucleotide composition.

The second method of DNA signal representation is the matrix representation of a DNA sequence which consists of four rows. Each row represents the occurrence of one DNA base (A, C, G or T). For example the DNA sequence with length 13 bases:

ATTCCGGAGCTAG

It will become in matrix representation as:

$$\begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{pmatrix} 1000000100010 \\ 0001100001000 \\ 0000011010001 \\ 0110000000100 \end{pmatrix}$$

As we examine the above shown matrix, the first row represents the occurrences of the (A) nucleotide in the DNA sequence and the second, third and fourth rows represent the (C), (G) and (T) occurrences respectively.
We can visually represent this matrix as a dot for each occurrence of the digit one and a blank for each occurrence of the digit zero that is very similar to QR (Quick Response) code.

Therefore we can apply the Haar wavelet transform on the resulting QR image (signal) [6]. This model had been applied on Humans and Chimps and the results were Chimps and Humans belong to the same evolutionary family. The fourth trend can be thought of as the species family and the various levels as the sub-species and so on until reaching the individual using the original DNA signal. As a matter of fact, this type of classification is the visual representations of taxonomic species groups. Studying the visual representation of the DNA provides an insight into the evolutionary process of a specific species [1].

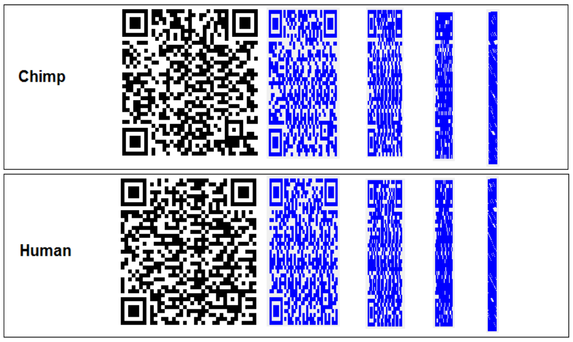
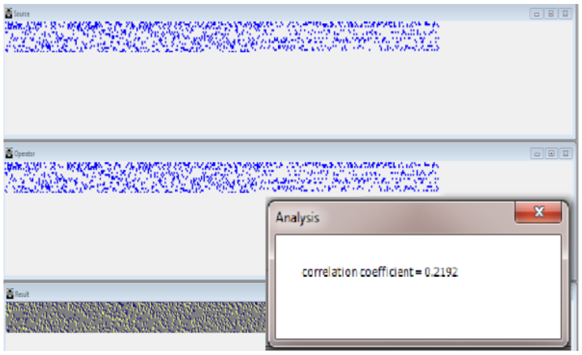


Fig.2: The QR code comparisons for trends' signals of DNA sequences of Chimps and Humans.

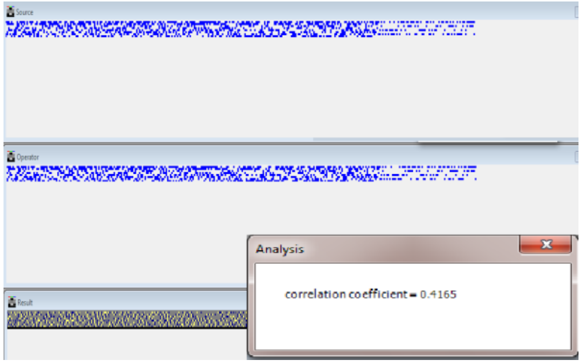
In this work we apply the Haar wavelet transform on the QR representation of the Onion Yellow Dwarf, which is an ssRNA positive strand virus, the viral sequences AB000472 and AB000473 (encoding 3` terminal region of coat protein) increased the correlation coefficient as depicted in Figures 3 to 7. The highest correlation coefficient was achieved at the fourth trend as seen from Figure 6. Therefore, the degree of relationship between the images increases when applying the Haar wavelet transform. This corresponds to the results represented in Table 1 and Table 2.
Afterwards, BLAST search [15] was applied on the viral sequences. The similarity ranged from 77% and 95%. The distance tree is represented in Figure B.1 shown in Appendix B. The relation was also affected by both length and nucleotide composition as seen from Table A.1 in Appendix A.



correlation coefficient = 0.2192

Fig.3: Correlation coefficient between the QR representations without applying Haar.

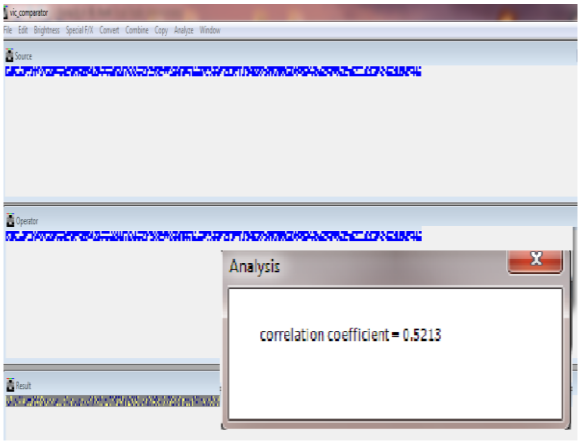
Figure 3 compares the QR representation of the first two sequences of the Onion yellow dwarf virus and computes the correlation coefficient to obtain the degree of relationship between the two images.



correlation coefficient = 0.4165

Fig.4: Correlation coefficient between the QR representations of the first trends.

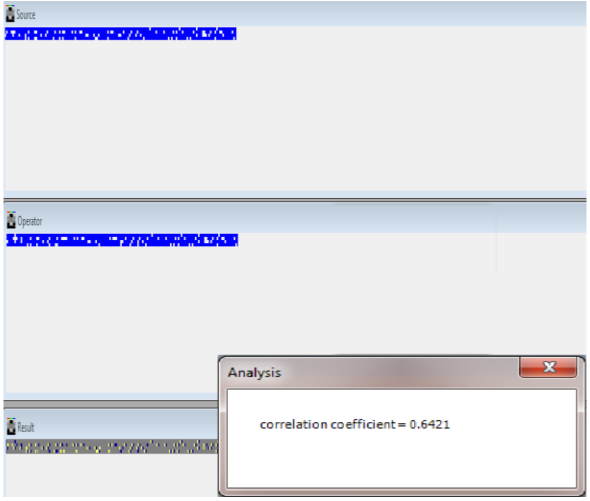
Figure 4 compares the QR representation of the first trend signals after applying one cycle of Haar Wavelet. As observed the correlation coefficient is increased, therefore the relation between the two sequences getting closer.


vic_comparator
File Edit Brightness Special F/X Convert Combine Copy Analyze Window
Source
Operator
Result
Analysis
X

correlation coefficient = 0.5213

Fig.5: Correlation coefficient between the QR representations of the second trends.

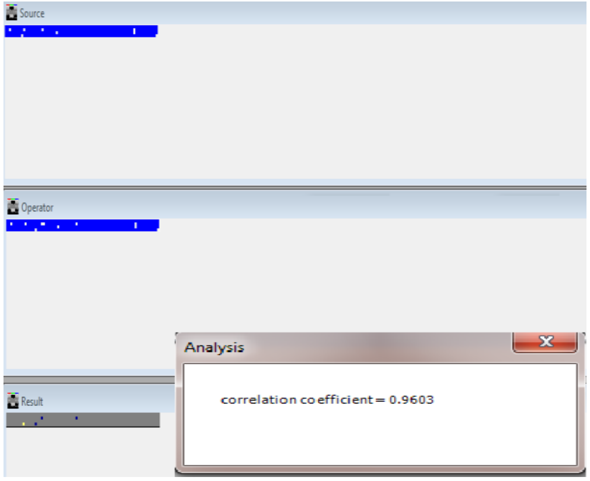
As observed from figure 5 the correlation coefficient is increased, therefore the relation between the two sequences getting closer.



correlation coefficient = 0.6421

Fig.6: Correlation coefficient between the QR representations of the third trends.

Applying the Haar Wavelet for the third time on the resulting QR images and calculates the correlation coefficient reveal that the two images getting very closer as shown in figure 6.



correlation coefficient = 0.9603

Fig.7: Correlation coefficient between the QR representations of the fourth trends.

As observed from the previous figures the more applying Haar Wavelet on QR images the more relationship between the QR images. Therefore the classification levels become more specific towards the bottom and become more generalized towards the top. Table 3 will summarize the previous figures.

Table 3
The correlation coefficients of QR images

Figure No.	**Haar Wavelet**	**Correlation Coefficient**	**Description**
3	No Haar	0.2192	The relationship is weak
4	First round	0.4165	The relationship is better
5	Second round	0.5213	The relationship is much better
6	Third round	0.6421	The relationship is getting closer
7	Fourth round	0.9603	The relationship is very close

Summary & Conclusion

Based on the previous discussion, one establishes the following:

- Applying the Haar Wavelet transform on the viruses' sequences, the proposed method provides a numeric value that identifies various evolutionary groups.
- The classification levels become more particular towards the bottom and become more general towards the top.
- The degree of relationship between the QR representations of viruses' sequences increases while applying more rounds of Haar Wavelet.
- BLAST search was performed for verification and the similarity ranged from 77% and 95%.

The proposed methodology serves in disclosing an unexampled classification technique based on the Haar Transform of DNA signal representation, sequence alignment and phylogenetic analysis.

References

[1] Gibbs A., "Plant Virus Classification," Advances in Virus Research, Volume 14, pp. 263-328, 1969.

[2] Ward CW., "Progress Towards a Higher Taxonomy of Viruses," Research in Virology 144, pp. 419-53, 1993.

[3] Hull R., "Nomenclature and Classification of Plant Viruses," Matthews' Plant Virology, 4th Ed., pp. 13-45, 2002.

[4] Adanson M., "Familles des Plantes," Vol 1, Vincent, Paris, 1963.

[5] Regenmortel MHV., "Virus Species and Virus Identification: past and current controversies," Infect Genet Evol, pp. 133-44, 2007.

[6] El-Zanaty M., Saeb M., A. Baith Mohamed, Guirguis S. K., "Haar Wavelet Transform of the Signal Representation of DNA Sequences," The International Journal of Computer Science & Communication Security (IJCSCS), Volume 1, pp. 56-62, July, 2011.

[7] Saeb M., El-Abd E., El-Zanaty M., "DNA Steganography using DNA Recombinant and DNA Mutagenesis Techniques," WSEAS Transaction on Computer Research, Vol. 2, pp. 50 -56, 2007.

[8] Smith WM. DNA-Based Steganography for Security Marking," Xix International Security Printers' conferences, Montreux, 14-16 May, 2003.

[9] Gehani A, LaBean T, Reif J., "DNA-based Cryptography," IMACS DNA-based computers V, American Mathematical Society, US., 2000.

[10] What is DNA Barcoding?
www. barcodeoflife.org,, 2010.

[11] Nielsen R, Matz M., "Statistical Approaches for DNA Barcoding," Society of Systematic Biologists, US., 2006.

[12] Law C, So S., "QR codes in Education," Journal of Educational Technology Development and Exchange (Hong Kong), pp. 85-100, 2010.

[13] Haimovich AD, Byrne B, Ramaswamy R, Welsh WJ., "Wavelet Analysis of DNA Walks," J Comput Biol , pp. 1289-1298, 2006.

[14] Arneodo A, D'Aubenton-Carafa Y, Audit B, Bacry E, Muzy JF, Thermes C., "What can we learn with wavelets about DNA sequences?," Elsevier Science, Physics A 249, pp. 439-448, 1998.

[15] Arneodo A, D'Aubenton-Carafa Y, Audit B, Bacry E, Vaillant C, Thermes C., "Extracting Structural and Dynamical Information from Wavelet-based Analysis of DNA Sequences," Coll. Group Theoretical Methods in Physics, IOP Publishing Ltd, France, 2003.

[16] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ., "Basic Local Alignment Search Tool,". J Mol Biol, pp. 403–410, 1990.

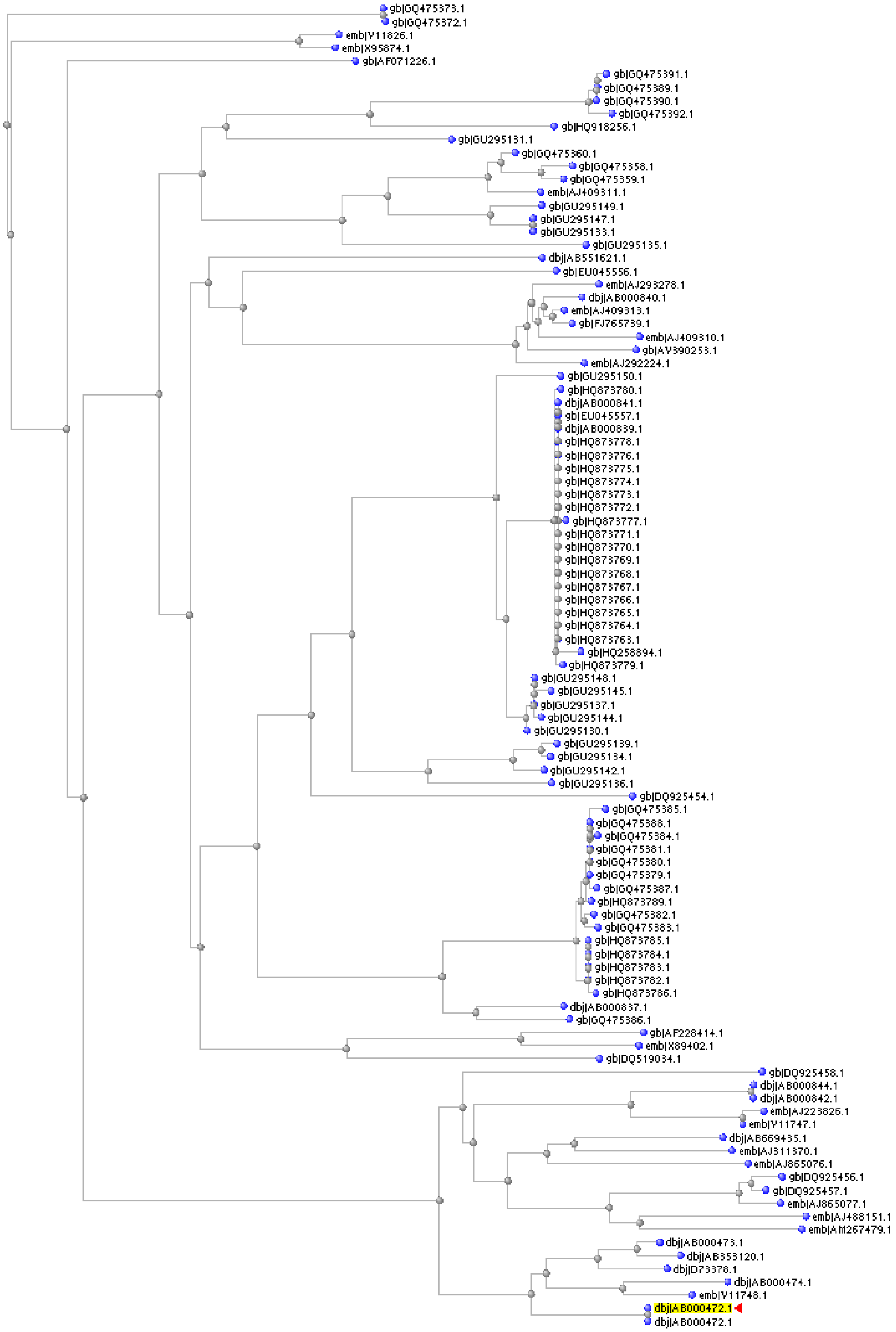
Appendix A: Length and base composition of viral sequences.

Table A.1: Length and base composition of viral sequences

Accession number	Length	N trend		Gene	Virus name	Virus type	Base composition			
							A	C	G	T
AB000472	1198bp	13.2361550603357	√	Coat protein, partial sequence, 3`terminal region	Onion yellow dwarf	ssRNA positive strand	34.56	18.2	22.29	24.46
AB000473	1198bp	13.2361550603357	√				34.22	17.7	23.21	24.87
AB000474	1199bp	13.2472036037917	√				33.61	17.26	23.77	25.35
AB027007	9760bp	38.125	×	Complete genome	Japanese yam mosaic virus	ssRNA positive strand	34.21	18.95	22.24	24.59
AB027008	873bp	13.640625	√	Coat protein, partial sequence, strain J1B			35.62	19.13	24.17	21.08
AB027009	876bp	13.6875	√	strain J2			35.16	17.58	25.11	22.15
AB027010	876bp	13.6875	√	strain J3			34.59	17.47	25.23	22.72
AB004456	1292bp	14.2747181452034	×	Coat protein & nucleic acid binding protein, clone CLC-1	Shallot latent virus	ssRNA positive strand	28.25	20.20	23.76	27.79
AB004457	1379bp	15.2359414258789	√	clone GC-2			28.14	20.38	23.71	27.77
AB004544	1378bp	15.2248928824228	√	clone WOC-2			27.87	20.39	23.73	28.01
AB098341	1098bp	12.1313007147317	√	Coat protein&3`UTR, isolate KK1	Chinese yam necrotic mosaic virus	ssRNA positive strand	32.24	19.85	23.68	24.23
AB098342	1098bp	12.1313007147317	√	isolate MD3			32.51	19.67	23.59	24.23
AB098343	1098bp	12.1313007147317	√	isolate NS2			32.33	19.67	23.63	24.32
AB098344	1098bp	12.1313007147317	√	isolate IB4			32.60	19.67	23.41	24.32
HQ588899	2633bp	20.5703125	√	Circular genomic DNA complete sequence of DNA-A segment	Abutilon mosaic virus	ssDNA viruses	24.23	23.05	23.32	29.40
HQ588900	2635bp	20.5859375	√				24.40	23.34	23.34	29.18
HQ588901	2637bp	20.6015625	√				24.33	23.06	23.28	29.28

Appendix B: Distance tree of viral sequences based on BLAST search.

Figure B.1: Distance tree of viral sequences based on BLAST search.


gb|GQ475373.1
gb|GQ475372.1
emb|Y11826.1
emb|X95874.1
gb|AF071226.1
gb|GQ475391.1
gb|GQ475389.1
gb|GQ475390.1
gb|GQ475392.1
gb|HQ918256.1
gb|GU295131.1
gb|GQ475360.1
gb|GQ475358.1
gb|GQ475359.1
emb|AJ409311.1
gb|GU295149.1
gb|GU295147.1
gb|GU295133.1
gb|GU295135.1
dbj|AB551621.1
gb|EU045556.1
emb|AJ293278.1
dbj|AB000840.1
emb|AJ409313.1
gb|FJ765739.1
emb|AJ409310.1
gb|AY390253.1
emb|AJ292224.1
gb|GU295150.1
gb|HQ873780.1
dbj|AB000841.1
gb|EU045557.1
dbj|AB000839.1
gb|HQ873778.1
gb|HQ873776.1
gb|HQ873775.1
gb|HQ873774.1
gb|HQ873773.1
gb|HQ873772.1
gb|HQ873777.1
gb|HQ873771.1
gb|HQ873770.1
gb|HQ873769.1
gb|HQ873768.1
gb|HQ873767.1
gb|HQ873766.1
gb|HQ873765.1
gb|HQ873764.1
gb|HQ873763.1
gb|HQ258894.1
gb|HQ873779.1
gb|GU295148.1
gb|GU295145.1
gb|GU295137.1
gb|GU295144.1
gb|GU295130.1
gb|GU295139.1
gb|GU295134.1
gb|GU295142.1
gb|GU295136.1
gb|DQ925454.1
gb|GQ475385.1
gb|GQ475388.1
gb|GQ475384.1
gb|GQ475381.1
gb|GQ475380.1
gb|GQ475379.1
gb|GQ475387.1
gb|HQ873789.1
gb|GQ475382.1
gb|GQ475383.1
gb|HQ873785.1
gb|HQ873784.1
gb|HQ873783.1
gb|HQ873782.1
gb|HQ873786.1
dbj|AB000837.1
gb|GQ475386.1
gb|AF228414.1
emb|X89402.1
gb|DQ519034.1
gb|DQ925458.1
dbj|AB000844.1
dbj|AB000842.1
emb|AJ223826.1
emb|Y11747.1
dbj|AB669435.1
emb|AJ311370.1
emb|AJ865076.1
gb|DQ925456.1
gb|DQ925457.1
emb|AJ865077.1
emb|AJ488151.1
emb|AM267479.1
dbj|AB000473.1
dbj|AB353120.1
dbj|D73378.1
dbj|AB000474.1
emb|Y11748.1
dbj|AB000472.1
dbj|AB000472.1