Thailand Statistician
July 2020; 18(3): 373-380
http://statassoc.or.th
Short Communication

Using Simple Statistics to Compare Genetic Sequences

Hossam Farag Abou-Shaara
Department of Plant Protection, Faculty of Agriculture, Damanhour University, Damanhour, Egypt.
*Corresponding author; e-mail: hossam.farag@agr.dmu.edu.eg

Received: 28 November 2018
Revised: 21 April 2019
Accepted: 9 November 2019

Abstract

The aim of this study was to compare different sequences using simple statistics. Firstly, sequences of four viruses were downloaded from the National Center for Biotechnology Information (NCBI). Then, these sequences were arranged in Excel sheets as numbers, and subjected to the statistical analysis using parametric and non-parametric tests. The obtained results were compared with those obtained by the phylogenetic analysis and gene cluster analysis for these viruses. The results of the statistical analysis, from ANOVA and Kruskal-Wallis test, were similar to those of phylogenetic relationships and shared gene clusters. It was possible to get additional information from the sequences using simple statistics either using parametric or non-parametric tests. The results of this study could help software developer and bioinformatics specialists to develop simple analytical methods to acquire information from the sequences.

Keywords: Parametric, non-parametric, significant, bioinformatics, phylogenetic.

1. Introduction

At the present time, there are many resources for genetic sequences, and fortunately most of these resources are free and available online. Also, the full sequences for many organisms are available including, human, fruit fly, honey bees, some bee viruses, plants and others (e.g. Spanos et al. 2000, Venter et al. 2001, Honeybee Genome Sequencing Consortium 2006, Beye et al. 2006, Jia et al. 2013). On the other side, there are various programs and websites that can be used to analyze the sequences. The available websites can be used to extract proteins, and gene clusters from sequences beside other tasks (Kanehisa and Goto 2000, Edgar et al. 2002, Gene Ontology Consortium 2004) while many softwares can be utilized to construct phylogenetic trees between studied organisms beside other options (Librado and Rozas 2009, Kearse et al. 2012, Kumar et al. 2016). These programs are based on many mathematical models and their results may vary for the same sequences. There is a need for more methods to extract information from the genetic sequences especially simple ones.

In fact, statistical analysis constitutes an essential part in genetic data analysis (e.g. Roff and Bentzen 1989, Schbath et al. 1995, Brown et al. 2001). Some programs can be used to perform the statistical analysis for data including SPSS. Such programs are used by many researchers to perform different tasks including means comparison for different treatments. It could be said that these statistical programs are well known for researchers than other specific programs for genetic analysis.

In this article, methods to compare sequences using parametric and non-parametric tests are presented. The obtained results were validated in light of the results obtained from other genetic programs. This study may encourage researchers to incorporate simple statistics in their genetic studies to compare sequences, and to develop new methods based on this study.

2. Methods

2.1. Sequences

The sequences of four viruses were downloaded from the National Center for Biotechnology Information (NCBI). These viruses were sacbrood virus strain II-9 (SBV1) (GenBank: JX270800.1), sacbrood virus strain S2 (SBV2) (GenBank: JX270799.1), black queen cell virus isolate JL1 (BQCV) (GenBank: KP119603.1), and Kakugo virus (KV) (GenBank: AB070959.1).

2.2. Sequences preparation in Excel sheets

The sequences were arranged in one column in Excel sheet (Figure 1A), and then nucleotides were changed into numbers as A = 1, G = 2, C = 3, and T = 4 (Figure 1B). These numbers were then used in the statistical analysis to compare sequences. Also, the nucleotides according to type were arranged in columns (one column to A, G, C, and T) in separated sheet, and inside each column number 1 was given to specific nucleotide type while number 0 to the other nucleotides (for example, in A column, only A = 1 and the other nucleotides G, C and T = 0) as shown in Figure 1C. These data were used to compare the nucleotides between viruses. Moreover, each 1000 nucleotides were placed together in one column (Figure 1D), to identify the differences between nucleotide sets of studied viruses. Number 0 was used to complete the missing nucleotide within the set (in case of sets with number of nucleotides < 1000). The statistical analysis was done using SPSS for windows version 16.0 (SPSS Inc. 2007) and the variations were considered significant when $p \leq 0.05$.

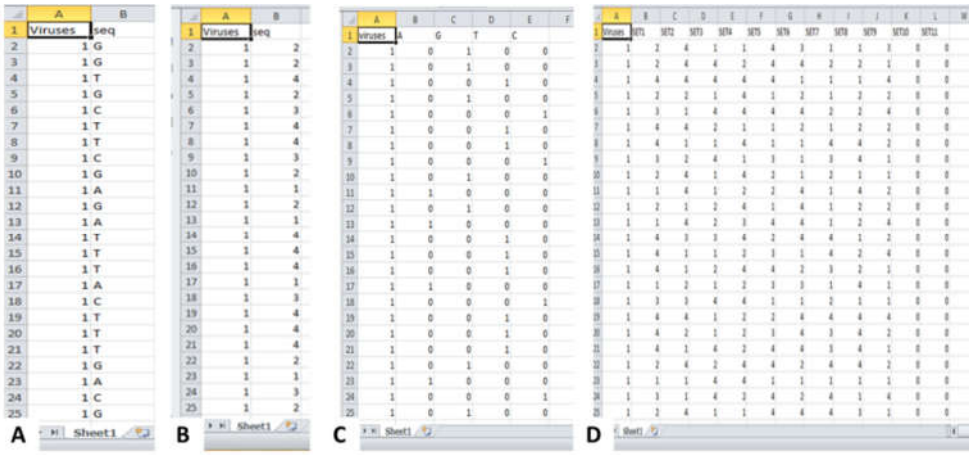

Figure 1 Preparation of sequences in Excel sheets; nucleotides in one column (A), nucleotides as numbers (B), nucleotides A, G, T, C in columns and numbers, and nucleotide sets as numbers (D)

2.3. Comparing sequences and nucleotides

Parametric test (One way ANOVA followed by Tukey test) and non-parametric test (Kruskal-Wallis test (k independent samples)) were used to identify the significant differences between full sequences, and nucleotides. To validate the obtained results, the phylogenetic relationships between viruses were done and compared with the statistical results. The phylogenetic analysis was done using MEGA7 (Kumar et al. 2016) based on maximum likelihood method and the Jukes-Cantor model (Jukes and Cantor 1969) after the alignment of sequences using ClustalW.

2.4. Comparing nucleotide sets

The nucleotide sets were analyzed using the previously mentioned parametric and non-parametric tests to identify the presence of significant differences between viruses. To validate the performed analysis, some nucleotide sets were subjected to gene cluster analysis. Nucleotide sets (1000 nucleotides per each set); 2 with significant and 2 with insignificant differences between viruses as shown from the statistical analysis were subjected to gene cluster analysis. Firstly, proteins for these sets were downloaded from Uniprot (http://www.uniprot.org), and then uploaded to OrthoVenn (http://www.bioinfogenome.net) to identify the gene cluster families at E-Value 1e-5, and inflation value of 1.5. The obtained results from gene cluster analysis were compared with those obtained from the statistical analysis.

3. Results and Discussion

3.1. Comparing sequences and nucleotides

The statistical analysis using the parametric test showed the presence of significant differences between sequences of the tested viruses (DF = 3, DF error = 35988, F = 5.56, $p = 0.001$), and the separated means (Table 1) showed the absence of significant differences between SBV1 and SBV2, and between BQCV and KV. The same significant differences were obtained by the non-parametric test (Table 1). The results obtained from the statistical analysis are supported by the phylogenetic relationships showed in Figure 2. The phylogenetic relationships show that SBV1 and SBV2 are very close to each other, while the other two viruses BQCV and KV are far from them. Therefore, the statistical analysis, parametric or non-parametric, of sequences can indicate the degree of relationships between studied organisms.

The variations in availability of nucleotides (A, G, T and C) between viruses were statistically tested (Table 1). No differences were detected between viruses in regard to nucleotide A (DF = 3, DF error = 35988, F = 0.26, $p = 0.854$) while significant differences were found between viruses in regard to nucleotide G (DF = 3, DF error = 35988, F = 9, $p = 0.00$), nucleotide T (DF = 3, DF error = 35988, F = 8.26, $p = 0.00$), and nucleotide C (DF = 3, DF error = 35988, F = 7.26, $p = 0.00$) according to parametric and non-parametric tests. From the separated means (Table 1), it is clear that KV was significantly different than SBV1 and SBV2 in regard to nucleotide G, T, and C. This result is in line with the sequence comparison and phylogenetic analysis, as SBV1 and SBV2 were separated than KV.

As shown from Table 2, no significant differences ($p > 0.05$) were found between means in all sets except set 1 (DF = 3, DF error = 3996, F = 4.57, $p = 0.003$), set 8 (DF = 3, DF error = 3996, F = 3.30, $p = 0.02$), set 9 (DF = 3, DF error = 3996, F = 199.37, $p = 0.00$), set 10 (DF = 3, DF error = 3996, F = 4.14, $p = 0.00$), and set 11 (DF = 3, DF error = 3996, F = 140.27, $p = 0.00$) according to the parametric test. Also, the same results were obtained by the non-parametric test (Table 2). The significant sets showed the presence of similarities between SBV1 and SBV2 unlike SK according to

the separated means (Table 2). This is an additional statistical confirmation on the close relationship between SBV1 and SBV2 than the other viruses.

Table 1 Means ± s.e. of sequences and nucleotides for sacbrood virus strain II-9 (SBV1), sacbrood virus strain S2 (SBV2), black queen cell virus isolate JL1 (BQCV), and kakugo virus (KV). Means were separated using Tukey test, and means followed by the same letter in the same column are not significantly different.

Viruses	Sequences	Nucleotides			
		A	G	T	C
SBV1	2.46 ± 0.01^{bc}	0.30 ± 0.005^{a}	0.24 ± 0.005^{a}	0.30 ± 0.005^{b}	0.16 ± 0.004^{b}
SBV2	2.45 ± 0.01^{c}	0.30 ± 0.005^{a}	0.24 ± 0.005^{a}	0.29 ± 0.005^{b}	0.17 ± 0.004^{ab}
BQCV	2.50 ± 0.01^{ab}	0.29 ± 0.005^{a}	0.22 ± 0.005^{b}	0.31 ± 0.005^{ab}	0.18 ± 0.004^{a}
KV	2.51 ± 0.01^{a}	0.29 ± 0.005^{a}	0.22 ± 0.004^{b}	0.32 ± 0.005^{a}	0.16 ± 0.004^{b}
KW test (Chi-Square, p-value)	(15.22, 0.002)	(0.77, 0.850)	(26.99, 0.000)	(24.77, 0.000)	(15.22, 0.000)

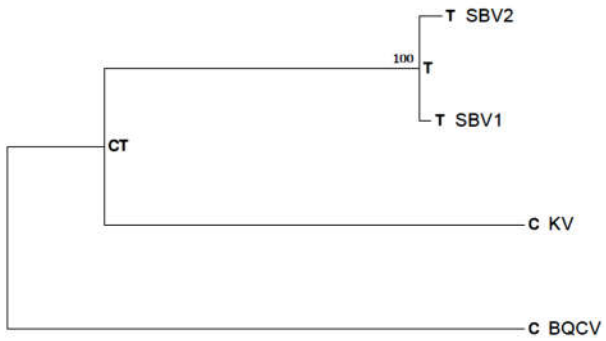

Figure 2 Phylogenetic relationships between viruses; Sacbrood virus strain II-9 (SBV1), Sacbrood virus strain S2 (SBV2), Black queen cell virus isolate JL1 (BQCV), and Kakugo virus (KV). The letters C and T denote the expected common ancestors

The sums of nucleotide values, considering A = 1, G = 2, C = 3, and T = 4, for the viruses are shown in Figure 3. Each set contains 1000 nucleotides, hence the high sum is an indicator to the abundance of nucleotide T. From this figure, it is clear that KV had more nucleotide T than the other viruses in all sets except set 4, 6, 7 and 8 as SBV2 and BQCV had high number of nucleotide T.

Sets 8 and 9 showed significant differences between viruses, and sets 2 and 5 showed insignificant differences between viruses were subjected to gene cluster analysis (Figures 4 and 5, respectively). The significant sets (Figure 4) showed the overlapping between SBV1 and SBV2 in 276 and 167 clusters for sets 8 and 9, respectively, while the other viruses shared few numbers of clusters from 1 to 21. The insignificant sets (Figure 5) showed the overlapping between SBV1 and SBV2 in 316 and 113 clusters for set 2 and 5, respectively, while the other viruses shared many clusters from 1 to 66. This analysis supports the close relationship between SBV1 and SBV2 as shown by the previous analyses. Also, significant sets between viruses based on the statistical analysis had few overlapping gene clusters than insignificant sets. Thus, the statistical analysis of nucleotide sets can help to detect the sets with and without high overlapping between studied organisms.

Table 2 Means ± s.e. of nucleotide sets (each set contain 1000 nucleotides) for Sacbrood virus strain II-9 (SBV1), Sacbrood virus strain S2 (SBV2), Black queen cell virus isolate JL1 (BQCV), and Kakugo virus (KV). Means were separated using Tukey test, and means followed by the same letter in the same column are not significantly different. Significant differences according to Kruskal-Wallis test are presented in bold

Viruses	Nucleotide sets										
	1	2	3	4	5	6	7	8	9	10	11
SBV1	2.42±0.03^{b}	2.45±0.03^{a}	2.45±0.03^{a}	2.51±0.03^{a}	2.48±0.03^{a}	2.45±0.03^{a}	2.49±0.03a	2.49±0.03^{a}	1.80±0.04^{b}	0.00±0.00^{b}	0.00±0.00^{b}
SBV2	2.39±0.03^{b}	2.44±0.03^{a}	2.42±0.03^{a}	2.54±0.03^{a}	2.48±0.03^{a}	2.47±0.03^{a}	2.49±0.03^{a}	2.41±0.03^{ab}	1.78±0.04^{b}	0.00±0.00^{b}	0.00±0.00^{b}
BQCV	2.43±0.03^{b}	2.49±0.03^{a}	2.46±0.03^{a}	2.51±0.03^{a}	2.48±0.03^{a}	2.49±0.03^{a}	2.55±0.03^{a}	2.53±0.03^{a}	0.94±0.04^{c}	0.00±0.00^{b}	0.00±0.00^{b}
KV	2.58±0.03^{a}	2.51±0.03^{a}	2.54±0.03^{a}	2.53±0.03^{a}	2.54±0.03^{a}	2.48±0.03^{a}	2.42±0.03^{a}	2.50±0.03a	2.49±0.03^{a}	2.49±0.03^{a}	0.39±0.03^{a}
KW test											
Chi-Square	13.41	1.97	5.11	0.54	1.77	0.23	6.03	9.17	631.80	3.89	473.60
p-value	0.004	0.570	0.160	0.910	0.620	0.970	0.100	0.020	0.000	0.000	0.000

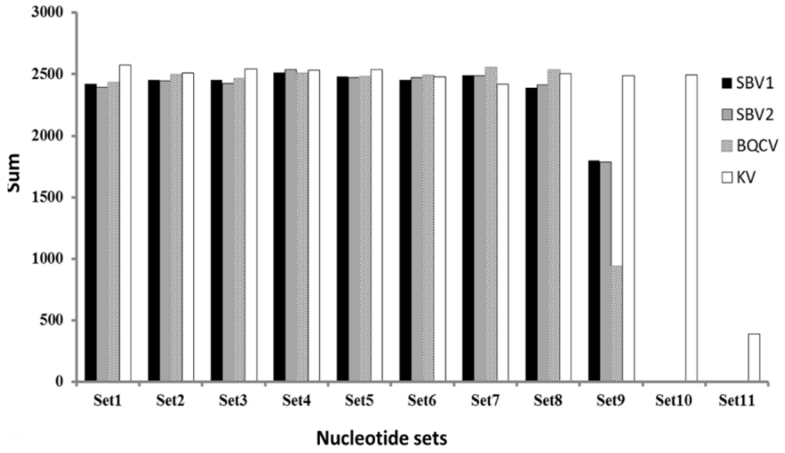

Figure 3 Sum of nucleotide values for 11 sets (set = 1000 nucleotides as numbers; A = 1, G = 2, C = 3, and T = 4). Sacbrood virus strain II-9 (SBV1), Sacbrood virus strain S2 (SBV2), Black queen cell virus isolate JL1 (BQCV), and Kakugo virus (KV)

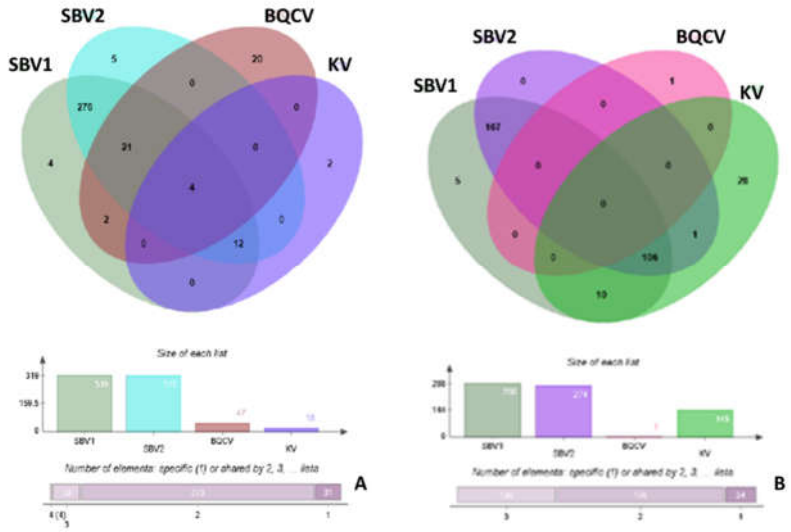

Figure 4 Gene clusters resulted from set 8 (A) and set 9 (B). Sacbrood virus strain II-9 (SBV1), Sacbrood virus strain S2 (SBV2), Black queen cell virus isolate JL1 (BQCV), and Kakugo virus (KV)

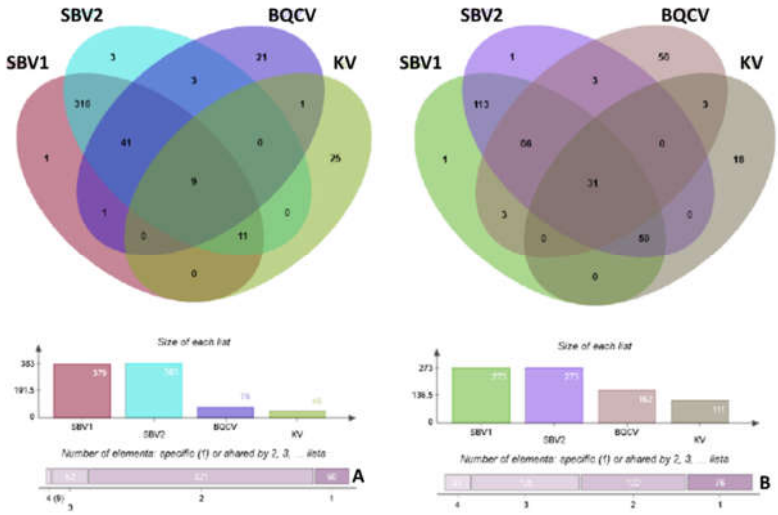

Figure 5 Gene clusters resulted from set 2 (A) and set 5 (B). Sacbrood virus strain II-9 (SBV1), sacbrood virus strain S2 (SBV2), black queen cell virus isolate JL1 (BQCV), and Kakugo virus (KV)

4. Conclusions

This study highlights the possibility of using parametric (ANOVA) and non-parametric (Kruskal-Wallis test) tests to compare full sequences and nucleotides, and to extract information from sequences using simple methods. The results obtained from simple statistics in this study were in line with those obtained from genetic programs. The ideas and methods presented in this study can inspire researchers to generate simple software to manage sequence data, and to develop simple statistical methods to analyze sequences to extract information.

Acknowledgements

Thanks are given to the anonymous reviewers for their comments on the manuscript. Also, the author wish to thank colleagues at the Genetics Division, Department of Plant Pathology for their support during the study.

References

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, Sirviö A. Exceptionally high levels of recombination across the honey bee genome. Gen Res. 2006; 16(11): 1339-1344.

Brown CS, Goodwin PC, Sorger PK. Image metrics in the statistical analysis of DNA microarray data. Proc Nat Acad Sci. 2001; 98(16): 8944-8949.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1): 207-210.

Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004; 32(1): 258-261.

Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006; 443(7114): 931-949.

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013; 496(7443): 91-95.

Jukes TH, Cantor CR. Evolution of protein molecules. In Munro HN, editor, Mammalian Protein Metabolism, New York: Academic Press, 1969.

Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28(1): 27-30.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 2012; 28(12): 1647-1649.

Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016; 33: 1870-1874.

Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11): 1451-1452.

Roff DA, Bentzen P. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol Biol Evol. 1989; 6(5): 539-545.

Schbath S, Prum B, de Turckheim E. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comput Biol. 1995; 2(3): 417-437.

Spanos L, Koutroumbas G, Kotsyfakis M, Louis C. The mitochondrial genome of the Mediterranean fruit fly, Ceratitis capitata. Insect Mol Biol. 2000; 9(2): 139-144.

SPSS Inc. SPSS for windows, version 16.0. Chicago: SPSS Inc., 2007.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD. The sequence of the human genome. Science. 2001; 291(5507): 1304-1351.