Virus Genes 28:1, 41–53, 2004

Base Usage and Dinucleotide Frequency of Infectious Bursal Disease Virus

TAN DO YEW,* MOHD HAIR BEJO, AINI IDERIS, ABDUL RAHMAN OMAR & GOH YONG MENG

Department of Veterinary Pathology and Microbiology, Faculty of Veterinary Medicine, Universiti Putra Malaysia, 43400, UPM, Serdang, Selangor, Malaysia

Received July 10, 2003; Accepted August 5, 2003

Abstract. Base usage and dinucleotide frequency have been extensively studied in many eukaryotic organisms and bacteria, but not for viruses. In this paper, a comprehensive analysis of these aspects for infectious bursal disease virus (IBDV) was presented. The analysis of base usage indicated that all of the IBDV genes possess equivalent overall nucleotide distributions. However when the base usage at each codon positions was analysed by using cluster analysis, the VP5 open reading frame (ORF) formed a different cluster isolated from the other genes. The unusual base usage of VP5 ORF may indicate that the gene was originated by the virus "overprinting strategy", a strategy in which virus may create novel gene by utilizing the unused reading frames of its existing genes. Meanwhile, the GC content of the IBDV genes and the chicken's coding sequences was comparable; suggesting the virus imitation of the host to increase its translational efficiency. The analysis of dinucleotide frequency indicated that IBDV genome had dinucleotide bias: the frequencies of CpG and TpA were lower and the TpG was higher than the expected. Classical methylation pathway, a process where CpG converted to TpG, may explain the significant correlation between the CpG deficiency and TpG abundance. "Principal component analysis of the dinucleotide frequencies" (DF-PCA) was used to analyse the overall dinucleotide frequencies of IBDV genome. DF-PCA on the hypervariable region and polyprotein (VPX-VP4-VP3) gene showed that the very virulent IBDV (vvIBDV) was segregated from other strains; which meant vvIBDV had a unique dinucleotide pattern. In summary, the study of base usage and dinucleotide frequency had unravelled many overlooked genomic properties of the virus.

Key words: base usage, bioinformatics, dinucleotide frequency, infectious bursal disease virus, principal component analysis, sequence analysis

Introduction

Infectious bursal disease (IBD) is an immunosuppressive disease that affects young chickens characterized by the destruction of bursa of Fabricius. Reviews of the disease have been published elsewhere [1–5]. IBD is caused by infectious bursal disease virus (IBDV), which is a double-stranded RNA (dsRNA) virus [6,7]. IBDV belongs to the genus *Avibirnavirus* [8] under the *Birnaviridae* family. Other genera of *Birnaviridae* are *Aquabirnavirus* and *Entomobirnavirus* [8]. IBDV genome consists of two segments, designated as segment A and B [6,7]. The genome is enclosed within a non-enveloped icosahedral capsid approximately 60 nm in diameter [9].

The complete nucleotide sequence of segment A is 3,261 bp [10] that contains two open reading frames (ORFs) of 3,036 bp [11] and 438 bp

*Author for all correspondence:
E-mail: dytan@vet.upm.edu.my

respectively, in which the smaller ORF partially overlaps at the 5′ end [12]. The large ORF encodes a precursor polyprotein (NH_2-VPX-VP4-VP3-COOH), which is autoproteolytically processed by *cis*-acting viral protease VP4 into VPX (48 kDa), VP3 (32 kDa), and VP4 (24 kDa) [13]. VPX, as a precursor protein, will undergo a second independent proteolytic processing step to yield a smaller matured product known as VP2 [14]. VP2 and VP3 form the viral capsid [15]. High conformational epitopes present in VP2 protein are responsible for the production of neutralizing antibody to protect the chicken from IBDV infection [16,17]. VP3 is the minor structural protein recognized by the non-neutralizing antibodies [18,19] and can efficiently bind to ssRNA and dsRNA [20].

The small ORF in segment A encodes VP5 protein with unknown function [21]. VP5 might be important in the pathogenesis [22] but is unessential for the viral replication and infection [22,23]. VP5 might also be involved in the release of viral progeny from infected cells [24]. VP5 gene overlaps VPX gene at its 35th nucleotide, therefore almost all of its nucleotides are within the VPX.

Segment B (2,827 bp [10]) consists of a single ORF that encodes for VP1 (90 kDa), a RNA-dependent RNA polymerase [25–27] with capping activities [28]. It has been reported that birnaviruses' polymerases formed a defined subgroup of polymerase by the lacking a GDD motif [29]. The formation of VP1–VP3 complexes plays a critical role in IBDV replication [30].

There are two serotypes of IBDV, namely serotype 1 and 2 [31,32]. In addition to serological classification, IBDV strains are also grouped according to their virulence (mortality and bursal lesions) [5]. The very virulent IBDV strain (vvIBDV) can cause up to 100% mortality and severe bursal lesions in specific-pathogen-free (SPF) chickens [33,34]. The classical virulent strain (cvIBDV) may cause bursal damage and mortality up to 30% [35]. Chickens infected by the variant strain (vaIBDV) may rapidly develop bursal atrophy without the inflammation phase [36] but the mortality caused by the vaIBDV can be less than 5% [5,37]. Attenuated strain (atIBDV) is usually derived from the attenuation of cvIBDV isolate and typically used as a vaccine; however, despite being attenuated, it may still capable of causing lesions in the bursa [37]. The newly emerged a typical (ayIBDV) strain that has unusual amino acid substitutions in the VP2 gene is also being documented [38–40]. Meanwhile, the serotype 2 isolates are usually isolated from turkeys and are apathogenic to both chickens and turkeys [18]. IBDV has also being classified based on its sequence characteristics such as the presence of certain restriction enzyme sites and unique amino acid residues in its VP2 gene [40–42].

The diversity of the IBDV strains had complicated the control and prevention of IBD, for example birds vaccinated against cvIBDV strain may not have adequate protection against other strains [43,44]. Therefore, analysis of the common genomic properties of the various IBDV strains will contribute greatly towards the understanding of the virus and the subsequent control and prevention efforts.

Although many sequence analyses papers had been published, base usage and dinucleotide frequency of IBDV remained unexplored. By studying the base usage, it was found that the genomic GC content of flaviviruses was associated with its vector specificity [45]. In thermophilic bacteria, high genomic GC content had been associated with the greater genomic stability (stronger bond of G–C pairs compared with A–T pairs) as a result of evolutionary adaptation to the hot environment [46]. And for human immunodeficiency virus (HIV) and other lentiviruses, unknown mechanisms had driven these viruses in having a strong bias for adenine nucleotide [47].

Non-random dinucleotide biases of the genome constitute a "general design" or genomic signature [48–51]. Genomic signature reflects the DNA properties in terms of its stacking energies, modification, replication, and repair mechanisms [51]. Moreover, genomic signature is useful for the detection of pathogenicity islands in bacterial genomes [51]. Generally, CpG (or 5′-CG -3′) and TpA dinucleotides are scarce [52–54]. CpG deficiency is typically associated with the classical methylation pathway, in which susceptible CpG dinucleotides will be methylated and subsequently converted to TpG [55]. TpA dinucleotides are unfavourable because the UA in mRNA is susceptible to Rnase activity [56]. Furthermore, avoiding TpA dinucleotides might reduce the occurrence of stop codons since two out of three stop codons are coded by TAA and TAG.

This paper had unveiled several fundamental characteristics of the IBDV genome. The base usage at each codon positions was described. The extracted information from the base usage was then utilized to investigate the origin of the overlapping VP5 gene. Comparison of the viral GC content with that of the host gave an insight into the virus–host interaction. The viral dinucleotide frequencies and their significance were also discussed.

Materials and Methods

IBDV Gene Sequences

All IBDV sequences (433 sequences), except the patented sequences, were downloaded from the GenBank release 131.0. Duplicated sequences, non-coding sequences, and sequences with unresolved/ambiguous sites were discarded. Sequences were then grouped into eight groups in reference to the different regions of the IBDV genome – namely VP1 (n = 25), VPX (36), VP2 (40), VP3 (34), VP4 (35), VP5 (28), polyprotein gene (33), and hypervariable region (HVR) (130) groups. Other sequences that cannot fit into the groups were excluded from the analysis. Selected sequences were edited and aligned by using BioEdit software version 5.0.9 [57] and ClustalX software [58]. Since most of the GenBank's IBDV entries did not clearly state that which strain (pathotype) the isolates belonged to, rather than merely based on molecular markers, strain identification was done manually by extensive literature search. Among the IBDV sequences in the GenBank, only few of the isolates had been completely sequenced; whereas the majority others were not. Since the grouping of the sequences was based on different regions of the IBDV genome, and since a fully sequenced isolate will cover all of the regions of the genome, then an isolate might simultaneously being included into different groups. Meanwhile for most isolates, only their HVRs were sequenced and therefore they only formed part of the "HVR group". There was 1 serotype 2 isolate in VP5 dataset whereas 2 isolates in all other datasets. In summary, regardless of the groupings, the nucleotide sequences of 131 IBDV isolates were analyzed. The accession numbers of these isolates were: AB024076, AF006694, AF006695, AF006696, AF006697, AF006698, AF006699, AF051837, AF051838, AF051839, AF076223, AF076224, AF076225, AF076226, AF076227, AF076228, AF076229, AF076230, AF076231, AF076232, AF076233, AF076234, AF076235, AF076236, AF083094, AF091097, AF091098, AF091099, AF109154, AF121256, AF133904, AF140705, AF155123, AF159207, AF159208, AF159209, AF159210, AF159211, AF159212, AF159214, AF159215, AF159216, AF159217, AF159218, AF165149, AF165150, AF165151, AF194428, AF240686, AF247006, AF260317, AF262030, AF279287, AF279288, AF281651, AF303219, AF321054, AF321055, AF321056, AF362747, AF362771, AF362773, AF362776, AF413069, AF413070, AF413071, AF413072, AF413073, AF413074, AF413075, AF413076, AF416620, AF416622, AF416623, AF416624, AF416625, AF416626, AF427103, AF454945, AF464901, AF498628, AF498629, AF498631, AF498632, AF498633, AF527039, AJ001941, AJ001942, AJ001943, AJ001944, AJ001945, AJ001948, AJ238647, AJ245885, AJ245886, AJ249517, AJ249519, AJ249520, AJ249523, AJ249524, AJ277801, AJ310185, AY029166, AY115569, AY115570, AY134874, D00499, D00867, D00868, D00869, D49706, D83985, L42284, M64285, M66722, M97346, X03993, X54858, X84034, X89570, X92760, X95883, Y14955, Y14956, Y14957, Y14962, Y14963, Y18612, Y18682, Z25481 and Z25482.

Host Coding Sequences and Highly Expressed Genes in the Bursa of Fabricius

Hosts (chicken and turkey) genomic coding sequences were obtained from Codon Usage Database (http://www.kazusa.or.jp/codon/) GenBank release 129.0. Bursal EST database [59] was referred to identify the highly expressed genes specifically found in the B-cells of the bursa of Fabricius. Since the database was constructed using a non-normalized cDNA library, the most frequently identified chicken (*Callus gallus*) genes will be the most abundantly (or highly) expressed genes in the bursa [59]. In addition, highly expressed bursal genes from other sources [60] were also included. Therefore the 28 highly expressed genes used in the analysis were ribosomal (16 sequences), heat shock (two sequences), elongation factor 1-α, β-actin, Ig rearranged light-chain VJC, chicken germ line Ig light chain, DEAD-box RNA helicase, non-histone chromosomal protein HMG-17, MHC B complex, ATF4, Bu-la, and chBl genes. All of the sequences were downloaded from GenBank and being meticulously edited, intron-excised, and analysed for the GC content.

Software Used in the Analyses

Base usage and dinucleotide frequencies were calculated by using CodonW 1.3 (software by John Penden and available at ftp://molbiol.ox.ac.uk/Win95.codonW.zip) and DAMBE version 4.0.98 (by Xuhua Xia and available at http://web.hku.hk/~xxia/software/Installation.htm). Both programs were used concurrently to ensure high reproducibility. Data editing and various analyses (correlation, cluster analysis, and principal component analysis) were done by using Microsoft Excel 2002, STATISTICA version 6, and SPSS version 11 software.

Base Usage and Dinucleotide Frequency Calculations

The overall base usage was calculated for each virus gene. In addition, base usage at the first (P1), second (P2), and third codon positions (P3) were also computed. Similarly, dinucleotide frequency was calculated for each of the reading frames (1:2, 2:3, 3:1) and as the overall measurement (at all codon positions). Dinucleotide index (DnI) was computed as the ratio of observed (O_d) to expected (E_d) dinucleotide frequencies:

$$\mathrm{DnI} = \frac{O_d}{E_d}$$

The expected frequency (E_d) of the dinucleotides at sites P1and P2 was calculated as

$$E_{d,\mathrm{P1,P2}} = \mathrm{p}(n_1) \times \mathrm{p}(n_2)$$

where $\mathrm{p}(n_1)$ and $\mathrm{p}(n_2)$ were the proportions of the nucleotides n_1 and n_2 at P1 and P2 respectively. If there was no dinucleotide bias, DnI value will be 1.

Results and Discussion

Base Usage of Serotype 1 IBDV Genes

Base usage or the relative distribution of each nucleotide (A, T, G, and C) at each codon positions was calculated for VP1, VPX, VP2, VP3, VP4, and VP5 genes. Subsequently, a rank of 1 (least frequently used) to 4 (most frequently used) was assigned to each nucleotide distribution in reference to its relative base usage percentage. The base usage patterns became pronounced after the shading (coloured as grey) of the higher ranks (rank 3 and 4) versus the lower ranks (rank 1 and 2) (non-coloured) as shown in Table 1.

Generally, base usage at each codon positions (P1, P2, and P3) would not be equal because the base usage of the coding sequences was not random. Moreover, base usage at P1 and P2 was constrained by the coding amino acids. Indeed, only 4% of P1 mutations were synonymous and all P2 mutations were non-synonymous [61]; these resulted in the inflexibility of the base usage at P1 and P2. However, the P3 was expected to have a more variable base usage because 69% of P3 mutations were silent [61].

Referring to Table 1, Thymine (T) was the least preferred nucleotide at P1. Considering all stop codons begin with T (TAA, TAA, and TGA), avoidance of T at P1 was understandable to prevent the unwanted occurrence of stop codon in the viral coding sequence (CDS). Except for VP5 gene, Guanine (G) was comparatively high at P1. This showed the inclination of IBDV to encode aliphatic amino acids (alanine, valine, and glycine). Intriguingly, the general base usage patterns at P1 were comparable for all IBDV genes.

At P2, all viral genes had the lowest G nucleotide except VP5; which had the lowest T nucleotide. Deficiency of G at P2 might attribute to the virus' efforts to prevent the occurrence of stop codon. Unlike P1, base usage at P2 was more varied because any P2's mutation will alter the encoded amino acid. In this case, maintaining the physiochemical properties of the virus proteins, most probably by evolutionary forces, would be more important than maintaining a similar base usage.

At P3, all viral genes were devoid of T, excluding VP4 and VP5 genes. In addition, C

Table 1. Base usage percentage and the ranking in the serotype 1 IBDV genes (mean ± SD)

Gene	Position	T	C	A	G
VP1	P1	14.7 ± 0.3	23.6 ± 0.2	30.0 ± 0.2	31.7 ± 0.2
VPX	P1	13.3 ± 0.2	20.4 ± 0.3	31.5 ± 0.3	34.8 ± 0.3
VP2	P1	13.6 ± 0.2	19.9 ± 0.2	33.5 ± 0.3	33.0 ± 0.3
VP3	P1	11.0 ± 0.3	26.0 ± 0.2	27.0 ± 0.4	36.0 ± 0.5
VP4	P1	12.1 ± 0.3	25.5 ± 0.4	25.8 ± 0.3	36.5 ± 0.3
VP5	P1	17.4 ± 0.6	32.4 ± 0.7	21.4 ± 0.5	28.8 ± 0.4
VP1	P2	26.3 ± 0.1	24.8 ± 0.1	32.0 ± 0.2	16.9 ± 0.2
VPX	P2	29.4 ± 0.3	28.3 ± 0.3	23.3 ± 0.2	18.9 ± 0.2
VP2	P2	30.3 ± 0.3	26.4 ± 0.4	24.4 ± 0.2	18.9 ± 0.2
VP3	P2	20.1 ± 0.4	28.1 ± 0.4	34.2 ± 0.3	17.6 ± 0.3
VP4	P2	30.0 ± 0.2	25.5 ± 0.1	26.3 ± 0.3	18.1 ± 0.3
VP5	P2	15.2 ± 0.2	24.8 ± 0.1	34.3 ± 0.4	25.6 ± 0.3
VP1	P3	16.4 ± 1.0	32.2 ± 0.4	23.5 ± 1.0	27.9 ± 0.3
VPX	P3	18.9 ± 0.7	33.2 ± 0.6	24.5 ± 0.7	23.3 ± 0.6
VP2	P3	18.6 ± 0.8	34.3 ± 0.6	23.8 ± 0.7	23.3 ± 0.6
VP3	P3	13.5 ± 0.7	29.2 ± 1.2	28.4 ± 1.7	28.9 ± 0.7
VP4	P3	22.4 ± 1.3	35.3 ± 1.7	23.1 ± 1.4	19.2 ± 1.2
VP5	P3	29.5 ± 0.4	27.1 ± 0.3	24.0 ± 0.2	19.4 ± 0.2
VP1	Total	19.1 ± 0.3	26.9 ± 0.2	28.5 ± 0.3	25.5 ± 0.1
VPX	Total	20.6 ± 0.3	27.3 ± 0.2	26.4 ± 0.3	25.7 ± 0.2
VP2	Total	20.8 ± 0.3	26.9 ± 0.2	27.2 ± 0.3	25.1 ± 0.2
VP3	Total	14.9 ± 0.3	27.7 ± 0.4	29.9 ± 0.6	27.5 ± 0.3
VP4	Total	21.5 ± 0.5	28.8 ± 0.6	25.1 ± 0.4	24.6 ± 0.4
VP5	Total	20.7 ± 0.2	28.1 ± 0.2	26.6 ± 0.2	24.6 ± 0.2

Note. Codon positions were represented as P1, P2, P3, and total (at all codon positions). The numbers of sequences used in the calculation, excluding the serotype 2 isolates, were: 23 (VP1), 34 (VPX), 38 (VP2), 32 (VP3), 33 (VP4), and 27 (VP5). A rank of 1 (least frequently used) to 4 (most frequently used) was assigned to each nucleotide distribution depending on its relative percentage. Shades areas (grey) denoted the upper ranks (rank 3 and 4) while the non-shaded areas denoted lower ranks (rank 1 and 2). Note that the overall base usage patterns (Total) for all IBDV genes were similar, where C and A nucleotides dominated whereas T was unfavourable.

(cytosine) appeared to be the preferred nucleotide. The bias towards C was an interesting feature because most P3's mutations were silent [61]. Base usage bias at P3 might confer certain selective advantageous to IBDV; perhaps by having the bias, the virus would be able to match up its codon usage with the host. If so, the virus may improve its translational efficiency and this may lead to increased fitness. Meanwhile, it was suggested that favouring of C at P3 would increase the coding ability or new ORF formation, considering none of the stop codons contain C nucleotide [62]. However, the dearth of G nucleotide in VP4 gene remained to be investigated.

Unexpectedly, the overall (total) base usage of all IBDV genes was similar, despite some discrepancies at each codon positions. Moreover, although being physically separated, the VP1 still resembled other genes. It was also found that C and A (adenine) were the most preferred nucleotides, whereas T was the least preferred. Given that RNA virus had high mutation rate [63] and short generation time, why did the virus maintain a similar base usage pattern for all its genes? Perhaps this could be the virus strategy to optimise its genes expression. It had been shown that virus could take the advantage of the codon composition to regulate its own programs of gene expression [64] while utilizing the cellular machinery to replicate its genome.

Base usage of the serotype 2 genome was separately analyzed because only two isolates (OH and 23/82) were available from the GenBank 131.0. Results indicated that the serotype 2's base usage was comparable to serotype 1's (data not shown). As in serotype 1, serotype 2's VP5 gene had peculiar base usage pattern.

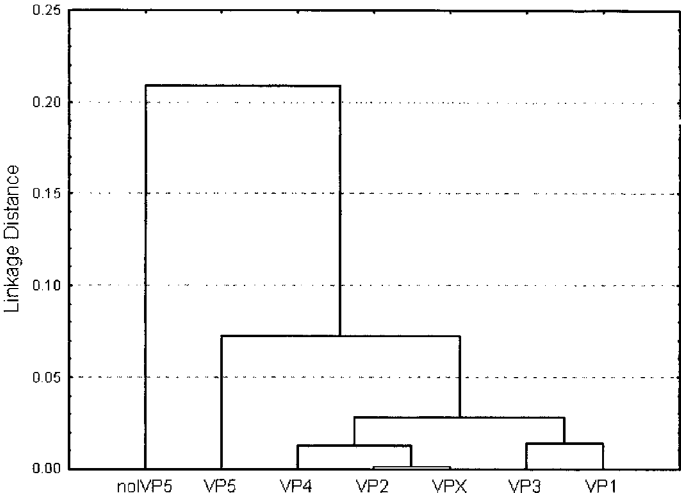

Fig. 1. Results of the cluster analysis of serotype 1 IBDV genes based on nucleotide composition. Note that the VP5 gene and its non-overlapping region (nolVP5) formed different clusters from other viral genes.

Peculiarity of the Base Usage Pattern in the VP5 gene

VP5 gene's peculiar base usage could be partly explained by its overlapping region within the VPX. Indeed, 92.24% (404/438 bp) of the VP5 gene was embedded within VPX. Consequently in the overlapping region, VPX's P3 became VPS's P1. Thus, this explained why VP5's P1 was high in C, P2 was low in T, and P3 was rich in T. Further analysis of the VP5's non-overlapping region (nolVPS) (11 codons, 34 bp) revealed that although its P3 was also rich in C (>30%), it was richest in T (33.7%); which was differed from other genes. These findings were in agreement with the previous report [62] where overlapping genes showed significant bias in their base usage.

To study the relationships among the virus genes, cluster analysis was performed on the virus genes' nucleotide compositions. The virus genes were treated as the 'columns' (seven columns: VP1, VPX, VP2, VP3, VP4, VP5, and nolVP5) and the nucleotide compositions (presented as mean percentages) at each codon positions were treated as the 'attributes' in Q-type cluster analysis. Since there were three different codon positions and four types of nucleotides, therefore there were 12 attributes: for example, the percentage of adenine at P1, the percentage of guanine at P1... the percentage of cytosine in P3, and so forth. Squared Euclidean distances were then computed and a tree was constructed using unweighted pair-group average (UPGMA) amalgamation rule (Fig. 1). Cutting the tree at 0.05 linkage distance, it was clear that VP5 gene and its non-overlapping region formed different clusters compared with other viral genes. This led us to suspect that VP5 gene's peculiar base usage was due to its origin; where most likely it was originated by overprinting the 'original' (or existing) viral genes.

To generate a novel gene, the virus may either need to synthesize an entirely new nucleotide sequences or alternatively, it may utilize the unused reading frames of the existing genes, a process first proposed by Grasse [65], who called it "overprinting" [65]. In tymoviruses, overlapping gene arose by overprinting the "original" replicase gene after the virus had diverged from its sister groups from a common ancestor [66]. In the *Birnaviridae* family, VP5 gene was found only in *Avibirnavirus* (IBDV) and *Aquabirnavirus* (infectious pancreatic necrosis virus or IPNV). The other genus, *Entomobirnavirus* (Drosophila X virus or DXV) had no equivalent ORF to overlap at the 5′ terminus of VPX [67]. For DXV, the predicted overlapping non-structural protein (believed to be a VP5 homolog) resides in between VP4 and VP3 genes. With regard to the birnaviruses evolution, the most parsimonious explanation appeared to be the polyprotein gene was the birnaviruses' "original gene" and VP5 gene arose after the vertebrate birnaviruses (IBDV and IPNV) and the insect

Table 2. Hosts and IBDV genomic GC content comparisons (means)

	Overall GC%	GC% at various codon positions			
		P1	P2	P3	N[a]
Chicken coding sequences	52.2	55.1	41.6	59.8	2170
Chicken highly expressed genes[b]	52.3	54.0	40.2	62.7	28
Overall IBDV serotype 1 genome	53.1	57.0	44.0	58.3	16
IBDV Segment A[c]	53.6	58.5	46.0	56.3	23
IBDV Segment B[c]	52.4	55.3	41.8	60.2	23
IBDV Hypervariable region[c]	52.2	54.3	47.5	54.9	128
Turkey coding sequences	50.9	52.3	39.8	60.7	46
Overall serotype 2 genome	53.7	57.3	44.2	59.5	5

[a] Number of sequences used in the analysis.
[b] Highly expressed genes in the chicken B-cell.
[c] Coding sequences of serotype 1 strains.

birnavirus (DXV) had diverged from their common ancestor. It was unlikely for DXV to initially possess VP5 gene, to lose it subsequently after the divergence, and to create another new ORF in order to replace the lost gene's function.

Due to the frame shift of overprinting gene, the gene will have an unusual codon usage and encodes new protein with physiochemically-biased properties [62]. VP5 protein had been shown to play a role in IBDV pathogenesis [22] and in the release of viral progeny from infected cells [24]. VP5-defective virus had exhibited a slight delay in replication [22]; but the VP5 gene was inessential for the virus *in vitro* [23] and *in vivo* replication [22,68]. Simply put, the acquaintance of VP5 gene as a "new gene" by overprinting strategy in birnaviruses evolutionary history, although inessential, may give the virus certain survival advantages to retain the VP5 gene in its genome.

GC Content of IBDV Genome is Very Similar to the Host

GC content (GC%) for many double-stranded DNA (dsDNA) viruses differed markedly from the GC content of the host cells they infected [69]. To investigate if the same phenomenon applies to IBDV (dsRNA), we compared the virus' GC% with the host (Table 2).

Results showed the overall GC% of IBDV genome was comparable to the chicken (*Gallus gallus*), in which it was around 52–53%. Interestingly, in spite of high mutation rate of the hypervariable region, its GC% nearly matched the host highly expressed genes' GC%. Similarly, segment B's GC% was very close to the chicken highly expressed genes' GC%. Meanwhile, serotype 2's GC% was differed more to turkey (*Meleagris gallopavo*) than to chicken, although serotype 2 isolates usually isolated from turkey. The reason for this discrepancy remained to be answered.

A general pattern of GC% for both virus and host was observed: high GC% in P1, low in P2, and high in P3. These findings would suggest the virus attempt in mimicking the host GC%, particularly P3 GC%, probably in order to optimise its codon usage for translational efficiency and continue to thrive as a successful intracellular parasite. In contrast to the dsDNA virus, GC% of the IBDV and the host was comparable.

IBDV Genome is Devoid of CpG and TpA Dinucleotides

Apart from the CpG islands in mammalian genome, CpG dinucleotides were usually under-represented because of two main reasons. First, the classical methylation pathway that converts CpG to TpG [55]. The pathway works by methylating the 5′ cytosine of CpG and subsequently deaminates the 5-methylcytosine leading to the mutation of CpG and convert to TpG [55]. Second, CpG dinucleotides exhibit the greatest thermodynamic stacking energy of all dinucleotides [70,71]; therefore, reducing its frequency might facilitate nucleic acids replication and transcription [72]. Thus, it will be interesting to investigate if IBDV genome was also devoid of CpG dinucleotides.

To study the IBDV's dinucleotide frequencies, three datasets were analysed, namely the polyprotein gene (VPX–VP4–VP3), hypervariable region, and segment B sequences. The VP5 gene was excluded because it was highly conserved (14/28 isolates have identical sequences) and most of its nucleotides were embedded within the VPX gene. The null hypothesis in this study was that there was no selective pressure against CpG dinucleotides or meaning that all dinucleotides pairs had equal chance of occurrence with the reference to the base composition. The Mann–Whitney U test was used to demonstrate if CpG dinucleotides had significantly deviated from the expected proportion.

Results from Table 1 showed that P3 and P1 were highest in C and G, respectively. Thus, if there was no dinucleotide bias, one would expect high CpG dinucleotides at the intercodon position (P3:P1). However, results from the analysis of the three datasets showed that the dinucleotide bias did occur where the expected intercodon CpG dinucleotides were significantly lower than the observed ($p < 0.01$). This succinctly showed the avoidance of CpG dinucleotides in IBDV genome. This finding was in accordance with Karlin et al. [73] where virtually all small eukaryotic viruses were deficient in CpG dinucleotides. Meanwhile, TpG intercodon dinucleotide frequency was significantly higher than the expected ($p < 0.01$).

Further analysis of the dinucleotide frequency at all possible codon positions gave the same results where the CpG was lower and TpG was higher than expected. Moreover, TpA dinucleotides were also found to be lower than the expected ($p < 0.01$). The dearth of TpA could be due to the susceptibility of UA in mRNA to RNase activity [56] (but see [74]). TpA was also less energically stable than all other dinucleotides [70,71], which rendered the nucleic acids to be more flexible in bending and untwisting. This explained why TATA sequences at the sites of replication origin were very easy to unwind and interact with other molecules [75]. Hence, the restriction of TpA dinucleotides may help in avoiding inappropriate binding of cellular factors to the viral nucleic acids. Furthermore, given the fact that two out of three stop codons have TpA dinucleotides, reducing the genomic TpA dinucleotides would certainly help in avoiding the occurrence of unwanted mutation-derived stop codons.

Deficiency of CpG is Correlated with the Abundance of TpG

The relationship between CpG and TpG dinucleotides were studied further by using correlation. For each dinucleotide pairs, the value of dinucleotide index (DnI) was calculated as the ratio of observed dinucleotides versus the expected dinucleotides. Results indicated that the number of CpG dinucleotides was negatively correlated with TpG dinucleotides. The *R*-values for segment B and polyprotein dataset were −0.803 and −0.815 ($p < 0.0001$), respectively. Correlation for HVR dataset ($R = -0.406$, $p < 0.0001$) was however weaker; probably due to its shorter sequence. We were fully aware that correlation did not imply causation, but based on the fact and our empirical results, we concluded that the deficiency of CpG probably contributed to the abundance of TpG in the IBDV genome through the conversion of methylated CpG to TpG [55].

The vertebrate immune system had apparently evolved the ability to recognize the unmethylated-CpG motifs and responds with a rapid and coordinated cytokine response leading to the induction of humoral and cell-mediated immunity [76,77]. Moreover, CpG-based adjuvant had shown to trigger protective antiviral cytotoxic T cell responses [78]. Therefore, we proposed that by avoiding the CpG dinucleotides, IBDV might be able to minimize its antigenicity and avoid undesirable host immune response. On a different perspective, we suggested the use of CpG-based adjuvant in IBD killed vaccine; considering the virus attempts in avoiding CpG dinucleotides. It had been shown that CpG oligonucleotides could be a valuable adjuvant for poultry vaccines [79]. Thus, the potential usage of CpG-based adjuvant in IBD killed vaccine may be the future research interest.

Dinucleotide Patterns are Different Among IBDV Strains

Classifying IBDV strains was indispensable for the control and prevention of IBD. Apart from pathological and serological classification, IBDV had been grouped by its sequence characteristics [40,42]; where each IBDV strains had its own characteristic restriction enzymes sites [41] and molecular markers [40]. IBDV dinucleotide usage (or dinucleotide patterns) was however unknown, despite many sequence analysis papers on IBDV genome had been published. In coronaviruses, analysis of dinucleotide frequency had separated the virus into two groups that roughly reflect its taxonomic origins [80]. Thus, the current study was to investigate if dinucleotide patterns differed among the IBDV strains and the practicality of "principal component analysis of the dinucleotide frequencies" (DF-PCA) approach in studying the IBDV dinucleotide patterns.

DnI was calculated for each of the 16-types of dinucleotide pairs. Since DnI was a relative measure of dinucleotide frequency, PCA rather than the correspondence analysis was used in the analysis [81]. The concepts and principles of PCA have been extensively described in most multivariate analysis textbooks, so it will not be discussed here. All the datasets (hypervariable region, polyprotein and segment B) were analysed by the DF-PCA approach. For hypervariable region and polyprotein datasets, three outliers namely the Australian cvIBDV (00/273) and serotypes 2 (OH, 23/82) isolates were excluded because of their unique sequence characteristics.

Results of DF-PCA were depicted as a graph plot in which the axes represent the amount of "extracted variation" (Fig. 2). In Fig. 2A, the first two axes accounted for 52% (35.84% + 15.34%) of the total variance, or in other words it explained 52% of the total variation observed from the dinucleotide patterns of the hypervariable region. Noticeably, there were two distinct groups separated along the first axis: a very virulent group on the left and attenuated group on the right. Other strains were remained in between the two major groups. There was no clear separation between classical and attenuated strain. This probably because many attenuated isolates originated from the attenuation of classical isolates. The bold capital V and A were OKYM (vvIBDV) and OKYMT (attenuated form of OKYM) isolates respectively [82]. Interestingly, it appeared to be a subtle "right-shift" of OKYM towards the attenuated strains after the attenuation process, but not to the extent of total separation from the vvIBDV cluster. While the impact of the attenuation on the IBDV's dinucleotide patterns remained to be investigated, the inability of OKYMT to be within the atlBDV cluster reflected that DF-PCA was in fact influenced by the virus evolutionary relationship. However, there was no evidence that IBDV isolates situated on the extreme left will be the "most virulent" vvIBDV and the extreme right isolate will be the "most attenuated" atIBDV.

Incorrectly classified isolates could be quickly detected on the DF-PCA graph due to their odd positions. It was found that the classifications of ZJ2000 (GenBanK accession no. AF321056) and GZ902 (AF006699) isolates were inappropriate. ZJ2000 was reported as a highly virulent IBDV [83] but its position in the graph (Fig. 2A and 2B) seemed to be related more to the attenuated or classical strain than to the vvIBDV strain. To examine this problem closely, sequence analysis for ZJ2000 was done. It was found that none of the important vvIBDV markers (242Ile, 256Ile, and 294Ile) [40] and serine-rich heptapeptide virulent marker "SWSASGS" [84] were present in ZJ2000. In addition, ZJ2000 had 253His and 284Thr that were closely related with the attenuated strain than to the virulent strain [40]. For the GZ902 ("variant strain"), its hypervariable region sequence was found to be identical with another attenuated strain GZ29112 (AF051837) and located exactly at the same position in the map (circle in Fig. 2A was the location for both GZ902 and GZ29112). Sequence analysis on both isolates found that GZ29112 was grouped correctly whereas GZ902 should be grouped as the attenuated strain by referring to the molecular markers.

Fig. 2B and C showed the DF-PCA results for polyprotein and segment B datasets. The first two axes of polyprotein and segment B datasets explained about 57% and 66% of total variation, respectively. We found that DF-PCA on hypervariable region sequences could yield comparable result as the longer polyprotein gene sequences. This probably because the virulence molecular determinants, cell tropism, and pathogenic phenotype of IBDV all fall within the hypervariable region [85]. Meanwhile, atypical isolates (UPM94/273 and K310) were located closely with the vvIBDV isolates as shown in Fig. 2B. This was

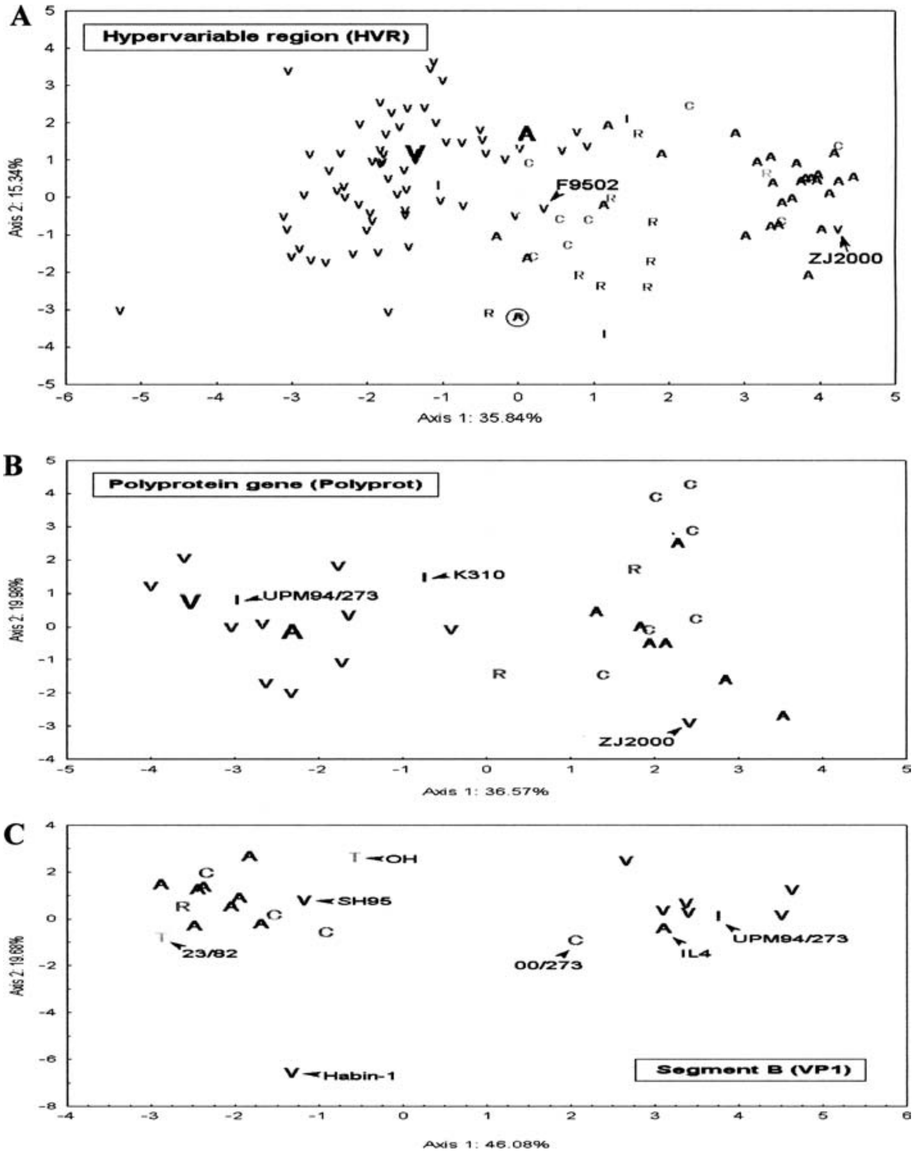

Fig. 2. Results from the DF-PCA of IBDV genome. (A) Hypervariable region. (B) Polyprotein gene (VPX-VP4-VP3). (C) Segment B (VPI gene). V = very virulent strain, I = atypical strain, C = classical strain, R = variant strain, A = attenuated strain, T = serotype *2* strain. The bold capitals V and A denoting OKYM (very virulent IBDV) and OKYMT (attenuated form of OKYM) isolates, respectively.

understandable because atypical strain was considered as a subset of vvIBDV strain [86].

VP1 gene had an intricate dinucleotide pattern (Fig. 2C) where different IBDV serotypes and strains were intermingled with each other on the graph. Intriguingly, rather than forming an isolated cluster, serotype 2 isolates (OH and 23/82) located near the cvIBDV and atIBDV isolates. In addition, the vvIBDV isolates (SH95 and Habin-1), cvIBDV isolate (00/273), and atIBDV (IL4) had unique dinucleotide patterns whereby they did not belong to any significant cluster. These findings disagreed with Islam et al. [87] where vvIBDV's VP1 gene distinctly separated from other strains. Perhaps this was because the number of sequences used in this study (n = 25) was larger compared with Islam's (n = 18). New vvIBDV isolates such as SH95 (AY134875) and Habin-1 (AF455136) were not included in the previous study. Furthermore, DF-PCA approach was differed from the phylogenetic approach because DF-PCA analysed the inter-relationships of the 16-dinucleotide pairs, whereas the phylogeny method (specifically distance method) calculated evolutionary distances based on a chosen substitution (or evolutionary) model. The substitution model chosen by Islam and co-workers in the construction of their VP1-phylogenetic tree was however not stated in their report. In a different viewpoint, it should be remembered that IBDV is a bisegmented virus and whether the bewildering dinucleotide patterns of VP1 gene were due to the inter-strains gene reassortment remained to be investigated.

The use of DF-PCA was unintended to be a substitute for the current strains classification methods, even though it was granted with some abilities in grouping the IBDV strains. In this study we used DF-PCA to demonstrate the unique characteristics of each IBDV strains by its dinucleotide frequency. DF-PCA analyzed the delicate inter-relationships among the dinucleotide pairs and visually projected the results in a form of graph or "map". The result from the DF-PCA analysis was not solely dependent on the sequence's identity percentage, albeit this was an important factor. For example, although OKYM shared a 92.9% of sequence identity with both F9502 and ZJ2000 isolates, ZJ2000 was located far away from OKYM in comparison with F9502 (see Fig. 2A).

Although many underlying biological properties of DF-PCA remained to be investigated, we believed that the results of DF-PCA reflected the evolutionary history of the virus considering each dinucleotide pairs were influenced by the evolutionary forces (and thus constituted the genomic signature). In phylogenetic analysis, particularly clustering algorithm, evolutionary relationships were studied by grouping the taxa into various groups or clades. And with regard to IBDV, these clades usually reflect the strain of the virus; for example, very virulent isolates are grouped together but not with the variant isolates.Therefore, a taxon must either be in or out from a given clade. In contrary, by using DF-PCA, the inter-relationships among the IBDV isolates were visually displayed as "points" on the graph rather than forming the distinct clusters. Thus, DF-PCA allowed the shades of grey and may promote further insight into the virus evolutionary history.

The virus genome is packed with information and it means everything for the virus survival. In this study we had uncovered many genomic properties of IBDV by analysing its base usage and dinucleotide frequency. We envisaged that similar approach could be adopted to study other viruses' genes to the understanding of the fundamental properties of the viruses.

Acknowledgments

This work was generously supported by IRPA Grant 54091 from the Malaysian government.

References

1. Saif Y.M., Vet Immunol Immunopathol *30*, 45–50, 1991.
2. Muller H., Schnitzler D., Bernstein F., Becht H., Cornelissen D., and Lutticken D.H., Vet Microbiol *33*, 175–183, 1992.
3. Nagarajan M.M., and Kibenge F.S., Can J Vet Res *61*, 81–88, 1997.
4. Sharma J.M., Kim I.J., Rautenschlein S., and Yeh H.Y., Dev Comp Immunol *24,* 223–235, 2000.
5. van den Berg T.P., Eterradossi N., Toquin D., and Meulemans G., Rev Sci Tech *19*, 509–543, 2000.
6. Dobos P., Hill B.J., Hallett R., Kells D.T., Becht H., and Teninges D., J Virol *32*, 593–605, 1979.
7. Muller H., Scholtissek C., and Becht H., J Virol *31*, 584–589, 1979.
8. Leong J.C., Brown D., Dobos P., Kibenge F.S.B., Ludert J.E., Muller H., Mundt E., and Nicholson B., in van Regenmortel M.H.V., Fauget C.M., Bishop D.H.L., Carstens E.B., Estes M.K., Lemon S.M., Maniloff J., Mayo A., McGeoch D.J., Pringle C.R., and Wickner R.B. (ed), Virus Taxonomy: Seventh Report of the International Committee on the Taxonomy of Viruses, Academic Press, 2000, pp. 481–490.
9. Bottcher B., Kiselev N.A., Stel'Mashchuk V.Y., Perevozchikova N.A., Borisov A.V., and Crowther R.A., J Virol *71*, 325–330, 1997.
10. Mundt E., and Muller H., Virology *209*, 10–18, 1995.
11. Bayliss C.D., Spies U., Shaw K., Peters R.W., Papageorgiou A., Muller H., and Boursnell M.E., J Gen Virol *71*, 1303–1312. 1990.
12. Kibenge F.S., Jackwood D.J., and Mercado C.C., J Gen Virol *71*, 569–577, 1990.
13. Birghan C., Mundt E., and Gorbalenya A., EMBO J *19*, 114–123, 2000.

14. Kibenge F.S., Qian B., Cleghorn J.R., and Martin C.K., Arch Virol *142*, 2401–2419, 1997.
15. Chevalier C., Lepault J., Erk I., Da Costa B., and Delmas B., J Virol *76*, 2384–2392, 2002.
16. Fahey K.J., Erny K., and Crooks, J., J Gen Virol *70*, 1473–1481, 1989.
17. Heine H.G., Haritou M., Failla P., Fahey K., and Azad A., J Gen Virol *72*, 1835–1843, 1991.
18. Becht H., Muller H., and Muller H.K., J Gen Virol *69*, 631–640, 1988.
19. Oppling V., Muller H., and Becht H., Arch Virol *119*, 211–223, 1991.
20. Kochan G., Gonzalez D., and Rodriguez J.F., Arch Virol *148*, 723–744, 2003.
21. Mundt E., Beyer J., and Muller H., J Gen Virol *76*, 437–443, 1995.
22. Yao K., Goodwin M.A., and Vakharia V.N., J Virol *72*, 2647–2654, 1998.
23. Mundt E., Kollner B., and Kretzschmar D., J Virol *71*, 5647–5651, 1997.
24. Lombardo E., Maraver A., Espinosa L., Fernandez-Arias A., and Rodriguez J.F., Virology *277*, 345–357, 2000.
25. Muller H., and Nitschke R., Virology *159*, 174–177, 1987.
26. Spies U., Muller H., and Becht H., Virus Res *8*, 127–140, 1987.
27. Morgan M.M., Macreadie I.G., Harley V.R., Hudson P.J., and Azad A.A., Virology *163*, 240–242, 1988.
28. Spies U., and Muller H., J Gen Virol *71*, 977–981, 1990.
29. Shwed P.S., Dobos P., Cameron L.A., Vakharia V.N., and Duncan R., Virology *296*, 241–250, 2002.
30. Maraver A., Clemente R., Rodriguez J.F., and Lombardo E., J Virol *77*, 2459–2468, 2003.
31. McFerran J.B., MeNulty M., Killop E.R., Corner T.J., McCracken R.M., Collins P.S., and Allan G.M., Avian Pathol *9*, 395–404, 1980.
32. Oppling V., Muller H., and Becht H., Arch Virol *119*, 211–223, 1991.
33. Nunoya T., Otaki T., Tajuma M., Hirag M., and Saito T., Avian Dis *36*, 597–609, 1992.
34. van den Berg T.P., Gonze M., and Meulemans G., Avian Pathol *20*, 133–143, 1991.
35. Lasher H.N., and Shane S.M., World Poult Sci J *50*, 133–166, 1994.
36. Rosenberger J.K., and Cloud S.S., J Am Vet Med Assoc *189*, 357, 1986.
37. van den Berg T.P., Avian Pathol *29*, 175–193, 2000.
38. Kwon H.M., Kim D.K., Hahn T.W., Han J.H., and Jackwood D.J., Avian Dis *44*, 691–696, 2000.
39. Kong L.L., Omar A.R., Hair-Bejo M., and Aini I., 2nd International Congress/13th VAM congress and CVA-Australasia/Oceania Regional Symposium; 2001 Aug 27–30; Kuala Lumpur; Vet Assoc Malaysia; 2001, pp. 77–79.
40. Rudd M.F., Heine H.G., Sapats S.I., Parede L., and Ignjatovic J., Arch Virol *147*, 1303–1322, 2002.
41. Sapats S.I., and Ignjatovic J., Avian Pathol *31*, 559–566, 2002.
42. Kim T.K., and Yeo S.G., Virus Genes *26*, 97–106, 2003.
43. Chettle N., Stuart J.C., and Wyeth P.J., Vet Rec *125*, 271–272, 1989.
44. Synder D.B., Lana D.P., Savage P.K., Yancey F.S., Mengel S.A., and Marquardt W.W., Avian Dis *32*, 535–539, 1988.
45. Jenkins G.M., Pagel M., Gould E.A., Zanotto P.M.A., and Holmes B.C., J Mol Evol *52*, 383–390, 2001.
46. Argos P., Rossmann M.G., Grau U.M., Zuber A., Franck G., and Tratschin J.D., Biochemistry *18*, 5698–5703, 1979.
47. Hemert F.J., and Berkhout B., J Mol Evol *41*, 132–140, 1995.
48. Russell G.J., Walker P.M.B., Elton R.A., and Subak-Sharpe J.H., J Mol Biol *108*, 1–23, 1976.
49. Viotti A., Balducci C., and Wdil J.H., Biochim Biophys Acta *517*, 125–132, 1978.
50. Karlin S., Mrazek J., and Campbell A.M., J Bacteriol *179*, 3899–3913, 1997.
51. Karlin S., Curr Opin Microbiol *1*, 598–610, 1998.
52. De Amicis F., and Marchetti S., Nucleic Acids Res *28*, 3339–3345, 2000.
53. Montero L.M., Salinas J., Matassi G., and Bernardi G., Nucleic Acids Res *18*, 1859–1867, 1990.
54. Shioiri C., and Takahata N., J Mol Evol *53*, 364–376, 2001.
55. Bird A.P., Nucleic Acids Res *8*, 1499–1504, 1980.
56. Beutler E., Gelbart T., Ran J.H., Koziol, J.A., and Beutler B., Proc Natl Acad Sci USA *86*, 192–196, 1989.
57. Hall T.A., Nucl Acids Symp Ser. *41*, 95–98, 1999.
58. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., and Higgins D.G., Nucleic Acids Res *24*, 4876–4882, 1997.
59. Abdrakhmanov I., Lodygin D., Geroth P., Arakawa H., Law A., Plachy J., Korn B., and Buerstedde J.M., Genome Res *10*, 2062–2069, 2000.
60. Goitsuka R., Mamada H., Kitamura D., Cooper M.D., and Chen C.I., J Immunol *167*, 1454–1460, 2001.
61. Li W.H., and Graur D. (ed), Fundamental of Molecular Evolution, Sinauer Associates, Sunderland, 1991, p. 15.
62. Keese P.K., and Gibbs A., Proc Natl Acad Sci USA *89*, 9489–9493, 1992.
63. Drake J.W., and Holland J.J., Proc Natl Acad Sci USA *96*, 13910–13913, 1999.
64. Zhou J., Liu W.J., Peng S.W., Sun X.Y., and Frazer I., J Virol *73*, 4972–4982, 1999.
65. Grasse P.P. (ed), Evolution of Living Organisms, Academic Press, New York, 1977, p. 297.
66. Keese P., Mackerzie A., and Gibbs A., Virology *172*, 536–546, 1989.
67. Chung H.K., Kordyban S., Cameron L., and Dobos P., Virology *225*, 359–368, 1996.
68. Weber S., Fichtner D., Mettenleiter T.C., and Mundt E., J Gen Virol *82*, 805–812, 2001.
69. Strauss E.G., Strauss J.H., and Levine A.J., Virology. Lippincott-Raven, Philadelphia, 1995, pp. 153–171.
70. Breslauer K.J., Frank R., Blocker H., and Marky L.A., Proc Natl Acad Sci USA *83*, 3746–3750, 1986.
71. Delcourt S.G, and Blake R.D., J Biol Chem *266*, 15160–15169, 1991.
72. Karlin S., and Burge C., Trends Genet *11*, 283–290, 1995.
73. Karlin S., Doerfler W., and Cardon L.R., J Virol *68*, 2889–2897, 1994.

74. Duret L., and Galtier N., Mol Biol Evol *17*, 1620–1625, 2000.
75. Hunter C.A., J Mol Biol *230*, 1025–1054, 1993.
76. Krieg A.M., Trends Microbiol *4*, 73–76, 1996.
77. Krieg A.M., Yi A.K., Schorr J., and Davis H.L., Trends Microbiol *6*, 23–27, 1998.
78. Vabulus R.M., Pircher H., Lipford G.B., Hacker H., and Wagner H., J Immunol *164*, 2372–2378, 2000.
79. Vleugels B., Ververken C., and Goddeeris B.M., Poult Sci *81*, 1317–1321, 2002.
80. Tobler K., and Ackermann M., Adv Exp Med Biol *400*, 801–804, 1998.
81. Perriere G., and Thioulouse J., Nucleic Acids Res *30*, 4548–4555, 2002.
82. Yamaguchi T., Ogawa M., Inoshima Y., Miyoshi M., Fukushi H., and Hirai K., Virology *223*, 219–223, 1996.
83. Yu L., Li J.R., Huang Y.W., Dikki J., and Deng R., Avian Dis *45*, 862–874, 2001.
84. Cao Y.C., Yeung W.S., Law M., Bi Y.Z., Leung F.C., and Lim B.I., Avian Dis *42*, 340–351, 1998.
85. Brandt M., Yao K., Liu M., Heckert R.A., and Vakharia V.N., J Virol *75*, 11974–11982, 2001.
86. Eterradossi N., Arnauld C., Toquin D., and Rivallan G., Arch Virol *143*, 1627–1636, 1998.
87. Islam M.R., Zierenberg K., and Muller H., Arch Virol *146*, 2481–2492, 2001.