Probability of Z-DNA Forming Subsequences and B-Z Transition Sites in the Epstein-Barr Virus and Others

JAMES R. BOZEMAN[1], JESUS E. GARCIA[2], V. A. GONZALEZ-LOPEZ[2]
[1]American University of Malta,
Bormla BML 1013,
MALTA

[2]Department of Statistics, University of Campinas,
Sergio Buarque de Holanda, 651, Campinas, S.P., CEP: 13083-859,
BRAZIL

Abstract: - It is well known what sequences of bases in DNA are potentially in the left-handed form, i.e. prone to be in the Z-DNA conformation. In particular, $d(GC)n > d(CA)n > d(CGGG)n > d(AT)n$. These Z-DNA forming sequences (ZFS) have been found in the full DNA sequences of SARS CoV-2, rodent parvoviruses, salmonella, and some carcinogens. During a Short Term Scientific Mission (STSM, see the Introduction) we examined the stochastic profiles of such sequences in the FASTA format to determine the probability of these occurring. In the sequences studied we found CA more prevalent than GC, and also AT more common than CGGG. Novelly, we found such sequences in the Epstein-Barr virus (EBV), which to the best of our knowledge had not been thoroughly checked previously for ZFS, and we calculated the probabilities of those subsequences occurring. Note that in the EBV case, GC pairs were more prevalent than CA pairs. We also checked Dengue and HIV for potential ZFS and found many potential sites in HV8. Finally, we present our current work, including implications for the 3-D conformation of the DNA molecule and applying the idea of microsatellites to the repeated sequences known to be in left-handed form, especially since these inform an analysis of the possible transition sites from the B-form of DNA to the Z-form, and vice versa. These transitions at CG pairs have been the most studied, but the flips seem to happen more at TG sites. AT pairs are also possible. As above, we find the probabilities of potential B-Z transitions at these locations using our methodology. Our results can help researchers hone in on the regions in genomes where Z-DNA formation is likely.

Key-Words: - Z-DNA, B-Z transition, Epstein-Barr virus, HIV, Markov models, Microsatellites, 3-D Conformation.

Received: April 23, 2024. Revised: November 24, 2024. Accepted: December 13, 2024. Published: February 7, 2025.

1 Introduction

In this paper, we report on results from a Short Term Scientific Mission (STSM) done under the auspices of the EU-COST Action grant EUTOPIA (CA17139), as well as on more recent work in this area. The STSM was performed by the first author at the Universidad de Campinas, Brazil, with the second and third authors, who used a local metric, developed by the second and third authors, from a distance measure between samples coming from discrete Markovian processes to decide if 2 independent samples are governed by the same stochastic law, [1]. This technique was applied to the stochastic profile of strains of the Zika virus utilizing the 4 bases of DNA, finding the probability of one base following another in the genomic sequence. Subsequent to this, the methodology was applied to SARS CoV-2, [2].

The purpose of this STSM was to complete work already begun on finding the probability that sequences of bases in DNA that are likely to be in the left-handed form, i.e. the Z-DNA conformation, occur in certain full DNA sequences. We find the probability of one base or collection of bases following or proceeding with others in the genomic sequence. This last is accomplished in a computer program written by the second author. While in Brazil these ideas were applied to find the probability of potential Z-DNA forming sequences (ZFS) occurring in rodent parvoviruses, salmonella, dengue,

and HIV, and the work already begun on SARS CoV-2 was completed. We also studied the Epstein-Barr virus, which has not been thoroughly examined for ZFS previously. Our results in this direction are especially emphasized in the current paper.

'Z-DNA is a high-energy conformation of the double helix with a zigzag backbone; this structure can be formed in alternating purine-pyrimidine sequences by negative supercoiling induced by transcription or unwrapping of nucleosomes. The discovery of antibodies and enzymes that bind specifically to Z-DNA has allowed the investigation of the biological significance of Z- DNA and the interdependence between transcription and Z-DNA formation', [3].

'Base modifications are known to affect the structure and function of DNA. C8-guanine adducts from various carcinogenic compounds have been shown to be potent Z-DNA inducers. Hence, it has been hypothesized that Z-DNA plays a role in cancer and other genetic diseases. The discoveries of Z-DNA binding proteins including ADAR1, E3L, DLM1, and PKZ have suggested the relevance of Z-DNA in living systems. In addition, increasing evidence on the Z-DNA connection to gene transcription and inhibition reveals potential biological functions of the left-handed DNA. C8-guanine adducts that promote Z-DNA formation can be used as a tool to explore the Z-DNA function and its role in carcinogenesis', [4].

The first author investigates Z-DNA from a topological perspective, [5]. There is a topological approach to deciding if Z-DNA is present by studying its 3-dimensional conformation. Some of the Z-DNA-forming conditions that are relevant in vivo are the presence of DNA supercoiling, Z-DNA-binding proteins, and base modifications, as above. 'In early *in vitro* studies on plasmid DNA, it was found that negative supercoiling favors Z-DNA formation. Z to B transition could be facilitated by the addition of topoisomerases that can relax negatively supercoiled DNA. Since Z-DNA formation is energy-intensive, BZ junction formation is critical to alleviate torsional stress and stabilize the Z-DNA. This is relevant *in vivo* because when transcription occurs, the movement of RNA polymerase II along the DNA strand generates positive supercoiling in front of, and negative supercoiling behind, the polymerase', [6].

This paper is organized as follows: Section 2 provides the preliminaries, in particular a description of the Partition Markov Model. Section 3 gives information on the five sequences studied of the Epstein-Barr Virus (EBV) in the FASTA format, utilizing the techniques applied by the second and third authors. Section 4 contains the probabilities when looking for potential ZFS in EBV and for B-Z transition sites in EBV. In Section 5, we present the other sequences studied, especially HIV which has findings similar to the ones for EBV. We analyze our findings, mention potential further work, and conclude in Section 6.

2 Preliminaries

In [1] the authors develop a Bayesian Information Criterion (BIC)-based consistent metric for Markovian processes. They show that the metric is statistically consistent and, in the case where the stochastic laws are not the same, they use the metric to find where the discrepancies are. In this particular paper, their results are used to decide if 2 lines of production are equivalent, in this case, the different columns used to produce fuel from sugar cane. Moreover, using the metric defined above, they are able to identify the strings that mark discrepancies between the processes. As the industry would want each column to perform similarly, this is important information to ensure this.

In [1] the authors show that the BIC can be used to obtain a consistent estimation of the partition of a Markov process. This partition comes from an equivalence relation based on the transition probabilities of elements of the state space. The authors then go on to introduce a measure to quantify the distance between the parts of a partition. This measure also forms a metric. The authors use their results to model internet navigation patterns. They identify strings that can be considered equivalent in terms of the next step of internet surfers and find the minimal partition. This information can be important in determining user profiles as desired by browser companies. In [5], the work of the authors above was examined from a topological perspective, forming the metric space and then the topological space from the distance measure. Possible further applications of this methodology are then introduced. Since the development of these techniques is covered completely in the previous 2 citations, we will not include that description here. We recommend that the reader read any of those works.

The authors of the last 2 cited papers, along with others [7] [8], have utilized the results in many and

varied areas. For example, the authors of [2] have examined the Stochastic Profile of Strains of the Zika virus utilizing the 4 bases of DNA, $\Delta = \{a, c, g, t\}$, adenine (a), cytosine (c), guanine (g) and thymine (t). They find the probability of one base following another in the genomic sequence. The first author of the current paper was involved in an EU-Cost Action grant IMAAC (CA 16227) investigating such vector-borne diseases and, through that, saw the second and third authors' work on the Zika virus. We realized this examination of base order in genomic sequences could be applied to investigations into Z-DNA, the left-handed version of the usually right-handed molecule, as it has been experimentally determined what orders of bases are likely to be in left-handed conformation, [9]. The technique can be used to find the probability of such sequences occurring. Hence, the STSM through the EU-COST Action grant EUTOPIA (CA17139). We studied SARS CoV-2, rodent parvoviruses, salmonella, HIV, and EBV. We also applied the analysis to another vector-borne disease, namely dengue, and it could be applied to Eastern Equine Encephalitis, malaria, West Nile virus, and Yellow Fever. Finally, we have also investigated the B-Z transition sites.

3 Epstein Barr Virus Sequences

In this Section, we provide an analysis of different versions of Epstein-Barr virus sequences similar to the investigation of the Zika virus ¯ mentioned above, [10], [11], [12], [13], [14], [15]. Table 1 (Appendix) reports the occurrences in five sequences in the FASTA format of three specific strings s composed by the concatenation of two elements coming from $\Delta = \{a, c, g, t\}$, the genetic alphabet, namely $gc, ca,$ and at. Note that for EBV the prevalence of these sequences follows the order for the propensity of Z-DNA forming subsequences mentioned in the Abstract ($cggg$ is omitted here). As mentioned above, this result will be different for other sequences we examined.

Table 2 (Appendix) shows the probability of elements of Δ following the sequences examined in Table 1 in Appendix ($cggg$ is again omitted). This is again similar to work done previously on the Zika virus. Here we can see the likelihood of the particular sequence repeating and/or the likelihood of a B-Z transition site (note that Section 4 will provide a much finer analysis). For example, the most likely repeating sequence will be ca, as opposed to gc. The most likely sequence followed by a B-Z transition (see the Abstract and Section 4.1) will be gc in the cg case; at in the tg case; and gc again in the at case.

4 Probabilities of Z-DNA Forming Subsequences and B-Z Transitions in EBV

The next tables in the Appendix report on the number of subsequences of base pairs mentioned in the Abstract that may be prone to Z-DNA formation of a given length. For example, in Table 3 (Appendix), the subsequence d(CA)3 appears 63 times in the Epstein-Barr Virus (EBV) genome AY961628.3. The rest of the table then gives the probabilities that this subsequence is followed by the base a or c or g or t. Since B-Z transitions have been studied most at CG sites, for example, we see that the probability of these 63 subsequences being followed by C is the highest, just under 50%. Other studies, however, have shown that AT sites could be where the transition takes place. These 63 subsequences being followed by A have the next highest probability, just under 25%. (Note that TG sites could also be where these transitions occur.) In our work, we concentrated on AT sites for the B-Z transition. The rationale for this is in the next section.

4.1 B-Z Transition

We examine the literature around the B-Z transition. Then we look at the probabilities for these transitions occurring in the sequences studied:

'Whenever B-DNA transforms into Z-DNA two B-Z junctions form. The crystal structure of these junctions revealed two extruded bases, adenine, and thymine at the junction. A crucial finding from this structure is that a right-handed DNA can transform to a left-handed DNA or vice versa by the disruption and extrusion of a base pair. It has also been suggested that the extruded base pairs at the B-Z DNA junction may be sites for DNA modification', [16].

'An important non-canonical structure associated with Z-DNA is a BZ junction that forms where B-DNA and Z-DNA meet with an extrusion of bases at the junction between the B- and Z-DNA. Accordingly, when Z-DNA forms in the genome, two BZ junctions flank each side of the Z-DNA-forming site (ZFS). Several studies have shown a clear sequence preference regarding the bases that can flip out at the junction, with A-T showing the highest propensity for

extrusion', [6].

'Although CG repeats, with the least propagation free energy in the B-to-Z transition (hence, free energy cost for Z-state relative to B-state), have been the most extensively studied over the last three decades, thymine-guanine (TG) repeats, simple repeat sequences with the second least cost, exist more frequently in the gene regulatory regions of eukaryotes, e.g., near rodent globin, immunoglobulin, and galactokinase genes and the human globin and actin genes. TG repeats are mainly located upstream of the first expressed exons. Besides, TG repeats are the most frequent microsatellite sequences in plants and are also common in other higher organisms. Microsatellite sequences have drawn much attention because they are recombination hotspots leading to genetic instability and serve as genetic markers. Expansion of microsatellites over generations is thought to aggravate the symptoms associated with genetic instabilities. Recently, the DNA fragility in the parallel evolution of pelvic reduction in stickleback fish has been attributed to Z-DNA formation by TG repeat sequences, which is, to date, the most straightforward evidence for the biological functions of Z-DNA. The dynamic B–Z transition observed under physiologically relevant tension and torsion indicates the physical advantage of Z-DNA, in particular, TG-repeat-based Z-DNA for its potential biological roles', [3]. We now examine our results in this direction, emphasizing the A-T extrusions.

Table 3 (Appendix) shows that for EBV the pair GC is more prevalent, but at longer lengths, CA occurs more often. It also indicates more than 500 sites where GC is preceded by or is followed by the AT pairing. These then are areas ripe for Z-DNA formation.

Since the results are similar for the other 4 versions of the EBV genome, we place their probability tables in Appendix in the Table 7, Table 8, Table 9, Table 10. We invite the reader to verify the conclusions.

5 Other Genomes

In this section, we examine the probabilities, as above, for other genomes studied. In the tables for this section, we find the pair CA more prevalent than GC, unlike EBV and the well-known indicator in the Abstract. In Table 4 (Appendix), for HIV, which follows, however, the longest sequences with the propensity to be in the Z-form are the GC repeats. All the genomes are checked for B-Z transitions using the AT paradigm. Of note is that many of the previous candidates for potential Z-DNA formation do not have as many possible sites as the EBV strains or HIV.

Table 5 (Appendix) examines one of the SARS CoV-2 strains out of the 4, [2] not examined previously to our knowledge (although Sars and Z-RNA have, [17]). (Note that the authors of [2] did an analysis of these 4 strains similar to the one carried out in Section 3.) We see again more propensity for CA pairs than for GC, but of special note is that AT pairs are quite prevalent. This indicates a need for further study on the other 3 strains.

Table 6 (Appendix) contains our results for the Dengue virus, another genome not mentioned previously as a possible place to look for Z-DNA. Once more CA is more prevalent than GC, as is AT when compared to CGGG.

In the Appendix, we put the probability tables for salmonella (Table 11 in Appendix), which has more sites than Dengue but fewer than HV8, and parvovirus (Table 12 in Appendix), which has the fewest sites, since these have been examined by others for ZFS previously.

6 Conclusions and Future Work

Since the discovery of Z-DNA in vitro in 1979, it has been hypothesized that this left-handed version of DNA exists *in vivo* and has potentially deleterious effects.

As mentioned previously, it is conjectured that Z-DNA plays a role in cancer and other genetic diseases. The fact that there are antibodies, proteins, and enzymes that specifically bind to Z-DNA has aided in these investigations. But up to now the only specific evidence for the biological functions of Z-DNA has been found in stickleback fish, as stated above.

The search for Z-DNA forming sequences (ZFS) has been proceeding apace since its discovery. For example, in [9] 391 sites were found and examined *in vivo*. We continue these efforts in this work by applying probability and Partition Markov Models. In particular, we find that the Epstein-Barr virus and HIV have many more sites that are Z-DNA prone.

Moreover, the probability of B-Z transition sites is highest in these genomes among the ones studied. These results are novel as far as we can determine. It is also the case that the longest sequences prone to Z-DNA formation were found in EBV and HIV, with the

longest being in the HV8 genomic sequence. These results then lead us to explore HPV, the Human Papilloma Virus, via [18].

Except for EBV, we find that the CA pairing is more prevalent than GC. The same is true for AT pairs as opposed to CGGG. So a reconsideration of where to look for ZFS may be in order. This observation, and the more important findings in the previous paragraph will help DNA researchers focus on more exact areas when looking for Z-DNA *in vivo* and for potential B-Z transition sites.

There are many more avenues to explore based on our work to date. We have already mentioned that the prevalence of AT pairs in the SARS CoV-2 genome merits further attention. This is due to the fact that extrusions at A-T sites are where B-Z transitions are prone to occur. Examination of the other 3 strains of SARS CoV-2 is hence called for and a Ph.D. student in Brazil is currently doing just that.

This same student is also investigating the other potential B-Z transition sites probabilistically. Namely the CG and TG pairs.

Finally, during the STSM some carcinogens were studied. In particular, the third author examined strains that included ZFS and ones that did not. The Partition Markov Model is being applied here to form a classification. This could also be done in the EBV and Sars cases.

Acknowledgements:
The STSM was funded by the EU-COST Action EUTOPIA (CA17139).

The first author was supported in part by this grant and another EU-COST Action grant IMAAC (CA 16227). Other work was performed by the first author at the American University of Malta, and this author is thankful for its support.

The second and third authors performed this study in the Department of Statistics at the University of Campinas, Brazil, and they are thankful for its support.

Finally, we thank the linguist Silvia Storti Shelyta for assisting with the writing of this paper.

References:

[1] Garcia JE, Gholizadeh R, Gonzalez-Lopez VA. A BIC-based consistent metric between Markovian processes. *Appl Stochastic Models Bus Ind.*, 2018, 34, 868-878. https://doi.org/10.1002/asmb.2346.

[2] Garcia JE, Gonz´alez-Lo´pez VA, Tasca, GH. Multiple partition Markov model for B.1.1.7, B.1.351, B.1.617.2, and P.1 variants of SARS-CoV 2 virus. *Computational Statistics*, 2022. https://doi.org/10.1007/s00180- 022-01291-8.

[3] Kim SH, Jung HJ, Lee IB, Lee NK, Hong SC. Sequence-dependent cost for Z-form shapes the torsion-driven B–Z transition via close interplay of Z-DNA and DNA bubble. *Nucleic Acids Research*, 2021, 49(7), 3651–3660, https://doi.org/10.1093/nar/gkab153.

[4] Vongsutilers V, Gannett PM. C8-Guanine modifications effect on Z- DNA formation and its role in cancer. *Org Biomol Chem.*, 2018, 16(13), 2198-2209. https://doi.org/10.1039/c8ob00030a.

[5] Bozeman JR. On the metric topology induced from the BIC-based consistent metric between Markovian processes. *Adv. Stud. Cont. Math.,* 2022, 32(3), 347-359, http://dx.doi.org/10.17777/ascm2022.32.3.347

[6] Ravichandran S, Subramani VK, Kim KK. Z-DNA in the genome: from structure to disease. *Biophys Rev.*, 2019 Jun, 11(3), 383–387. https://doi.org/10.1007/s12551-019-00534-1.

[7] Zhao LC, Dorea CCY, Gon¸calves CR. On determination of the order of a Markov chain. *Statistical inference for stochastic processes.* 2001; 4(3), 273-282. https://doi.org/10.1023/A:1012245821183.

[8] Pereira DFS. Efficient Determination Criterion for Estimation of Minimum Partition Markov Chains (Critério de Determinação Eficiente para Estimação de Cadeias de Markov de Partição Mínima), 2021; Msc Dissertation, [Online]. https://mat.unb.br/images/smat/repositorio/2022_02_02/Dissertacao_DiegoFelipe.pdf (Accessed Date: August 10, 2024).

[9] Shin SI, Ham S, Park J, Seo SH, Lim CH, Jeon H, Huh J, Roh TY. Z- DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Research*, 2016, 23(5). 477–

486, https://doi.org/10.1093/dnares/dsw031.

[10] Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ. The genome of Epstein-Barr virus type 2 strain AG876. *Virology*, 2006, 350 (1) p. 164, http://dx.doi.org/10.1016/j.virol.2006.01.015.

[11] Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, S´eguin C, Tuff PS, Barrell BG. DNA sequence and expression of the B95-8 Epstein-Barr virus genome, *Nature 1984*, 310 (5974) p. 207, doi: 10.1038/310207a0.

[12] Liu P, Fang X, Feng Z, Guo YM, Peng RJ, Liu T, Huang Z, Feng Y, Sun X, Xiong Z, Guo X, Pang SS, Wang B, Lv X, Feng FT, Li DJ, Chen LZ, Feng QS, Huang WL, Zeng MS, Bei JX, Zhang Y, Zeng YX. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *Journal of Virology*, 2011, 85 (21) p.11291, doi: 10.1128/JVI.00823-11.

[13] Sample J, Young L, Martin B, Chatman T, Kieff E, Rickinson A, Kieff E. Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA- 3B, and EBNA-3C genes. *Journal of Virology*. 1990, 64 (9) p. 4084, doi: 10.1128/JVI.64.9.4084-4092.

[14] Kwok H, Tong AH, Lin CH, Lok S, Farrel PJ, Kwong DL, Chiang AK. Genomic sequencing and comparative analysis of Epstein-Barr Virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE*, 2012, 7 (5): e36939,doi: 10.1371/journal.pone.0036939.

[15] Zeng MS, Li DJ, Liu QL, Song LB, Li MZ, Zhang RH, Yu XJ, Wang HM, Emberg I, Zeng YX. Genomic sequence analysis of Epstein-Barr Virus strain GD1 from a nasopharyngeal carcinoma patient. *Journal of Virology*, 2005, 79 (24) p. 15323, doi: 10.1128/JVI.79.24.15323-15330.2005.

[16] Ha SC, Lowenhaupt K, Rich A, Kim YG, Kim KK. Crystal struc- ture of a junction between B-DNA and Z-DNA reveals two ex-truded bases. Nature 2005 Oct 20, 437(7062), 1183-6. PMID: 16237447, doi: 10.1038/nature04088.

[17] Herbert A, Poptsova A. Z-RNA and the Flipside of the SARS Nsp13 Helicase: Is There a Role for Flipons in Coronavirus-Induced Pathology? *Front Immunol.*, 2022 Jun. 17, 13:912717 doi: 10.3389/fimmu.2022.912717.

[18] Stebbing J, Bower M. Epstein-Barr virus in Burkitt's lymphoma: the missing link. *The Lancet Oncology*, 2009, 10(4). 430, https://doi.org/10.1016/S1470-2045(09)70045-2.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The STSM was funded by the EU-COST Action EUTOPIA (CA17139).

The first author was supported in part by this grant and another EU-COST Action grant IMAAC (CA 16227).

Conflict of Interest

The authors have no conflicts of interest to declare.

APPENDIX

Table 1. Occurrences of the strings $s = at, ca$ and gc in sequences of Epstein-Barr Virus in the FASTA format. From left to right: (i) GenBank and version of the sequence, (ii) size of the sequence, (iii) reference for the sequence, (iv)-(vi) occurrences of the strings $gc, ca,$ and $at,$ respectively

GenBank/version	size	ref.	*gc*	*ca*	*at*
AY961628.3	171657	[6]	13860	11813	6622
DQ279927.1	172764	[1]	13954	11909	6603
HQ020558.1	164669	[3]	12899	11474	6521
JQ009376.2	171553	[5]	13025	11520	6511
NC 007605.1	171823	[2]	13835	11806	6621

Table 2. Transition probabilities from the string s to each element of Δ, from top to bottom: $s = gc, ca,$ and $at,$ respectively.

	$P(a\|gc)$	$P(c\|gc)$	$P(g\|gc)$	$P(t\|gc)$
AY961628.3	0.21861	0.39127	0.17727	0.21284
DQ279927.1	0.21793	0.39602	0.17522	0.21084
HQ020558.1	0.22095	0.39375	0.17660	0.20831
JQ009376.2	0.22234	0.38902	0.18219	0.20637
NC 007605.1	0.21735	0.39465	0.17636	0.21164
	$P(a\|ca)$	$P(c\|ca)$	$P(g\|ca)$	$P(t\|ca)$
AY961628.3	0.16744	0.25531	0.38127	0.19597
DQ279927.1	0.16668	0.25972	0.37904	0.19456
HQ020558.1	0.17222	0.25745	0.37197	0.19836
JQ009376.2	0.17092	0.25495	0.37648	0.19766
NC 007605.1	0.16881	0.25682	0.37945	0.19490
	$P(a\|at)$	$P(c\|at)$	$P(g\|at)$	$P(t\|at)$
AY961628.3	0.18076	0.25687	0.34174	0.22063
DQ279927.1	0.17992	0.25685	0.34363	0.21960
HQ020558.1	0.17620	0.26192	0.34274	0.21898
JQ009376.2	0.17954	0.25910	0.34357	0.21748
NC 007605.1	0.17958	0.25842	0.34179	0.22021

Table 3. Number of subsequences of given length followed by the probabilities of the base following for the Epstein-Barr virus genome AY961628.3

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	13860	0.21861	0.39126	0.17727	0.21284
d(CA)1	11813	0.16744	0.25531	0.38127	0.19597
d(CGGG)1	1356	0.11283	0.28761	0.37758	0.22197
d(AT)1	6622	0.18076	0.25687	0.34173	0.22062
d(GCAT)1	582	0.19072	0.31099	0.31615	0.18213
d(ATGC)1	533	0.23452	0.38461	0.14821	0.23264
d(GC)2	682	0.22434	0.39296	0.20967	0.17302
d(CA)2	692	0.18786	0.27601	0.30924	0.22687
d(CGGG)2	15	0.2	0.4	0.4	0
d(AT)2	281	0.20996	0.24199	0.32028	0.22775
d(GCAT)2	2	0	0.5	0.5	0
d(ATGC)2	2	0	0.5	0.5	0
d(GC)3	42	0.30952	0.42857	0.16666	0.09523
d(CA)3	63	0.23809	0.47619	0.20634	0.07936
d(CGGG)3	4	0	0	1	0
d(AT)3	8	0.125	0.375	0.25	0.25
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	2	0.5	0.5	0	0
d(CA)4	4	0.25	0.25	0.25	0.25
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN

Table 4. Number of subsequences of given length followed by the probabilities of the succeeding base for the HIV virus genome NC009333.1.hv8

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	9762	0.23837	0.29655	0.23960	0.22546
d(CA)1	10098	0.22390	0.27678	0.29094	0.20835
d(CGGG)1	605	0.20495	0.23140	0.35041	0.21322
d(AT)1	6938	0.23854	0.25454	0.28293	0.22398
d(GCAT)1	464	0.23491	0.24137	0.26724	0.25646
d(ATGC)1	469	0.25799	0.27505	0.21108	0.25586
d(GC)2	734	0.19209	0.35013	0.25068	0.20708
d(CA)2	718	0.23259	0.26462	0.28272	0.22005
d(CGGG)2	1	0	0	1	0
d(AT)2	466	0.26609	0.26824	0.22961	0.23605
d(GCAT)2	1	1	0	0	0
d(ATGC)2	0	NaN	NaN	NaN	NaN
d(GC)3	68	0.17647	0.41176	0.27941	0.13235
d(CA)3	63	0.23809	0.31746	0.22222	0.22222
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	33	0.45454	0.24242	0.15151	0.15151
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	9	0.33333	0.11111	0.33333	0.22222
d(CA)4	7	0.28571	0.42857	0	0.28571
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	9	0.88888	0.11111	0	0
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 5. Number of subsequences of given length followed by the probabilities of the succeeding base for the Covid virus genome NC045512.2.cov

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	1168	0.31849	0.16010	0.07534	0.44606
d(CA)1	2084	0.33733	0.22024	0.21017	0.23224
d(CGGG)1	10	0.1	0.3	0	0.6
d(AT)1	2308	0.20407	0.14688	0.31412	0.33492
d(GCAT)1	78	0.26923	0.17948	0.19230	0.35897
d(ATGC)1	153	0.32026	0.18954	0.05882	0.43137
d(GC)2	17	0.35294	0.05882	0.17647	0.41176
d(CA)2	156	0.32051	0.23076	0.21794	0.23076
d(CGGG)2	0	NaN	NaN	NaN	NaN
d(AT)2	111	0.22522	0.15315	0.29729	0.32432
d(GCAT)2	1	0	0	0	1
d(ATGC)2	1	1	0	0	0
d(GC)3	0	NaN	NaN	NaN	NaN
d(CA)3	8	0.375	0.375	0.125	0.125
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	8	0.25	0.25	0.125	0.375
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	0	NaN	NaN	NaN	NaN
d(CA)4	0	NaN	NaN	NaN	NaN
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 6. Number of subsequences of given length followed by the probabilities of the succeeding base for the Dengue virus genome KF8249021.den

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	511	0.36986	0.22896	0.10371	0.29745
d(CA)1	937	0.30522	0.21664	0.26040	0.21771
d(CGGG)1	12	0.25	0	0.41666	0.33333
d(AT)1	678	0.19616	0.17109	0.40560	0.22713
d(GCAT)1	35	0.17142	0.22857	0.37142	0.22857
d(ATGC)1	36	0.25	0.30555	0.11111	0.33333
d(GC)2	10	0.7	0	0	0.3
d(CA)2	90	0.34444	0.25555	0.34444	0.05555
d(CGGG)2	0	NaN	NaN	NaN	NaN
d(AT)2	27	0.07407	0.25925	0.37037	0.29629
d(GCAT)2	0	NaN	NaN	NaN	NaN
d(ATGC)2	0	NaN	NaN	NaN	NaN
d(GC)3	0	NaN	NaN	NaN	NaN
d(CA)3	8	0.125	0.25	0.625	0
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	1	0	0	1	0
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	0	NaN	NaN	NaN	NaN
d(CA)4	0	NaN	NaN	NaN	NaN
(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

More probabilities for EBV sequences and for parvovirus and salmonella.

Table 7. Number of subsequences of given length followed by the probabilities of the base following for the Epstein-Barr virus genome DQ279927.1.ebv

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	13954	0.21793	0.39601	0.17521	0.21083
d(CA)1	11909	0.16668	0.25971	0.37904	0.19455
d(CGGG)1	1384	0.11416	0.28251	0.37066	0.23265
d(AT)1	6603	0.17991	0.25685	0.34363	0.21959
d(GCAT)1	580	0.2	0.31379	0.29827	0.18793
d(ATGC)1	541	0.24029	0.39556	0.13678	0.22735
d(GC)2	686	0.23323	0.40379	0.18658	0.17638
d(CA)2	690	0.19420	0.27971	0.31014	0.21594
d(CGGG)2	10	0.3	0.5	0.2	0
d(AT)2	279	0.193548	0.25089	0.33333	0.22222
d(GCAT)2	1	0	0	1	0
d(ATGC)2	1	0	1	0	0
d(GC)3	40	0.3	0.475	0.125	0.1
d(CA)3	57	0.24561	0.49122	0.17543	0.08771
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	8	0.25	0.375	0.125	0.25
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	2	0.5	0.5	0	0
d(CA)4	4	0.25	0.25	0.25	0.25
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 8. Number of subsequences of given length followed by the probabilities of the base following for the Epstein-Barr virus genome HQ020558.1.ebv

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	12899	0.22094	0.39375	0.17660	0.20831
d(CA)1	11474	0.17221	0.25745	0.37197	0.19836
d(CGGG)1	1118	0.13774	0.32647	0.34347	0.19230
d(AT)1	6521	0.17619	0.26192	0.34273	0.21898
d(GCAT)1	566	0.19434	0.31448	0.31978	0.17137
d(ATGC)1	522	0.23563	0.39655	0.13984	0.22796
d(GC)2	643	0.23639	0.39035	0.19906	0.17418
d(CA)2	672	0.19642	0.28422	0.30208	0.21726
d(CGGG)2	7	0.42857	0.28571	0.28571	0
d(AT)2	272	0.17279	0.28308	0.30882	0.23529
d(GCAT)2	2	0	0.5	0.5	0
d(ATGC)2	2	0	0.5	0.5	0
d(GC)3	42	0.30952	0.40476	0.16666	0.11904
d(CA)3	63	0.23809	0.46031	0.20634	0.09523
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	14	0.14285	0.35714	0.21428	0.28571
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	2	0.5	0.5	0	0
d(CA)4	6	0.16666	0.16666	0.33333	0.33333
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	1	0	1	0	0
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 9. Number of subsequences of given length followed by the probabilities of the base following for the Epstein-Barr virus genome JQ009376.2.ebv

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	13025	0.22234	0.38902	0.18218	0.20637
d(CA)1	11520	0.17092	0.25494	0.37647	0.19765
d(CGGG)1	1174	0.13032	0.32879	0.36115	0.17972
d(AT)1	6511	0.17954	0.25909	0.34357	0.21747
d(GCAT)1	576	0.19965	0.30902	0.31597	0.17534
d(ATGC)1	527	0.23529	0.38330	0.14231	0.23908
d(GC)2	655	0.24122	0.37862	0.21221	0.16793
d(CA)2	673	0.19613	0.28083	0.30757	0.21545
d(CGGG)2	7	0.42857	0.28571	0.28571	0
d(AT)2	283	0.19787	0.25795	0.32508	0.21554
d(GCAT)2	2	0	0.5	0.5	0
d(ATGC)2	2	0	0.5	0.5	0
d(GC)3	44	0.29545	0.38636	0.22727	0.09090
d(CA)3	62	0.24193	0.46774	0.19354	0.09677
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	10	0.1	0.4	0.2	0.3
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	2	0.5	0.5	0	0
d(CA)4	6	0.16666	0.16666	0.33333	0.33333
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 10. Number of subsequences of given length followed by the probabilities of the base following for the Epstein-Barr virus genome NC007605.1.ebv

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	13835	0.21734	0.39465	0.17636	0.21163
d(CA)1	11806	0.16881	0.25681	0.37946	0.19490
d(CGGG)1	1387	0.11391	0.28406	0.38139	0.22062
d(AT)1	6621	0.17958	0.25842	0.34179	0.22020
d(GCAT)1	579	0.18998	0.31260	0.31778	0.17962
d(ATGC)1	533	0.22701	0.38836	0.14821	0.23639
d(GC)2	674	0.22551	0.40652	0.19287	0.17507
d(CA)2	684	0.19005	0.27923	0.31432	0.21637
d(CGGG)2	11	0.27272	0.54545	0.18181	0
d(AT)2	277	0.22743	0.23465	0.30685	0.23104
d(GCAT)2	2	0	0.5	0.5	0
d(ATGC)2 1	0	1	0	0	
d(GC)3	40	0.325	0.425	0.15	0.1
d(CA)3	60	0.23333	0.5	0.18333	0.08333
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	7	0.14285	0.42857	0.14285	0.28571
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	2	0.5	0.5	0	0
d(CA)4	4	0.25	0.25	0.25	0.25
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 11. Number of subsequences of given length followed by the probabilities of the base following for the parvovirus genome NC038545.1.par

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	232	0.38362	0.16810	0.09051	0.35775
d(CA)1	418	0.33971	0.20813	0.30382	0.14832
d(CGGG)1	4	0.5	0	0.5	0
d(AT)1	299	0.26086	0.15719	0.39464	0.18729
d(GCAT)1	7	0	0	0.71428	0.28571
d(ATGC)1	17	0.29411	0.17647	0	0.52941
d(GC)2	5	0	0	0.6	0.4
d(CA)2	41	0.19512	0.24390	0.34146	0.21951
d(CGGG)2	0	NaN	NaN	NaN	NaN
d(AT)2	16	0.1875	0.0625	0.375	0.375
d(GCAT)2	0	NaN	NaN	NaN	NaN
d(ATGC)2	0	NaN	NaN	NaN	NaN
d(GC)3	1	0	0	0	1
d(CA)3	6	0.16666	0.5	0.33333	0
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	0	NaN	NaN	NaN	NaN
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	0	NaN	NaN	NaN	NaN
d(CA)4	2	0	0	1	0
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN

Table 12. Number of subsequences of given length followed by the probabil- ities of the base following for the salmonella genome NC025443.1.sal

Sequence/length	amount	$P(a)$	$P(c)$	$P(g)$	$P(t)$
d(GC)1	2494	0.35164	0.15677	0.24097	0.25060
d(CA)1	3885	0.23603	0.19922	0.32483	0.23989
d(CGGG)1	140	0.28571	0.12857	0.22142	0.36428
d(AT)1	4290	0.24242	0.23170	0.24825	0.27762
d(GCAT)1	229	0.26200	0.31004	0.13537	0.29257
d(ATGC)1	212	0.35849	0.18396	0.24056	0.21698
d(GC)2	160	0.3125	0.2	0.225	0.2625
d(CA)2	212	0.22641	0.14622	0.36792	0.25943
d(CGGG)2	0	NaN	NaN	NaN	NaN
d(AT)2	310	0.18387	0.24516	0.22580	0.34516
d(GCAT)2	0	NaN	NaN	NaN	NaN
d(ATGC)2	0	NaN	NaN	NaN	NaN
d(GC)3	0	NaN	NaN	NaN	NaN
d(CA)3	5	0	0	0.2	0.8
d(CGGG)3	0	NaN	NaN	NaN	NaN
d(AT)3	11	0.27272	0	0.27272	0.45454
d(GCAT)3	0	NaN	NaN	NaN	NaN
d(ATGC)3	0	NaN	NaN	NaN	NaN
d(GC)4	0	NaN	NaN	NaN	NaN
d(CA)4	0	NaN	NaN	NaN	NaN
d(CGGG)4	0	NaN	NaN	NaN	NaN
d(AT)4	0	NaN	NaN	NaN	NaN
d(GCAT)4	0	NaN	NaN	NaN	NaN
d(ATGC)4	0	NaN	NaN	NaN	NaN