Magnus Representation of Genome Sequences

Chengyuan Wu[a,1,*], Shiquan Ren[a,*], Jie Wu[a,*], Kelin Xia[b,*]

[a] *Department of Mathematics, National University of Singapore, Singapore 119076*
[b] *Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371*

Abstract

We introduce an alignment-free method, the Magnus Representation, to analyze genome sequences. The Magnus Representation captures higher-order information in genome sequences. We combine our approach with the idea of k-mers to define an effectively computable Mean Magnus Vector. We perform phylogenetic analysis on two datasets: mosquito-borne viruses and filoviruses. Our results on ebolaviruses are consistent with previous phylogenetic analyses, and confirm the modern viewpoint that the 2014 West African Ebola outbreak likely originated from Central Africa. Our analysis also confirms the close relationship between *Bundibugyo ebolavirus* and *Taï Forest ebolavirus*.

1. Introduction

In the field of combinatorial group theory, Wilhelm Magnus studied representations of free groups by non-commutative power series (Lyndon and Schupp, 2015). For a free group F with basis $x_1, \ldots, x_n$ and a power series ring Π in indeterminates $\xi_1, \ldots, \xi_n$, Magnus showed that the map $\mu : x_i \mapsto 1 + \xi_i$ defines an isomorphism from F into the multiplicative group $\Pi^\times$ of units in Π. Using concepts from Magnus' work, we define the Magnus representation and Magnus vector of a DNA/RNA sequence, and apply them in the analysis of genomes (Huang, 2016; Dong et al., 2017; Kwan and Arniker, 2009).

2. Materials and methods

2.1. Summary of the procedures

We compute the mean Magnus vectors of non-overlapping k-mers of virus genomes, and their mutual Euclidean distances. We store the distances in a distance matrix and construct a phylogenetic tree (or dendrogram) using neighbor-joining or UPGMA.

2.2. Magnus Representation and Magnus Vector

The Magnus representation and Magnus vector of a DNA/RNA sequence are described as follows. Consider a DNA sequence $S = x_1x_2 \ldots x_N$ of length N, where the x_i lie in the set $\{A, C, G, T\}$ (or $\{A, C, G, U\}$ in the case of RNA). Our subsequent notation will mainly follow that of DNA sequences for convenience, but we emphasize that our methods work for RNA sequences as well. We define the Magnus representation of S, denoted $\rho(S)$, to be the product $\prod_{i=1}^{N}(1 + x_i)$ in the non-commutative polynomial algebra $R\langle A, C, G, T\rangle$, where R is a commutative ring. In practice, we may take R to be the set of real numbers $\mathbb{R}$, the set of integers $\mathbb{Z}$ or the ring of integers modulo 2, $\mathbb{Z}/2$.

The Magnus vector of a DNA sequence S, denoted by $v(S)$, is obtained by two steps:

1. Arrange the set of possible words over the alphabet $\{A, C, G, T\}$ of length less than or equal to N first by ascending order of length and then by lexicographic order.
2. With respect to the above arrangement, assign $c \in R$ for each term present in $\rho(S)$ with coefficient c, and 0 for each term not present in $\rho(S)$.

The Magnus vector is a $(\sum_{i=1}^{N} 4^i)$-tuple, or equivalently, a $(\frac{4^{N+1}-4}{3})$-tuple. We illustrate this in the following example. Consider the DNA sequence $S = AC$. Then, the Magnus representation is $\rho(S) = (1 + A)(1 + C) = 1 + A + C + AC$. The arrangement of the set of possible words of length less than or equal to 2 is: A, C, G, T, AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT. Hence, the Magnus vector is the 20-tuple $v(S) = (1, 1, 0, 0, 0, 1, 0, 0, \ldots, 0)$. Note that due to non-commutativity of the variables, if $S' = CA$, we observe that $\rho(S') = 1 + A + C + CA \neq \rho(S)$ and $v(S') = (1, 1, 0, 0, 0, 0, 0, 0, 1, 0, \ldots, 0) \neq v(S)$.

We show an example when $R = \mathbb{Z}/2$. If $S = AA$, then $\rho(S) = (1 + A)(1 + A) = 1 + AA$. By looking at the word with greatest length in $\rho(S)$, we observe that it is the DNA sequence itself. Hence, for $R = \mathbb{R}$, $\mathbb{Z}$, or $\mathbb{Z}/2$, the Magnus representation is faithful (namely injective). This means that the Magnus representation is able to distinguish between any two different DNA sequences, and detect all forms of DNA mutations.

It can be seen that the Magnus vector consists of many coordinates even for $N = 2$, however it is typically sparse.

[*] First authors
Email addresses: wuchengyuan@u.nus.edu (Chengyuan Wu), sren@u.nus.edu (Shiquan Ren), matwuj@nus.edu.sg (Jie Wu), xiakelin@ntu.edu.sg (Kelin Xia)
[1] Corresponding author

Hence, we introduce the short Magnus vector to compress the Magnus vector into fewer coordinates. The short Magnus vector is defined to be the vector whose first coordinate is the dimension of the Magnus vector (i.e. the number of coordinates $\sum_{i=1}^{N} 4^i$), and the subsequent entries indicate the position of the non-zero entries of the Magnus vector. We denote the short Magnus vector also by $v(S)$, since in practice there is no danger of confusion. In the example of $S = AC$, the short Magnus vector is $v(S) = (20, 1, 2, 6)$. For the example of $S' = CA$, the short Magnus vector is $v(S') = (20, 1, 2, 9)$. If $R = \mathbb{Z}/2$, there is a one-to-one correspondence between the set of Magnus vectors and the set of short Magnus vectors.

We remark that there are other theoretical ways to understand the algebra of formal power series on non-commuting variables (see Appendix A).

2.3. Algorithm for Magnus Vector

We use a quaternary (base 4) system to encode the DNA sequence. Namely, we let the digits 0, 1, 2, 3 represent the letters A, C, G, T respectively. We use the term *DNA subsequence* to denote a sequence that can be derived from the original DNA sequence by deleting some or no letters without changing the order of the remaining letters.

For a DNA sequence of length N, we count the number of occurrences of the subsequences “0”, “1”, “2”, “3”, “00”, “01”, etc. The counting can be done efficiently through dynamic programming. For each subsequence S (considered as a base 4 number), we convert it to a decimal (base 10) numeral d_S. Let α_S be the number of occurrences of the subsequence S (in the original DNA sequence), and l_S be the length (number of digits) of S. Define $p_S = \frac{4^{l_S}-4}{3} + d_S + 1$. Then, the p_S-th component of the Magnus vector is precisely α_S.

For instance, consider the DNA sequence CCGAG and the subsequence $S = CCG = 112$. Then, $\alpha_S = 2$, $l_S = 3$ and $d_S = 22$. Hence, $p_S = 43$ and the 43rd component of the Magnus vector is 2.

We implement the algorithm in Python.

2.4. k-mer and size selection

A k-mer is a segment of k consecutive nucleotides of a genome sequence (Huang, 2016; Koren et al., 2017; Rizk et al., 2013). Due to the length of the Magnus vector, it is not practical to compute it for the entire genome sequence of length N. Instead, we compute the Magnus vector for each k-mer, which are enumerated by non-overlapping sliding windows of size k, shifting k nucleotides each time until the entire sequence (possibly excluding up to $k-1$ nucleotides at the tail end if N is not divisible by k) is scanned.

We choose non-overlapping sliding windows due to the observation that the Magnus vectors of overlapping windows may contain similar information (counting the same subsequences). Hence, non-overlapping sliding windows will reduce the total number of windows without much loss of information. The value of k is chosen to be 5 as it corresponds to the smallest classification errors of the Baltimore and genus classification labels (Huang, 2016). This leads to the Magnus vector having length 1364.

2.5. Mean Magnus Vector and Euclidean distance

We calculate the *mean Magnus vector* of a genome sequence by dividing the sum of Magnus vectors of all k-mers, by the total number of windows. We use the Euclidean distance to calculate the distance between the mean Magnus vector of two different genome sequences.

2.6. Benefits of our approach

The Magnus Representation (and Magnus vector) contains higher-order information about the genome sequence, in the form of its subsequences. This is in contrast to lower-order information such as simply counting the number of letters ‘A’, ‘C’, etc., in the genome sequence. The non-commutativity of the variables enhances the discriminatory power of the Magnus Representation by distinguishing between permutations of subsequences.

By combining the Magnus Representation with the idea of k-mers, we improve the computability issue of our approach. (It is infeasible to compute the Magnus Representation of an entire long genome sequence.) By breaking up the genome sequence into k-mers, the mean Magnus vector of a large genome sequence can be effectively computed.

2.7. Data

2.7.1. Mosquito-borne viruses

We use virus genome data from GenBank (Benson et al., 2008, 2012), with a focus on mosquito-borne viruses such as dengue (Tuiskunen Bäck and Lundkvist, 2013). We also include two plant viruses (*tobacco mosaic virus* and *cauliflower mosaic virus*) for contrast. We compute their mean Magnus vector and the corresponding distance matrix. We then draw a phylogenetic tree (neighbor-joining) using MEGA7 (Kumar et al., 2016).

2.7.2. Ebolaviruses

Following the seminal paper by Hui Zheng, Stephen S.-T. Yau and coauthors (Zheng et al., 2015), we study 69 filoviruses, and draw the phylogenetic tree. Virus genome data are also taken from GenBank. The 69 filoviruses correspond exactly to the ones in Figure 2 and Supplementary Table S2 of Zheng et al. (2015).

3. Results

3.0.1. Mosquito-borne viruses

The distance matrix (presented as a lower-triangular matrix) is given in Table 1, with entries correct to 3 decimal places.

The phylogenetic tree (strictly speaking, a dendrogram) drawn using MEGA7 by neighbor-joining is presented in Figure 1.

Table 1: Distance matrix (lower-triangular)

	1	2	3	4	5	6	7	8	9	10
1										
2	0.189									
3	0.360	0.478								
4	0.437	0.559	0.146							
5	0.593	0.720	0.427	0.393						
6	0.569	0.700	0.419	0.380	0.249					
7	0.585	0.656	0.401	0.437	0.657	0.723				
8	0.584	0.718	0.386	0.346	0.188	0.283	0.596			
9	0.781	0.860	0.727	0.713	0.932	0.791	0.964	0.883		
10	0.887	0.792	1.046	1.111	1.398	1.336	1.087	1.374	1.029	

Legend:

1. *Dengue virus* type 1 strain 16007 (10735 bp)
2. *Dengue virus* type 2 strain 16681 (10723 bp)
3. *Dengue virus* type 3 vector p3(delta30) (15145 bp)
4. *Dengue virus* type 4 vector p4 (15270 bp)
5. *Zika virus* VEN/UF-1/2016 (10808 bp)
6. *Yellow fever virus* strain Trinidad 79A isolate 788379 (10760 bp)
7. *Chikungunya virus strain* 06113879 (11929 bp)
8. *West Nile virus* from USA (11030 bp)
9. *Tobacco mosaic virus* genome (variant 1) (6395 bp)
10. *Cauliflower mosaic virus* isolate NY8153 (8030 bp)

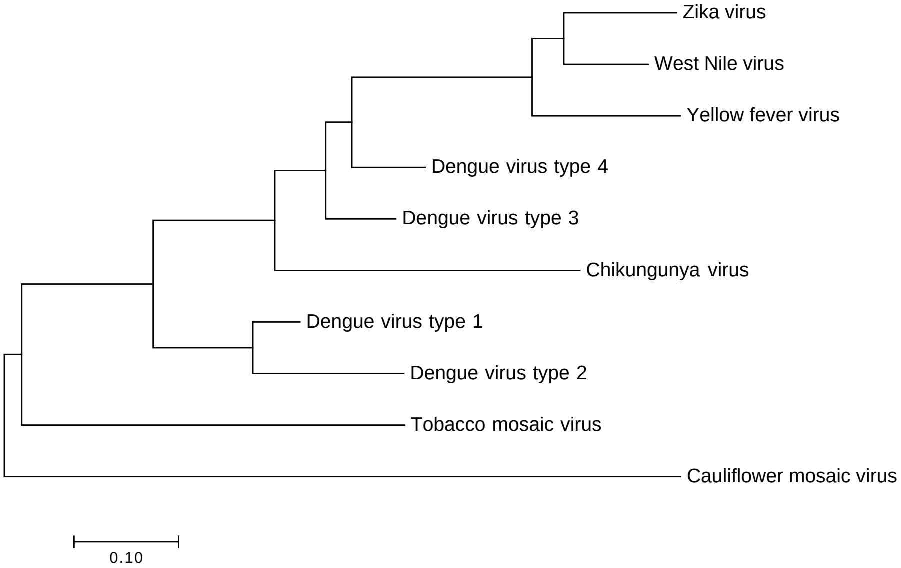

Figure 1: Phylogenetic tree (neighbor-joining) drawn using MEGA7.

3.0.2. Ebolaviruses

The phylogenetic tree was drawn using the UPGMA (unweighted pair group method with arithmetic mean) method, using MEGA7. The optimal tree with the sum of branch length = 2.09343667 is shown in Figure 2. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were obtained by calculating the Euclidean distance between the respective mean Magnus vectors.

We chose the UPGMA method to correspond to Figure 2 in Zheng et al. (2015).

In the process, we also discover that some viruses are labelled as distinct in GenBank, but are actually the same virus with identical DNA. They are the following pairs of viruses (denoted by their GenBank Accession Number): (FJ217161, NC_014373), (AF522874, NC_004161), (AY729654, NC_006432).

4. Discussion

4.0.1. Mosquito-borne viruses

Our results obtained by the Magnus vector approach make biological sense. Intuitively, the mosquito-borne viruses should be more or less similar within the group, but should have big differences compared to plant viruses such as *tobacco mosaic virus* and *cauliflower mosaic virus*. Our results clearly reflect this, as shown in the phylogenetic tree in Figure 1.

Our results also show some interesting phenomena among the mosquito-borne viruses. Among the 4 types of dengue viruses, Type 1 and Type 2 appear to be closely related. On the other hand, Type 3 and Type 4 also appear to be closely related, but having some differences from Types 1 and 2. In short, the 4 dengue viruses seem to form two clusters.

Zika virus and *West Nile virus* show signs of having close similarities as well. A plausible biological explanation for this may be the fact that both *Zika virus* and *West Nile virus* are believed to originate from the geographical region in or near Uganda (Central East Africa). *Zika virus* was first discovered in the Zika Forest of Uganda in 1947 (Schwartz, 2016). *West Nile virus* was also originally discovered in Uganda in the year 1937 (Johnston and Conly, 2000), in the West Nile area, which gives rise to its name.

4.0.2. Ebolaviruses

Our results are consistent (though not identical) with Zheng et al. (2015), which employs the reliable and highly successful alignment-free natural vector method (Yu et al., 2013).

In particular, our results also have the following properties shared by Zheng et al. (2015): The five species of the Ebolaviruses are separated well (EBOV, SUDV, RESTV, BDBV, TAFV). The viruses from the same country are (generally, with a few exceptions) classified together within each species. The MARV and EBOV genomes are closer than the LLOV and EBOV genomes. In the branch of EBOV, majority (seven out of eight) of the viruses from Guinea and Sierra Leone of the 2014 outbreak are separated from others. *Bundibugyo ebolavirus* and *Taï Forest ebolavirus* are in the same group.

It was considered unusual by Zheng et al. (2015) that *Bundibugyo ebolavirus* and *Taï Forest ebolavirus* are in the same group according to their classification using natural vectors. This is because *Bundibugyo ebolavirus* is a deadly species while the *Taï Forest ebolavirus* is not so deadly. Our results confirm that, though unusual, it seems that the two viruses are indeed closely related. Hence, it appears that the virulence of ebolaviruses may not directly correlate with their genetic proximity.

Our results show that the Reston virus (RESTV) is most closely related to the Sudan virus (SUDV). This result is consistent with previous known phylogenetic analyses (Cantoni et al., 2016).

Our results also show a close similarity between ebolaviruses from the 2014 West Africa Ebola outbreak and ebolaviruses from Central Africa. In particular, the virus KM233096 from Sierra Leone is classified in the same group as HQ613403 from the Democratic Republic of the Congo. This is consistent with the general consensus among experts that the 2014 West African virus likely spread from Central Africa within the past decade (Gire et al., 2014; Alexander et al., 2015).

According to Alexander et al. (2015), the outbreak in Sierra Leone is believed to have started from the introduction of two genetically different viruses from Guinea. Our results are consistent with this, since the virus KM233096 is in a distinct group from the other 4 viruses from Mano River in Sierra Leone.

5. Further Improvements

We outline some further improvements that can be made to our approach. Some of the improvements may lead to increased computational costs.

5.1. Overlapping windows

Instead of non-overlapping sliding windows, overlapping windows can be used to increase robustness against frameshift errors in genome sequencing.

5.2. Increasing window size

Increasing the window size (increasing k for the k-mers) also has the benefit of increasing robustness against frameshift errors. This is because with a larger window, subsequences will be more likely to stay in the window despite frameshifts, hence minimizing the effect on the mean Magnus vector.

An increase in the window size also allows the Magnus Vector to capture more higher-order information in the form of longer subsequences.

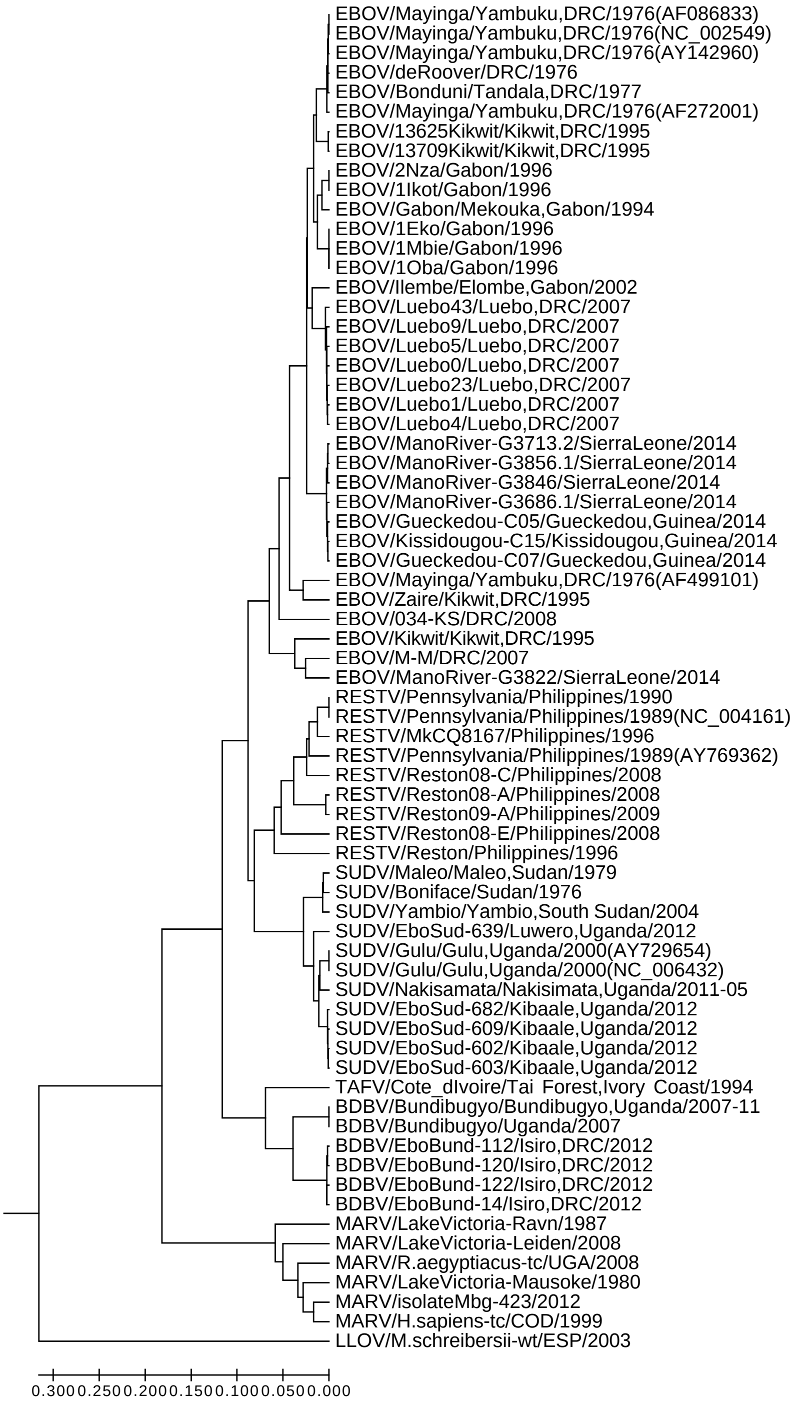
EBOV/Mayinga/Yambuku,DRC/1976(AF086833)
EBOV/Mayinga/Yambuku,DRC/1976(NC_002549)
EBOV/Mayinga/Yambuku,DRC/1976(AY142960)
EBOV/deRoover/DRC/1976
EBOV/Bonduni/Tandala,DRC/1977
EBOV/Mayinga/Yambuku,DRC/1976(AF272001)
EBOV/13625Kikwit/Kikwit,DRC/1995
EBOV/13709Kikwit/Kikwit,DRC/1995
EBOV/2Nza/Gabon/1996
EBOV/1Ikot/Gabon/1996
EBOV/Gabon/Mekouka,Gabon/1994
EBOV/1Eko/Gabon/1996
EBOV/1Mbie/Gabon/1996
EBOV/1Oba/Gabon/1996
EBOV/Ilembe/Elombe,Gabon/2002
EBOV/Luebo43/Luebo,DRC/2007
EBOV/Luebo9/Luebo,DRC/2007
EBOV/Luebo5/Luebo,DRC/2007
EBOV/Luebo0/Luebo,DRC/2007
EBOV/Luebo23/Luebo,DRC/2007
EBOV/Luebo1/Luebo,DRC/2007
EBOV/Luebo4/Luebo,DRC/2007
EBOV/ManoRiver-G3713.2/SierraLeone/2014
EBOV/ManoRiver-G3856.1/SierraLeone/2014
EBOV/ManoRiver-G3846/SierraLeone/2014
EBOV/ManoRiver-G3686.1/SierraLeone/2014
EBOV/Gueckedou-C05/Gueckedou,Guinea/2014
EBOV/Kissidougou-C15/Kissidougou,Guinea/2014
EBOV/Gueckedou-C07/Gueckedou,Guinea/2014
EBOV/Mayinga/Yambuku,DRC/1976(AF499101)
EBOV/Zaire/Kikwit,DRC/1995
EBOV/034-KS/DRC/2008
EBOV/Kikwit/Kikwit,DRC/1995
EBOV/M-M/DRC/2007
EBOV/ManoRiver-G3822/SierraLeone/2014
RESTV/Pennsylvania/Philippines/1990
RESTV/Pennsylvania/Philippines/1989(NC_004161)
RESTV/MkCQ8167/Philippines/1996
RESTV/Pennsylvania/Philippines/1989(AY769362)
RESTV/Reston08-C/Philippines/2008
RESTV/Reston08-A/Philippines/2008
RESTV/Reston09-A/Philippines/2009
RESTV/Reston08-E/Philippines/2008
RESTV/Reston/Philippines/1996
SUDV/Maleo/Maleo,Sudan/1979
SUDV/Boniface/Sudan/1976
SUDV/Yambio/Yambio,South Sudan/2004
SUDV/EboSud-639/Luwero,Uganda/2012
SUDV/Gulu/Gulu,Uganda/2000(AY729654)
SUDV/Gulu/Gulu,Uganda/2000(NC_006432)
SUDV/Nakisamata/Nakisimata,Uganda/2011-05
SUDV/EboSud-682/Kibaale,Uganda/2012
SUDV/EboSud-609/Kibaale,Uganda/2012
SUDV/EboSud-602/Kibaale,Uganda/2012
SUDV/EboSud-603/Kibaale,Uganda/2012
TAFV/Cote_dIvoire/Tai Forest,Ivory Coast/1994
BDBV/Bundibugyo/Bundibugyo,Uganda/2007-11
BDBV/Bundibugyo/Uganda/2007
BDBV/EboBund-112/Isiro,DRC/2012
BDBV/EboBund-120/Isiro,DRC/2012
BDBV/EboBund-122/Isiro,DRC/2012
BDBV/EboBund-14/Isiro,DRC/2012
MARV/LakeVictoria-Ravn/1987
MARV/LakeVictoria-Leiden/2008
MARV/R.aegyptiacus-tc/UGA/2008
MARV/LakeVictoria-Mausoke/1980
MARV/isolateMbg-423/2012
MARV/H.sapiens-tc/COD/1999
LLOV/M.schreibersii-wt/ESP/2003
0.300
0.250
0.200
0.150
0.100
0.050
0.000

Figure 2: Phylogenetic tree of 69 filoviruses drawn using MEGA7 (UPGMA).

5.3. Weighted Magnus Vector

The Hadamard product (entrywise product) of a Magnus Vector $(m_1, \ldots, m_n)$ with a weight vector $(w_1, \ldots, w_n)$ produces a weighted Magnus Vector $(m_1 w_1, \ldots, m_n w_n)$. This can be used to emphasize certain subsequences by increasing their weight.

Acknowledgements

The project was supported in part by the Singapore Ministry of Education research grant (AcRF Tier 1 WBS No. R-146-000-222-112). The first author was supported in part by the President's Graduate Fellowship of National University of Singapore. The third author was supported by a grant (No. 11329101) of NSFC of China. The fourth author was supported by Nanyang Technological University Startup Grants M4081842, Singapore Ministry of Education Academic Research Fund Tier 1 RG126/16 and RG31/18, Tier 2 MOE2018-T2-1-033.

Disclosure Statement

No competing financial interests exist.

Appendix A. Supplementary data and proofs

References

Alexander, K. A., Sanderson, C. E., Marathe, M., Lewis, B. L., Rivers, C. M., Shaman, J., Drake, J. M., Lofgren, E., Dato, V. M., Eisenberg, M. C., et al., 2015. What factors might have led to the emergence of Ebola in West Africa? PLoS neglected tropical diseases 9 (6), e0003652.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2012. Genbank. Nucleic acids research 41 (D1), D36–D42.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., 2008. Genbank. Nucleic acids research 36 (Database issue), D25.

Cantoni, D., Hamlet, A., Michaelis, M., Wass, M. N., Rossman, J. S., 2016. Risks posed by Reston, the forgotten ebolavirus. mSphere 1 (6), e00322–16.

Dong, R., Zheng, H., Tian, K., Yau, S.-C., Mao, W., Yu, W., Yin, C., Yu, C., He, R. L., Yang, J., Yau, S. S., 2017. Virus database and online inquiry system based on natural vectors. Evolutionary Bioinformatics 13, 1176934317746667.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al., 2014. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. Science 345 (6202), 1369–1372.

Huang, H.-H., 2016. An ensemble distance measure of k-mer and natural vector for the phylogenetic analysis of multiple-segmented viruses. Journal of theoretical biology 398, 136–144.

Johnston, B. L., Conly, J. M., 2000. West Nile virus-where did it come from and where might it go? Canadian Journal of Infectious Diseases and Medical Microbiology 11 (4), 175–178.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., Phillippy, A. M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, gr–215087.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution 33 (7), 1870–1874.

Kwan, H. K., Arniker, S. B., 2009. Numerical representation of dna sequences. In: Electro/Information Technology, 2009. eit'09. IEEE International Conference on. IEEE, pp. 307–310.

Lyndon, R. C., Schupp, P. E., 2015. Combinatorial group theory. Springer.

Rizk, G., Lavenier, D., Chikhi, R., 2013. Dsk: k-mer counting with very low memory usage. Bioinformatics 29 (5), 652–653.

Schwartz, D. A., 2016. The origins and emergence of Zika virus, the newest TORCH infection: what's old is new again. Archives of pathology & laboratory medicine 141 (1), 18–25.

Tuiskunen Bäck, A., Lundkvist, Å., 2013. Dengue viruses–an overview. Infection ecology & epidemiology 3 (1), 19839.

Yu, C., Hernandez, T., Zheng, H., Yau, S.-C., Huang, H.-H., He, R. L., Yang, J., Yau, S. S.-T., 2013. Real time classification of viruses in 12 dimensions. PloS one 8 (5), e64328.

Zheng, H., Yin, C., Hoang, T., He, R. L., Yang, J., Yau, S. S.-T., 2015. Ebolavirus classification based on natural vectors. DNA and cell biology 34 (6), 418–428.

APPENDIX A

CHENGYUAN WU, SHIQUAN REN, JIE WU, AND KELIN XIA

1. Algebra of formal power series on non-commuting variables

Let F be a free group on letters $x_1, x_2, \cdots$. Let R be a commutative ring. Consider the group ring $R(F)$, which is an algebra over R. Consider the augmentation ideal filtration of $R(F)$, where $IF = \ker(\epsilon : R(F) \to R)$ with $\epsilon(g) = 1$ for $g \in F$. The augmentation ideal filtration is given by $I^nF = (IF) \cdot (IF) \dots (IF)$, the n-fold product of IF. Let $A(F) = \varprojlim R(F)/I^nF$, the inverse limit. Then, one can prove that $A(F)$ is the algebra of formal power series on non-commuting variables $x_1, x_2, \dots$ using the property that F is a free group. The mapping $F \to A(F)$ is exactly the Magnus representation. We will describe it in greater detail in the following paragraphs.

Definition 1.1 (cf. [1, p. 132])**.** The homomorphism $\epsilon : R(F) \to R$ given by

$$\epsilon(\sum_{g\in F} a_g g) = \sum_{g \in F} a_g$$

is called the *augmentation mapping* of $R(F)$ and its kernel, denoted by IF, is called the *augmentation ideal* of $R(F)$.

Proposition 1.2 (cf. [1, p. 133])**.** The set $\{g - 1 \mid g \in F, g \neq 1\}$ is a basis of IF over R.

Thus, we can write

$$IF = \left\{ \sum_{g\in F} a_g(g-1) \mid g \in F, g \neq 1, a_g \in R \right\},$$

where the sums are finite sums.

Proof. If $\alpha = \sum_{g\in F} a_g g$ belongs to IF, then $\epsilon(\alpha) = \sum_{g\in F} a_g = 0$. Hence, α can be expressed in the form:

$$\alpha = \sum_{g\in F} a_g g - \sum_{g\in F} a_g = \sum_{g\in F} a_g(g-1).$$

We note that $\epsilon(g-1) = 0$ so that $g - 1 \in IF$. Hence, this implies that $\{g - 1 \mid g \in F, g \neq 1\}$ is a generating set for IF over R. If $a_1(g_1 - 1) + \cdots + a_n(g_n - 1) = 0$ for distinct $g_i \neq 1$, then $a_1g_1 + \cdots + a_ng_n - (a_1 + \cdots + a_n)1 = 0$. Since $g_i \neq 1$, hence we have $a_1 = \cdots = a_n = 0$. This shows linear independence of the generating set. Thus, we have shown that $\{g - 1 \mid g \in F, g \neq 1\}$ is a basis for IF. □

Corollary 1.3. The augmentation ideal IF is generated by

$$T := \{x_1 - 1, x_2 - 1, \dots\} \cup \{x_1^{-1} - 1, x_2^{-1} - 1, \dots\}$$

as an R-algebra.

That is, every element $\alpha \in IF$ can be expressed as a polynomial with indeterminates in T and coefficients in R.

Proof. By Proposition 1.2, any element in IF is of the form $\alpha = \sum_{g\in F} a_g(g-1)$. Since F is a free group, any element $g \in F$ is a word in T. By applying repeatedly the identity

$$ab - 1 = (a-1)(b-1) + (a-1) + (b-1)$$

for $a, b \in F$, we see that $g-1$ is a polynomial with indeterminates in T and coefficients in R. Hence, it follows that the same is true for α.

Since $\epsilon(x_i - 1) = \epsilon(x_i^{-1} - 1) = 0$, it is clear that conversely, any polynomial with indeterminates in T and coefficients in R is in IF. □

Corollary 1.4. The augmentation ideal IF is generated by

$$S = \{x_1 - 1, x_2 - 1, \dots\}$$

as an ideal of $R(F)$, or equivalently, as an $R(F)$-submodule.

Proof. By Proposition 1.2, any element $\alpha \in F$ is of the form $\alpha = \sum_{g\in F} a_g(g-1)$, where $g \in F$ is a word in S.

By the identities

$$ab - 1 = a(b-1) + (a-1)$$

and

$$x_i^{-1} - 1 = -x_i^{-1}(x_i - 1),$$

it follows that $g-1$ is a finite linear combination of $\{x_1 - 1, x_2 - 1, \dots\}$ with coefficients in $R(F)$. Hence, the same is true for α.

Conversely, since $\epsilon(x_i - 1) = 0$ and ϵ is an R-algebra homomorphism, it is clear that any linear combination of $\{x_1 - 1, x_2 - 1, \dots\}$ with coefficients in $R(F)$ is in IF. □

Definition 1.5 (cf. [2, p. 651])**.** An *inverse system* in the category of R-algebras **R-Alg** consists of an ordered pair $\{M_i, \psi_i^j\}$, where $(M_i)_{i\in I}$ is a family of R-algebras indexed by a partially ordered set $(I, \preceq)$ and $(\psi_i^j : M_j \to M_i)_{i\preceq j}$ is a family of R-algebra homomorphisms, such that the following diagram commutes whenever $i \preceq j \preceq k$:

$$\begin{tikzcd}
M_k \arrow[r, "\psi_i^k"] \arrow[dr, "\psi_j^k"'] & M_i \\
& M_j \arrow[u, "\psi_i^j"']
\end{tikzcd}$$

Proposition 1.6 (cf. [2, p. 652])**.** For $m \geq n$, define $\psi_n^m : R(F)/I^mF \to R(F)/I^nF$ by

$$\psi_n^m : \alpha + I^mF \mapsto \alpha + I^nF.$$

Then, $\{R(F)/I^nF, \psi_n^m\}$ is an inverse system over $\mathbb{N}$.

Proof. IF is the kernel of ϵ and thus a subalgebra of $R(F)$. Each I^nF is also a subalgebra and there is a decreasing filtration

$$R(F) \supseteq IF \supseteq I^2F \supseteq I^3F \supseteq \cdots.$$

Since $I^mF \subseteq I^nF$ for $m \geq n$, the maps ψ_n^m are well-defined. For $\alpha + I^kF \in R(F)/I^kF$, we have

$$\psi_i^k(\alpha + I^kF) = \alpha + I^iF = \psi_i^j\psi_j^k(\alpha + I^kF).$$

Hence, $\{R(F)/I^nF, \psi_n^m\}$ is an inverse system over $\mathbb{N}$. □

The categorical definition for inverse limit is stated in [2, p. 653]. The inverse limit is unique up to isomorphism, if it exists.

Proposition 1.7 (cf. [2, p. 669])**.** The inverse limit of an inverse system $\{M_i, \psi_i^j\}$ of R-algebras over a partially ordered index set I exists.

In particular,

$$\varprojlim M_i \cong \{(m_i) \in \prod M_i \mid m_i = \psi_i^j(m_j) \text{ whenever } i \preceq j\}.$$

Proof. The proof is similar to that of [2, p. 669]. □

Theorem 1.8. Let

$$A(F) = \varprojlim_{n\in\mathbb{N}} R(F)/I^nF,$$

where F is a free group with free generating set $\{x_1, x_2, \ldots\}$.

Then,

$$A(F) \cong R[[x_1, x_2, \ldots]],$$

the algebra of formal power series on non-commuting variables $x_1, x_2, \cdots$.

Proof. First, we observe that a change of variables $y_i - 1 = x_i$ induces an isomorphism

$$R[[x_1, x_2, \ldots]] \cong R[[x_1 - 1, x_2 - 1, \ldots]].$$

We can further view $R[[x_1 - 1, x_2 - 1, \ldots]]$ as the inverse limit

$$\varprojlim_{n\in\mathbb{N}} R[x_1, x_2, \ldots]/(x_1 - 1, x_2 - 1, \ldots)^n,$$

where $(x_1 - 1, x_2 - 1, \ldots)$ denotes the ideal of $R[x_1, x_2, \ldots]$ generated by $\{x_1 - 1, x_2 - 1, \ldots\}$.

Alternatively, we can view $R[[x_1 - 1, x_2 - 1, \ldots]]$ as the inverse limit

$$\varprojlim_{n\in\mathbb{N}} R[[x_1, x_2, \ldots]]/J^n,$$

where J denotes the ideal of $R[[x_1, x_2, \ldots]]$ generated by $\{x_1 - 1, x_2 - 1, \ldots\}$.

Consider the homomorphism $\phi : R(F) \to R[[x_1 - 1, x_2 - 1, \ldots]]$ defined on the generators of F by $x_i \mapsto x_i$ and extending to $R(F)$. The map is well-defined because

$$x_i^{-1} = \frac{1}{1 - (1 - x_i)} = \sum_{k=0}^{\infty} (-1)^k (x_i - 1)^k$$

lies in $R[[x_1 - 1, x_2 - 1, \ldots]]$.

Let $\alpha \in IF$. By Corollary 1.3, α can be expressed as a polynomial with indeterminates in

$$T = \{x_1 - 1, x_2 - 1, \ldots\} \cup \{x_1^{-1} - 1, x_2^{-1} - 1, \ldots\}$$

and coefficients in R. By the identity

$$x_i^{-1} - 1 = -x_i^{-1}(x_i - 1),$$

we see that $\phi(\alpha) \in J$. Hence, we have $\phi(IF) \subseteq J$. Similarly, we have $\phi(I^nF) \subseteq J^n$, for all $n \in \mathbb{N}$.

Hence, the homomorphism $\theta : R(F)/I^nF \to R[[x_1, x_2, \ldots]]/J^n$ defined by $\theta(\alpha + I^nF) = \alpha + J^n$ is well-defined. The homomorphism then extends to $\widetilde{\theta} : A(F) \to R[[x_1 - 1, x_2 - 1, \ldots]]$.

By Corollary 1.4, we see that $(x_1 - 1, x_2 - 1, \ldots)^n \subseteq I^nF$. Hence, the map $\psi : R[x_1, x_2, \ldots]/(x_1 - 1, x_2 - 1, \ldots)^n \to R(F)/I^nF$ defined by

$$\psi(\alpha + (x_1 - 1, x_2 - 1, \ldots)^n) = \alpha + I^nF$$

is well-defined. Subsequently, ψ extends to $\widetilde{\psi} : R[[x_1 - 1, x_2 - 1, \ldots]] \to A(F)$ which is the inverse of $\widetilde{\theta}$. Hence, $\widetilde{\theta}$ is an isomorphism, and this completes the proof. □

We then have the following corollary.

Corollary 1.9. The mapping $F \to A(F)$ defined on the generators by $x_i \mapsto 1 + x_i$ and extended to F is exactly the Magnus representation. □

References

1. César Polcino Milies and Sudarshan K. Sehgal, *An introduction to group rings*, vol. 1, Springer Science & Business Media, 2002.
2. Joseph J. Rotman, *Advanced modern algebra: Part 1*, vol. 165, American Mathematical Soc., 2015.

Department of Mathematics, National University of Singapore, Singapore 119076
E-mail address: wuchengyuan@u.nus.edu

Department of Mathematics, National University of Singapore, Singapore 119076
E-mail address: sren@u.nus.edu

Department of Mathematics, National University of Singapore, Singapore 119076
E-mail address: matwuj@nus.edu.sg

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371
E-mail address: xiakelin@ntu.edu.sg