Analytics Between Aligned Ancestral Bat to Microbat and Simulated Microbat Sequence Generated using EvoLSTM

Student: Soon Jae Jo (260738557)

Supervisor: Prof. Mathieu Blanchette

Co-Supervisor: DongJoon Lim

1 Introduction and Background

Random mutations are the basis of evolution. With even the earliest forms of bioinformatics, the study of probability on different types of mutations were deemed critical in understanding phylogenetic, biological, and evolutionary inferences. Better understanding of this subject allowed for inquiring important questions like in frameworks of sequence alignment and enhanced the advancements in simulating sequence evolution. This allowed for reaching new levels of understanding in reconstruction of ancestral genomes of many unavailable species (Blanchette *et al., 2004b*).

Although the probability of different mutations has been long understood to be dependent its' type, it has also been shown to be related to the region or nucleotides around at which it occurs (Arenas, 2015). The most known context dependent mutation is the C to T substitution with CpG island causing deamination of methylated cytosine to convert into thymine (Bird, 1980). Another example of it is in plants where different flanking nucleotide pattern of alternating pyrimidine-purine and purine-pyrimidine caused highly elevated mutation rates (Sung *et al.,* 2017). Not only does the context dependencies of flanking nucleotides affect mutations but also the situation or region at which it occurs in. For example, transcription coupled repair machinery exhibits high localization in gene rich region of the human genome (Surrallés *et al.,* 2002). Furthermore, mutations of different types can also be linked to context dependency. DNA polymerase slippage was observed to be linked with insertions and deletions rates (Messer and Arndt, 2007).

Many of classical sequence alignment algorithms such as BLAST (Altschul *et al.,* 1990), Needleman-Wunsch (Needleman and Wunsch, 1970) are sequence alignment methods that operate independently of context. Although these models are more effective and more efficient for sequences that are less diverged, that is not the case when considering the opposite. In the cases for aligning highly diverged sequences, having the right model for substitution, indels and other types of mutations are of high importance as it allows greater alignment accuracy and detecting remote homologies.

Especially for whole genome alignment these factors are of key in generating the most accurate alignment (Blanchette *et al.,* 2004a).

There are numerous models that have been proposed to calculate and understand the effect of context-dependencies. One of the earlier models, suggested a Markov chain Monte Carlo model which can be used to see the rate of substitutions, with respects to its flanking regions in the coding sequence and used the maximum likelihood to analyze its affects in the non-coding regions (Jenson and Pederson, 2000). This showed the effect of context dependencies in both the exon and the intron of the genome. Using this idea, a Bayesian network model was designed using the parameters of the flanking bases for ancestral genome reconstruction (Chachick and Tanay, 2012). However, the methods proposed have limitations, as the context dependencies it uses to calculate the mutation rates were limited to only one flanking nucleotide. The reason for this was the high computational cost and its exponential growth with larger sample size and context size. Although there were models that have been proposed to overcome these barriers with methods that takes parameters of larger flanking regions like with 5-mers or 7-mers, (Aggarwala and Voigt, 2016) (Zhu *et al*., 2017), it has been seen that for organisms as complex as mammals, the mutations are dependent not only on its neighboring flanking regions but also the relations it has with regions that are much further away (Aikens *et al.*, 2019).

EvoLSTM is an attempt to overcome these boundaries of probabilistic context dependent mutations using machine learning techniques. Its' goal is to provide a solution to context dependent mutations with larger flanking region and to provide a more accurate probabilistic model for different types of mutations.

In comparison to primates, bats are somatically very different. In general, even with other mammals, bats have some of the most unique adaptations and features. Although they possess the smallest genome across all extent of mammals, they also contribute to approximately twenty percent of diversity in mammals (Teeling *et al.,* 2018).

This is a continuation of an existing project. The overall objective is to expand further on the understanding of context dependent mutations using the data from evoLSTM to gain more insight. Previously, evoLSTM was used mainly between ancestral primates and humans. This research will try to provide more information with respect to species outside of primates, specifically using Microbat and its ancestor.

2 Method

2.1 EvoLSTM and its methodologies

EvoLSTM uses Long-term Short-term Memory, a machine learning technique modified to be used in a context-dependent probabilistic model. The general step is to first, pre-process the aligned ancestral/descendent sequence. Then using that data to

train the model, and finally, using the model to simulate evolution. The architecture and detailed methodology of EvoLSTM's can be found in Lim and Blanchette (2020)'s paper.

We used the data generated from the aligned sequence of David's Myotis bat (Myotis *davidii)* and its descendent Microbat *(Myotis luciferus)* to further investigate EvoLSTM's validity and its' effect on context-dependent mutations outside of primates.

2.2 Capturing K-mers of varying sizes in simulated and original aligned sequence

Then to analyze the differences of K-mers with respect to simulated Microbat sequence and the original aligned sequence, all possible combinations of K-mers were generated for varying sizes. For each size of K-mer, 4^k total combinations were generated using the nucleotides: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). Then for each unique combination of K-mer, its number of instances were counted in both the original aligned ancestral/descent and simulated sequences in a sliding manner. For example, in the sequence 'ATCG', for K-mer of size 2, one instance of 'AT', 'TC' and 'CG' were recorded. This was the chosen method to record the number of instances because it is the same way at which EvoLSTM was with its' sequences. The ratio of every K-mer in original aligned sequence versus simulated sequence were also recorded.

2.3 Capturing dinucleotide pattern in varying K-mer sizes from simulated and original aligned sequence

For each K-mers of size greater than 3, we separated the dataset into two groups. Those that contain a specific dinucleotide pattern, those that do not, for all possible combinations. And for each datapoint, we calculated the absolute distance from 1 using the ratio for each K-mer. Then, for each pair of datasets, one of K-mers capturing dinucleotide pattern and the other for not capturing the dinucleotide pattern, the z-test was performed. This method was used on K-mers up to size 7 as anything above had too little number of instances for each unique K-mer, thus its statistical value was too deviating to conclude any meaningful findings.

3 Results

3.1 Ratio of K-mers between original aligned sequence of David's bat and microbat and simulated sequence

For varying K-mer sizes, there are some discrepancy of K-mer ratios between the original sequence to simulated sequence. This discrepancy of K-mer ratios were more extreme with higher K-mer sizes (Fig 1.). This was expected as with K-mers of larger sizes, there are going to be less instances at which they are observed in the two sequences. Furthermore, the simulation of EvoLSTM does not generate a sequence at which perfectly aligns with the original sequence. Finally, the context-dependent mutation probabilities at which EvoLSTM considers when simulating, is likely going to cause a greater deviation with increasing sizes of K-mer.

In general, for most K-mers, it was observed that the ratio was close to one, meaning that in both the original sequence and the simulated sequence, the number of its' instances did not vary. However, in special cases, like with K-mers containing 'CG' dinucleotide pattern, there was a significant deviation of its' ratio from one (Fig 1.).

3.2 Absolute distance of ratios from one

One was specifically chosen as the number to subtract the ratio to as that is the expected value of the ratio given that there is no discrepancy in the analyzed sequences. Fig 2. shows that the average distance from one of varying K-mer sizes has the greatest contrast depending on whether the dinucleotide pattern of 'CG' is contained or not contained in its K-mer sequence. Because of its' location on the right most side, it shows that when 'CG' is contained in the K-mer sequence, there is a much higher discrepancy in original versus simulated sequence ratio in comparison to K-mers that do not contain the 'CG' dinucleotide pattern. Inversely, for K-mers containing the pattern 'AG' and 'TG' its' mean of ratio is very close to zero, which shows that the magnitude of mutations that occurred for those dinucleotides in comparison to others were very little.

Z-test for two sample means were conducted on these datasets for absolute distance from one. For confidence level of 95% (with alpha as 0.05), the only pair of datasets that rejected the null hypothesis were for K-mers of sizes 3 to 7 with respect to dinucleotide 'CG' (Supplementary Data 1). This further solidified that the presence of dinucleotide 'CG' plays a much more dominant role in mutation in comparison to other dinucleotide presence. Z-test was chosen versus that of the t-test as the sample sizes were large (greater than 30). K-mers with length greater than 7 were not chosen to be interpreted as the number of instances in which they are found in the sequences were too little to present conclusive interpretations.

4 Discussion and Conclusion

The variation of K-mer frequencies with dinucleotide 'CG' can be interpreted as the effect of CpG sites. In CpG sites of mammals, 70 to 80% of cytosines are methylated (Jabbari and Bernardi, 2004). Methylated cytosines have a much greater chance of substitution in comparison to that of a regular cytosine (Walsh and Xu, 2006). This shows that EvoLSTM can simulate and represent the CpG context dependency. Furthermore, it was shown by Hwang and Green (2004) that substitution from nucleotides G or C to A or T have a higher mutagenesis rate than the reverse. With this understanding, the presence of 'AG' and 'TG' nucleotide having less chances at mutations can be explained as, C or G substitution to A and T and out of C and G, C is more likely to go under mutagenesis with context dependencies, thus the combination of AG and TG dinucleotide presence can explain the lower mutagenesis.

Although many models have been proposed to computationally model and characterize context dependencies like, by using hidden Markov Models (Siepel and Haussler, 2003), they were unavailable to go beyond the limits of small flanking regions.

Using EvoLSTM, we were able to generate a simulated genome result which makes predictions on mutation probabilities for each nucleotide in each position, where it takes into 14 flanking nucleotides into consideration (Lim and Blanchette, 2020). EvoLSTM gave more insight in this subject by generating a model that takes into consideration a more complex context dependent mutations.

Due to the limitations of classical alignment methods, more recent studies propose the use of complex mutation models. It has shown to be highly valuable in whole genome alignment in highly diverged sequences (Blanchette *et al.,* 2004a). Furthermore, with tools like whole-genome aligner (Earl *et al.,* 2014), that serves as a guideline for varieties of bioinformatics tools, having a good model with accurate context dependencies is crucial.

Further areas of research can be made on other types of mammals. Especially given that the bat sequence in comparison with other mammals is much smaller, observing different mammals will allow a better statistical analysis on larger K-mers. Another application can be made with observing different regions in the genome, like intron versus exon regions. Or specific cell types like regular breast genes versus breast cancer genes or an invasive breast cancer genes versus non-invasive cancer genes.

Advancement of machine learning techniques is only going to be ever more evolving. Within the area of bioinformatics, it has only touched the surface of its' impact. Studies using EvoLSTM are colossal and is merely one of many ways that could have been taken to improve and integrate context dependent mutations into sequence evolution simulation, genomics, and other areas of biology.

Figures

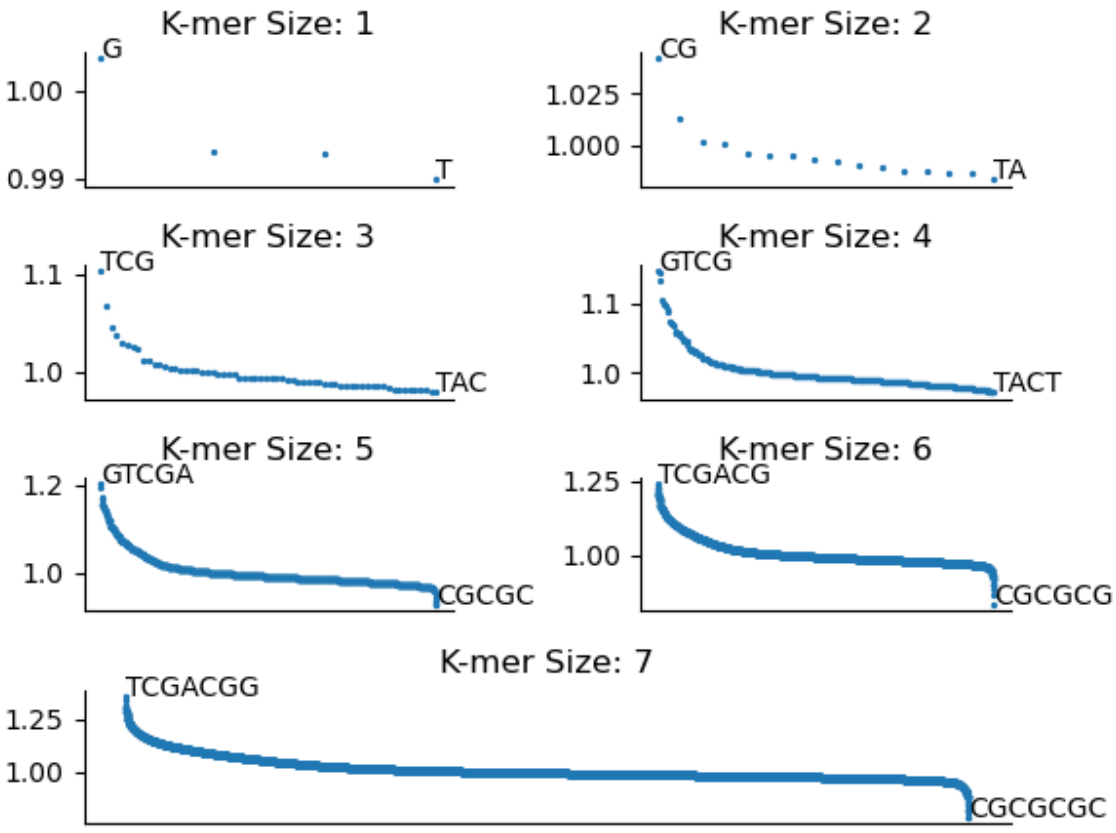

Fig. 1. Ratio between Aligned Ancestral Bat/Microbat to Simulated Microbat Sequence in Varying Sizes of K-mers. The x-axis is the arrangement of from highest ratio value to lowest ratio value. The y-values represent the ratio between the aligned sequence of ancestor/descendent sequence and simulated microbat sequence.

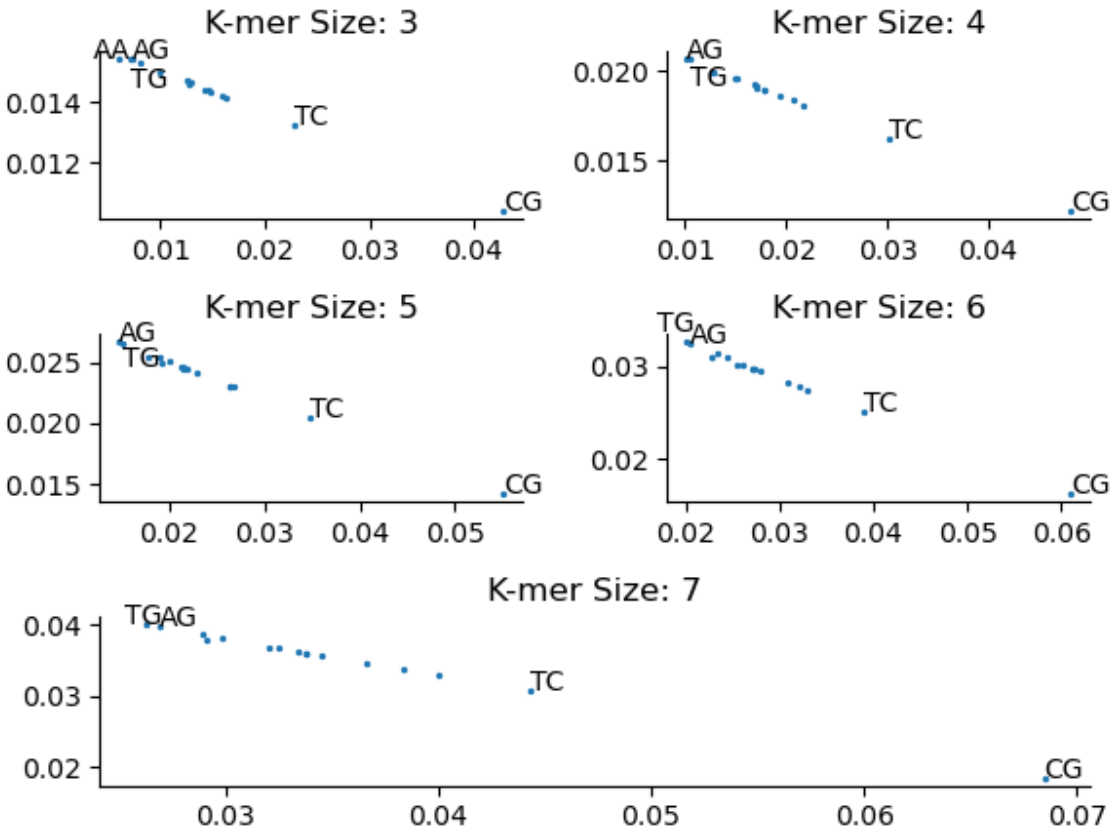

Fig. 2. The absolute average distance from 1 of which contains and do not contain specific dinucleotide pattern in varying K-mer sizes. The x-coordinate represents the average distance from 1 of K-mers which contains the specific dinucleotide pattern, and the y-coordinate represents the average distance from 1 of K-mers which do not contains the specific dinucleotide pattern.

Bibliography

Aggarwala, V. and Voight, B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.,* 48, 349–355.

Altschul, S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., 215, 403–410.

Arenas, M. (2015) Trends in substitution models of molecular evolution. *Front. Genet.*, 6, 319

Bird, A.P. (1980) DNA methylation and the frequency of CpG in animal DNA.

Nucleic Acids Res., 8, 1499–1504.

Blanchette, M. *et al*. (2004a) Aligning multiple genomic sequences with the threaded blockset aligner*. Genome Res*., 14, 708–715.

Blanchette, M. *et al*. (2004b) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.,* 14, 2412–2423.

Chachick, R. and Tanay, A. (2012) Inferring divergence of context-dependent substitution rates in drosophila genomes with applications to comparative genomics. *Mol. Biol. Evol., 2*9, 1769–1780.

Earl, D. et al. (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res*., 24, 2077–2089.

Hwang DG, Green P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA*.

Jabbari K., Bernardi, G. (2004). Cytosine methylation and CpG, TpG (CpA) and TpA frequencies, *Gene, Volume* 333, Pages 143-149.

Jensen, J.L. and Pedersen, A.-M.K. (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.,* 32, 499–517.

Lim, D.J., and Blanchette, M. (2020) EvoLSTM: context-dependent models of sequence evolution using a sequence-to-sequence LSTM, *Bioinformatics*, *Volume 36, Issue Supplement_1, July 2020*, Pages i353–i361.

Messer, P.W. and Arndt, P.F. (2007) The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol. Biol. Evol.,* 24, 1190–1197.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48, 443–453.

Siepel, A. and Haussler, D. (2003) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood*. Mol. Biol. Evol., 21*, 468–488.

Sung, W. *et al.* (2015) Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments, *Molecular Biology and Evolution*, Volume 32, Issue 7, Pages 1672–1683.

Surrallés, J. *et al*. (2002) Clusters of transcription-coupled repair in the human

genome. *Proc. Natl. Acad. Sci. USA*, 99, 10571–10574.

Teeling, E. *et al.* (2018) Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Bat1K Consortium. Annual Review of Animal Biosciences,* 23-46

Walsh C.P., Xu G.L. (2006) Cytosine Methylation and DNA Repair. In: Doerfler W., Böhm P. (eds) DNA Methylation: Basic Mechanisms*. Current Topics in Microbiology and Immunology, vol 301*.

Supplementary Data: Supplementary data is attached in submission.