B565-Data Mining
Homework #2

Due on Wednesday, Feb 15, 2023 08:00 p.m.

Dr. H. Kurban

Sarthak Mishra, Yu Mo, Aazin Asif Shaikh, Renu Jaiswal

March 20, 2023

Re-investigating the Molecular Evidence of Muller's Ratchet with lager data-set and different statistic [1]

Group Members

1. Yu Mo - moyu@iu.edu
2. Sarthak Mishra - samishr@iu.edu
3. Aazin Asif Shaikh - aazshaik@iu.edu
4. Renu Jaiswal - rjaiswal@iu.edu

Definition and background

There are four fundamental forces influencing evolution. These are Selection, Mutation, Recombination and Genetic Drift. Any evolutionary change can be attributed to the combination of these four forces. To understand the biological view of this project we have to delve into the definitions and types of these fundamental forces.

Selection

Selection is a process by which a living organism adapts to environmental changes. There are many kinds of selections. Some of them are positive selection, purging selection, Balancing selection, etc [19, 4]. This project will require knowing the definition of Positive and Purging selection. Some selection procedure are labeled as positive and some are labeled as negative. To quantify the positivity and negativity we need a measure. Fitness is one of the measures that shows the effect of selection and acts as the basis of the dichotomy.

Fitness

It is defined as the measure of individual reproductive success within a population. In other words, the probability of an individual from a population to leave an offspring. Fitness can be defined as 1- selection [4].

Positive selection

When the advantageous variation sweeps through the population causing their overall fitness to increase, such selection is called positive selection [19].

Purifying selection

When certain traits or genetic variations that negatively affect the overall fitness of the population are removed, then such selection is called purifying selection [4, 19].

Mutation

Mutation is a change in DNA sequence. Some mutations positively affect the fitness of organisms while some of them are deleterious. There are different regions of gene. Some of them are responsible for coding protein

[1] This is a research project, suggested by Dr. Matthew Hahn

and some are not. Hence mutation can be classified into two groups based on the location of mutation. Generally, the mutations on the protein-coding region are considered to be deleterious mutations [1]. These two kinds of mutations are of our concern.

Non-synonymous mutations

When the mutation occurs in coding region of the gene the mutation will alter the protein sequence. Any change in amino acids are considered harmful thus non-synonymous mutation is considered harmful/deleterious [5].

Synonymous mutations

When the mutations occur in the noncoding region of the gene the mutation will not alter the protein sequence. These genes are considered inconsequential/neutral [5].

Recombination

The rearrangement of DNA is called genetic recombination. It increases the genetic variation in the population [4].

Genetic Drift

Change in allele frequency of population across generations due to stochastic sampling is called Genetic Drift. Genetic drift generally decreases genetic variation [4, 2]. Genetic drift is modeled using two major stochastic process: Wright-Fisher Model and Moran Process. To understating neutral theory and the overall project, we will have to understand basic terms of Wright-Fisher model such as Fixation.

Wright-Fisher Model [14]

It is stochastic fluctuations of allele frequencies due to random sampling in finite populations.
Fixation Probability: When a random sampling takes place in a population of two or more individual/allele the frequency of alleles drifts in different direction. For any allele after certain number of generation/sampling there a chance that the frequency of allele in the population will be one. In this case the frequency of allele will remain the same. This is termed as fixation. Depending upon the frequency distribution of allele in current state there is a chance that the site will fix on one of the allele in the future state. This probability is called fixation probability.

Neutral theory of evolution [9]

Neutral theory of evolution , first introduced by Kimura 1991, says that advantageous mutations have a higher probability of fixation than neutral mutations, and deleterious mutations have a lower probability of fixation. It therefore follows that sequences subject to positive selection evolve faster than neutral sites, whereas sequences subject to negative selection evolve more slowly. When there is no selection acting on a region it is called a neutral region. While screening for selection Neutral expectation is considered to be the null hypothesis.

Rate of Evolution [7, 20]

Rate of evolution is the molecular change that causes different Taxa to diverges from one another. Rate of evolution is a function of time since the most recent common ancestor and the molecular substitution rate. For the purpose of our study, we take the time since Most recent common ancestor as fixed and assume only the substitution rates are differing.

Substitution Rate [7, 20]

The substitution rate is the number of new mutation that gets added in each generation multiplied by the probability each new mutation reaches fixation. Substitution rate can be understood as the rate by which actual number of mutations are being accumulated when moving from one generation to another. When the mutation is neutral substitution rate is equal to mutation rate. But for non neutral cases the effect of selection implied by neutral hypothesis will deviate the substitution rate from mutation rate.

Models of DNA evolution example [7, 20]

There are several models describing the models of DNA evolution. Generally, it is given by a matrix that denotes a probability of a base changing to another base. Most simple example of DNA evolution is Jukes Cantor model (also known as JC69) [7, 20]. In this model each DNA base has an equal chance of changing into another base. For example, A has an equal chance of changing into C or G, or T.

	A	C	T	G
A	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
C	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
T	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
G	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Codon Substitution Model [7]

Similar to DNA substitution model, codons also have a substitution model. This denotes the probability of a codon changing into another codon. For our project we are trying to infer the rate of evolution of genes from two different regions based on synonymous sites and nonsynonymous sites. You can find more information on the inference of codon substitution model in parameter estimation section.

Summary statistic [10]

d_N: It denotes the divergence non-synonymous regions of a genome. It is very hard to infer actual substitution rate hence d_N is generally estimated by finding the distance between the genomes in that region. d_S: It denotes the divergence synonymous regions of a genome. It is very hard to infer actual substitution rate hence d_S is generally estimated by finding the distance between the genomes in that region. Because d_S is subjected to selection and d_N is not subjected to selection, the ratio of d_N/d_S is considered to be a very good summary statistics to measure the effect of selection. In other words if $d_N/d_S < 1$ it is considered to be negative selection. $d_N/d_S = 1$ is considered as Neutrality and $d_N/d_S > 1$ is considered to be positive selection.
You can find more information about summary statistic and its estimation in parameter estimation section.

Mullers Ratchet

As-sexual population can undergo extinction due to irreversible mutational degradation in the absence of segregation and recombination. This phenomenon is called Muller's Ratchet(Muller1964)[11].

Problem Statement

In the past there have been multiple efforts to detect the molecular evidence of Mullers Ratchet. Notably, Lynch1996 [11] has claimed to have found the molecular evidence of Mullers Ratchet by comparing the mitochondria tRNA with nuclear tRNA. But results of later study by Cooper et al. [6] implies quite contradictory results. We propose to re-examine the problem with a larger data-set and better statistics. In doing so we are planning to compare the summary statistics for synonymous regions of mitochondrial tRNA and with the tRNA from Nuclear genome.

Literature Review

1. Lynch 1996 [11] showed the molecular evidence of Mullers Ratchet in mitochondrial Genome (mtDNA). The paper compared the evolution of Transfer RNA (tRNA) in mitochondrial and nuclear genomes. It shows that the t-RNA has a higher substitution rate in Mitochondrial Genome compared to Nuclear Genome. The author has provided a variety of reasons for this difference including the physical arrangement of strands of tRNA and varying loop sizes in the structure of Mitochondrial tRNA and Nuclear tRNA.

2. Cooper et al. [6] used *Drosophila Melanogaster*'s mitochondrial genome and nuclear genome. They could not find the evidence of recombination in mitochondrial genome but found the neutrality index is weakly significantly different between the mitochondrial and nuclear loci. They have attributed this difference to a larger proportion of beneficial mutations in X-linked relative to autosomes. But they could not find any difference whatsoever in the neutrality index of Mitochondria and autosome.

3. Popadin et al. [17] found that mammalian mitochondrial genomes differ from the nuclear genomes by maternal inheritance in the absence of recombination, and a higher mutation rate. They found Mitochondrial accumulate at least 5-folds more deleterious mutations compared to nuclear genome. This causes irreversibly degradation leading to decrease of organisms fitness in assexual lineages with low effective population size. They reach this concluding by comparing Kn/Ks in mitochondrial and nuclear regions. Kn/Ks are summary statistics similar compared to Dn/Ds but they need very strong signal to detect selection [22] also they cannot distinguish between positive and negative non-synonymous substitution [18].

Relevance to Data Mining

To complete this project we will be using different tools and methods that fall under the realm of data mining. Firstly we will scan through the species tree to find the species which have short branch lengths. This will imply the set of species has very low expected gene discordance. This requires mining the data from different regions of Whole genome alignment. Secondly, we will separate the coding and noncoding regions and infer the codon substitution rate for the species tree. Both of these sections employ data mining algorithms including but not limited to parameter inference algorithms and data cleaning procedures.

Data

The data set that we will utilize has been published [8]. It contains aligned genomes of 120 mammalian species, including primates, rodents, carnivores, ungulates, and marine mammals. It chooses human as the reference species, and 119 non-human species are aligned with it. As a direct result of this, genes from non-human species are eliminated if they are absent from human genome. In the meanwhile, it may lead to in-dels in non-human species, located in the position where genes are present in the human genome only.
In the context of this study, our primary research interest will be the protein-coding and tRNA genes present

in mitochondrial and autosome (non-sex chromosomes) genomes. To estimate the size of data briefly, we use human as an example. The length of the mtDNA is more than 16,000 base pairs (bp), and it includes 13 protein-coding genes, 22 tRNAs, 2 ribosomal components, and little noncoding DNA. The length of nuclear genome is more than 3×10^9 bp, including approximately 20,000 protein-coding genes, and over 400 tRNA genes. The total number of genes in mtDNA and autosome vary. Therefore, we anticipate less than 13 protein-coding genes and no greater than 22 tRNAs in mtDNA, and more than 500 protein-coding genes and fewer than 500 tRNAs in the autosome [3].

Methods

Phylogeny reconstruction

Phylogeny refers to the evolutionary history of a collection of species, which is often shown by a tree-like diagram. The leaves represent the extant species, and internal nodes reflect hypothetical ancestors that produced descendant lineages. The topology refers to its branching structure, indicating the connections among species. The length of a branch represents the amount of sequence divergence or the time period covered by the branch. To construct a phylogeny and understand the evolutionary history, we follow the general procedures listed below. First, sequences that have been aligned for the species of interest are gathered. Sequences may consist of DNA or RNA sequences, amino acid sequences, or other sorts of molecular data. A substitution model is a probabilistic model using continuous-time Markov chains, describing the evolution of the sequences over time. Different models are proposed for various forms of molecular data, with various assumptions and parameters. Then, a phylogenetic tree is inferred with Maximum Likelihood (ML) or Bayesian inference, given sequences, and a suitable substitution model.

Due to the aligned techniques employed to construct the dataset, it is not suitable to include all species in subsequent analyses. There are few overlapping genes across all species, resulting in incredibly short sequences that invalidate our test. Primarily, we will concentrate on the *Primate, Gilres,* and *Laurasiatheria* clades, each of them with at least twenty species. Because species inside a clade are more closely related than those outside of a clade. We expect closely related species to have more overlapping genes. Another notable factor for selecting species is to avoid incomplete lineage sorting (ILS). ILS indicates the genealogical history of some genes within a group of species is not concordant with the species tree, leading to biased estimation [13]. Therefore, we would like to select a subset of species to eliminate the factors mentioned above. Specifically, we will construct the phylogeny of all species, with a nucleotide substitution model selected by IQ-TREE [16] that fits the alignments the best. Then, we calculate a summary statistic called site Concordant Factor (sCF) [15] for each branch. The sCF ranges from zero to a hundred and is used to assess the level of ILS. Intuitively, a smaller sCF implies fewer sites are concordant with the inferred branch, indicating greater ILS. We would thus eliminate branches with sCF less than 90. The remaining phylogeny is utilized to estimate parameters with ML.

Parameter estimation

With protein-coding genes, we have the ability to identify synonymous and nonsynonymous substitutions. As natural selection primarily functions at the level of proteins, both mutations are fixed at vastly different rates. Thus comparison of their rates provides a means to understanding the effect of natural selection on the protein level [21]. For protein-coding genes, codon substitution models are implemented. Codon is a triplet of nucleotides that encodes a specific amino acid or a stop signal in a protein. The codon substitution model describes the relation of codons transition instead of nucleotides. To be specific, we are more interested in synonymous substitution rates (d_S) in this project. To estimate d_S, we will utilize a codon substitution

model [23] with protein-coding regions in mtDNA and autosome respectively, using the sequences and the phylogeny mentioned as input.
As for the nonsynonymous substitution rate, we will focus on tRNAs in mtDNA and autosome, denoted as d_t to distinguish from d_N, meaning nonsynonymous substitution rate in protein-coding genes. We could only use nucleotide substitution model with tRNAs because there are few nonsynonymous sites in tRNAs. Therefore, we will estimate d_t with the substitution model selected by IQ-TREE [16], using tRNAs in mtDNA and autosome respectively.

Test of selection

Inspired by McDonald-Kreitman test [12] for neutrality, we will construct a two-by-two contingency table of d_t and d_S, as shown in Table 1. Assuming mtDNA has a higher evolutionary rate compared to autosome [11], and the relative evolutionary rate within mtDNA and autosome would be the same, our null hypothesis is that the ratio of d_t to d_S is identical for mtDNA and autosome. We will test for significant deviations from null hypothesis with the Fisher's exact tests.

	mtDNA	autosome
tRNA	d_t^{mt}	d_t^A
protein-coding regions	d_S^{mt}	d_S^A

Table 1: Estimate for the hypothesis test.

Challenge

Since our study relies significantly on the phylogenetic tree. Inaccurate results may emerge from a misspecified tree. There are multiple reasons for an incorrect tree. For example, the size of data would be a remarkable element. We exclude species with high ILS to avoid discordance, which may result in fewer species present in subsequent analyses. Also, the tree would be less convincing if few relevant genes are present in all selected species, resulting in short sequences. We will first analyze the data and then choose the optimum option for it. In the worst scenario, we would have to replace the dataset if it is not capable.

Computational Resource

With the size of the data we are handling, we expect the computations to be extensive and not feasible for our local machines. To mitigate this issue, we are planning to request Campus Computing Cluster. We are planning a request for a head node and a compute node.

Timeline and milestones

Date	**Milestone**	**People assigned**
02/22/2023	Knowledge transfer (understand data and methods)	Aazin, Renu, Yu, Sarthak
02/22/2023	Initial data processing	Yu, Aazin
03/01/2023	Data processing, creating/filtering from species trees	Yu, Aazin
03/08/2023	Run preliminary experiments	Sarthak, Renu
03/22/2023	Hypothesis testing	Sarthak, Renu
03/29/2023	Result summary	Sarthak, Renu
04/05/2023	Report conclusion and project report	Yu, Sarthak, Renu, Aazin

Individual tasks

The typical steps while involved while building a data mining project are:
1. Data selection
2. Data Pre-processing
3. Data transformation
4. Data mining
5. Interpretation/Evaluation

Similarly, our project involves several tasks such as:
1. Understand data and methods
2. Data processing
3. Creating/filtering from species trees
4. Hypothesis testing
5. Result summary and final report

To ensure that each of the above stated task is executed properly, the below table shows the tentative division of tasks assigned to each team member:

Sr. no	**Team Members**	**Tasks**
1.	Yu Mo	Data processing, creating/filtering from species trees, Project report
2.	Sarthak Mishra	Hypothesis testing, Result summary and methods, Report conclusion
3.	Aazin Asif Shaikh	Understand data and methods, creating/filtering from species trees, Data processing, Project report
4.	Renu Jaiswal	Understand data and methods, Hypothesis testing, Result summary, Project report

References

[1] Mutation genetics-glossary https://www.genome.gov/genetics-glossary/mutation.

[2] Genetic drift - understanding evolution https://evolution.berkeley.edu/evolution-101/mechanisms-the-processes-of-evolution/genetic-drift/, Feb 2022.

[3] Abbott, J. A., Francklyn, C. S., and Robey-Bond, S. M. Transfer rna and human disease. *Frontiers in genetics 5* (2014), 158.

[4] Alberts B, Johnson A, L. J. e. a. Molecular biology of the cell. *New York: Garland Science 4th edition* (2002).

[5] Chu, D., and Wei, L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer 19*, 1 (Apr 2019), 359.

[6] Cooper, B. S., Burrus, C. R., Ji, C., Hahn, M. W., and Montooth, K. L. Similar Efficacies of Selection Shape Mitochondrial and Nuclear Genes in Both Drosophila melanogaster and Homo sapiens. *G3 Genes—Genomes—Genetics 5*, 10 (10 2015), 2165–2176.

[7] Erickson, K. The jukes-cantor model of molecular evolution. *PRIMUS 20*, 5 (2010), 438–445.

[8] Hecker, N., and Hiller, M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *Gigascience 9*, 1 (2020), giz159.

[9] KIMURA, M. The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics 66*, 4 (1991), 367–386.

[10] Kryazhimskiy, S., and Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet 4*, 12 (Dec. 2008), e1000304.

[11] Lynch, M. Mutation accumulation in transfer RNAs: molecular evidence for muller's ratchet in mitochondrial genomes. *Mol Biol Evol 13*, 1 (Jan. 1996), 209–220.

[12] McDonald, J. H., and Kreitman, M. Adaptive protein evolution at the adh locus in drosophila. *Nature 351*, 6328 (1991), 652–654.

[13] Mendes, F. K., Fuentes-González, J. A., Schraiber, J. G., and Hahn, M. W. A multispecies coalescent model for quantitative traits. *eLife 7* (jul 2018), e36482.

[14] Messer, P. Neutral models of genetic drift and mutation. 119–123.

[15] Minh, B. Q., Hahn, M. W., and Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Molecular biology and evolution 37*, 9 (2020), 2727–2733.

[16] Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution 32*, 1 (2015), 268–274.

[17] Popadin, K. Y., Nikolaev, S. I., Junier, T., Baranova, M., and Antonarakis, S. E. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Mol Biol Evol 30*, 2 (Sept. 2012), 347–355.

[18] Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., and Feil, E. J. Comparisons of dn/ds are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology 239*, 2 (2006), 226–235. Special Issue in Memory of John Maynard Smith.

[19] Thomas, J. H. Positive selection. *Genome Sciences, University of Washington 1*, 1 (2009), 218.

[20] Warnow, T. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, 2017.

[21] Yang, Z. *Computational molecular evolution*. OUP Oxford, 2006.

[22] Yang, Z., and Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol 15*, 12 (Dec. 2000), 496–503.

[23] Yang, Z., and Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution 17*, 1 (2000), 32–43.