A simple test to uncover signals of CpG hypermutability using posterior predictive simulations

Simon Laurin-Lemay[1] and Nicolas Rodrigue[1,2,3]

[1]Department of Biology, Carleton University, Ottawa, Canada
[2]Institute of Biochemistry, Carleton University, Ottawa, Canada
[3]School of Mathematics and Statistics, Carleton University, Ottawa, Canada

Keywords

DNA sequence evolution, Markov chain Monte Carlo, phylogenetics, misspecification, confounding factors

Declarations

Conflict of interest: The authors declare that they have no competing interests.

Correspondence

Simon Laurin-Lemay
209 Nesbitt Biology Building,
1125 Colonel By Drive Ottawa,
Ontario, CANADA
K1A 0C6
evol.simon@gmail.com

Abstract

CpG hypermutability is caused by the spontaneous deamination of methylated cytosines within CpG contexts and is known to impact vertebrate evolution. The phenomenon has been shown to confound tests of selection in protein-coding genes. In this work, we propose a simple test based on the use of posterior predictive sampling to detect the presence of CpG hypermutability using common nucleotide substitution models. Artificial data sets were generated using a jump-chain simulation algorithm with a model incorporating varying levels of CpG hypermutability. On simulations using realistic parameter values, we recovered between 0-8% of false positives and no false negatives. We then ran the test on 137 mammalian gene alignments, all of which were found to exhibit CpG hypermutability, corroborating previous studies based on elaborate and computationally expensive Monte Carlo methods. For greater confidence in its results, any study aimed at detecting signals of selection could easily be accompanied by this simple test.

Introduction

All models, by definition, make simplifying assumptions about the phenomena they attempt to capture. One of the risks associated with these simplifications is that true features that are unaccounted for could mislead the model-based inferences being conducted. For example, the hypermutability of CpGs due to the process of spontaneous deamination of methylated cytosines in CpG contexts (Bird, 1980; Burge et al, 1992) can confound model-based detection of selection, potentially leading to erroneous conclusions regarding codon usage preferences (Laurin-Lemay et al, 2018a), or positive selection at the amino acid level (Laurin-Lemay et al, 2022).

Most substitution models assume that sites evolve independently, which is particularly convenient for calculating the phylogenetic likelihood function; one only need multiply all the site likelihoods, calculated independently from each site of the alignment (Felsenstein, 1981). On the other hand, implementing substitution processes taking into account dependencies between sites (e.g., Robinson et al, 2003; Rodrigue et al, 2009; Laurin-Lemay et al, 2018b; Meyer et al, 2019), as required to parameterize the hypermutability of CpGs, is not a trivial task. Some of the strategies employed have included nested Markov chain Monte Carlo sampling (e.g., Robinson et al, 2003; Rodrigue et al, 2009) and simulation-based Approximate Bayesian Computation (e.g., Laurin-Lemay et al, 2018b). In all cases, the methods are computationally elaborate and costly.

Short of explicitly modeling site-dependencies, simulation-based methods, such as parametric bootstrapping (Efron, 1979; Efron and Tibshirani, 1993) or posterior predictive sampling (Gelman et al, 1996; Brown and Thomson, 2018), can be utilized to uncover site-dependent features having the potential to mislead inferences. Here, we propose a simple posterior predictive test based on the widely used GTR+Γ model to uncover a statistical signal of CpG hypermutability in mammalian genes. In the presence of CpG hypermutability, we expect the GTR+Γ substitution model to predict more CpG dinucleotides than would be observed in real alignments, because CpG hypermutability is not accounted for in the model definition, and hence CpG dinucleotides do not become depleted when simulating the evolutionary process over the phylogeny.

CpG hypermutability test

The CpG test consists of calculating the proportion of times, i.e., *p-value*, that the frequency of CpGs calculated from the real alignment is greater than the CpG frequency recovered from the posterior predictive alignments, which are generated by simulation over a sample of parameter values drawn from their posterior distribution by Markov chain Monte Carlo under the GTR+Γ model. The following equation details the CpG test:

$$p\text{-}value = \frac{1}{N}\sum_{i=1}^{N}\delta(freq_{CpG}^{real} > freq_{CpG}^{pred_i}), \quad (1)$$

where N is the predictive sample size, $freq_{CpG}^{real}$ refers to the frequency of CpG dinucleotides in the real alignment, $freq_{CpG}^{pred_i}$ is the CpG dinucleotide frequency of the ith posterior predictive alignment, and δ is an indicator function that returns 1 if the CpG frequency calculated from the real alignment is greater than the predictive CpG frequency and 0 otherwise. With this simple test-statistic, a *p-value* close to 0 suggests a rejection of a null hypothesis of absence of a CpG hypermutability process governing the real data relative to the simulations. Crudely, a threshold ($p\text{-}value < 0.05$) can be chosen to flag a data set as a *positive* for potential CpG hypermutability, and *negative* otherwise.

Simulation study

We analyzed 10 mammalian protein-coding gene alignments selected for their wide range of GC content (Laurin-Lemay et al, 2018b). We used these alignments to obtain parameter values for simulations under the GTR+Γ substitution model, providing the simplest *negative controls*: synthetic data sets that have no signal of CpG hypermutability. For each gene, we used ten draws from the posterior distribution, approximated using PhyloBayes-MPI (Lartillot et al, 2013), as parameters for simulations under GTR+Γ, conducted with AliSim (Ly-Trong et al, 2022), thereby yielding 100 simulated alignments in total.

We next conducted simulations for *positive controls*: synthetic data sets produced under an evolutionary process that includes CpG hypermutability. In this case, we used the MG-F1×4 codon substitution model (see, e.g., Rodrigue et al, 2008, for a detailed description) within Phylobayes-MPI (Rodrigue and Lartillot, 2014). The latter provided parameter values for nucleotide frequencies and pairwise exchangeabilities, but for two key parameters of our codon-level simulations, we explored two different values each: the nonsynonymous/synonymous rate ratio, ω, was set to either 0.2 or 1, and a parameter controlling CpG hypermutability, λ (see Laurin-Lemay et al, 2018b, for a detailed description), was set to either 4 or 8, reflecting empirical values commonly observed on real data (Laurin-Lemay et al, 2018b). Note that another negative control can be obtained by setting $\lambda = 1$, which we also investigated. From the 10 sets of parameter values drawn for the posterior distribution under MG-F1×4 for each of the 10 alignments, with all combinations of ω and λ values, we thus simulated 600 synthetic alignments.

Each negative control and positive control alignment was then analyzed with the CpG hypermutability test. Less than 5% of the tests were significant when performed on synthetic sequence alignments generated under the GTR+Γ substitution model, i.e., 2% (Table 1). Similarly, none of the CpG tests were significant

when performed on synthetic sequence alignments generated without CpG hypermutability and with $\omega = 0.2$ using the MG-F1×4 codon substitution model, and 8% were significant with $\omega = 1$ (Table 1). On the other hand, in the presence of CpG hypermutability all tests were positive.

Empirical study

We retrieved the 137 mammalian codon alignments from Laurin-Lemay et al (2018a) along with the tree topology. We applied the posterior predictive CpG hypermutability test to each of them and found 100% of the tested genes to be significant (*p-value*= 0, $\alpha = 0.05$). In other words, CpG frequencies are significantly reduced in all real alignments compared to frequencies recovered from predicted alignments generated by the GTR+Γ substitution model. As an example, in Figure 1, panel A, we show the discrepancy between the CpG frequencies predicted by GTR+Γ and the observed CpG frequency in one of the examined coding sequence alignments. On the same alignment, panel B shows that the true value of the frequency of another dinucleotide, ApT, falls within the posterior predictive distribution. The test thus suggests that something is unaccounted for with regards to CpG frequencies in particular.

Discussion

The approach proposed herein is crude, with uncalibrated posterior predictive *p-values*. However, it is quick and easy to implement, and clearly highlights the presences of CpG hypermutability. Knowing that CpG hypermutability can mislead tests for selection (Laurin-Lemay et al, 2018a, 2022), this simple test could be systematically applied when conducting selection analyses (e.g., Kosiol et al, 2008; Murrell et al, 2015; Davydov et al, 2019; Slodkowicz and Goldman, 2020), at relatively low cost.

More generally, however, we propose that numerous simple test-statistics could be employed within posterior predictive (or parametric bootstrap) analyses. These include all possible dinucleotides, trinucleotides, codon boundary contexts, adjacent amino acids, average effective number of amino acids (or nucleotides, or codons) across positions, compositional heterogeneity measures of several types, and many more. This could help contextualize the interpretation of evolutionary studies, either giving greater confidence in their results, or emphasizing where one should be cautious due to glaring model violations.

Such tests would provide a technically simple means of guiding modeling efforts. Historically, model developers have chosen to focus on one or another hypothesized aspect of the evolutionary process to build within an existing modeling framework, without much of a systematic outline of the most prevalent or most pronounced model violations. In many cases, newly proposed models have been years in the making, requiring the invention or recombination of novel Monte Carlo approaches, the development of new and

complex computing software, self-validation simulation checks, testing, and so forth. To date, it is quite unclear if these modeling efforts have been undertaken in rational manner, in terms of order of priority.

Our simple and crude test of CpG hypermutability is an example of how, with the tools and models we have had for decades, we can bring forth complex features of the evolutionary process from any multiple sequence alignment. This calls for a much broader simulation-based outline of modeling objectives, both to retrospectively evaluate the sequence of modeling achievements to date, and plan more carefully for future work.

Availability of data and materials

Workflow and data are available on the GitHub repository `https://github.com/Simonll/CpG-ppred-test.git`.

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council of Canada.

Contributions

Conceptualization: SLL, NR. Formal analysis: SLL Investigation: SLL, NR. Methodology: SLL, NR. Resources: SLL, NR. Supervision: NR Writing original draft: SLL, NR. Review and editing: SLL, NR.
All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Figures

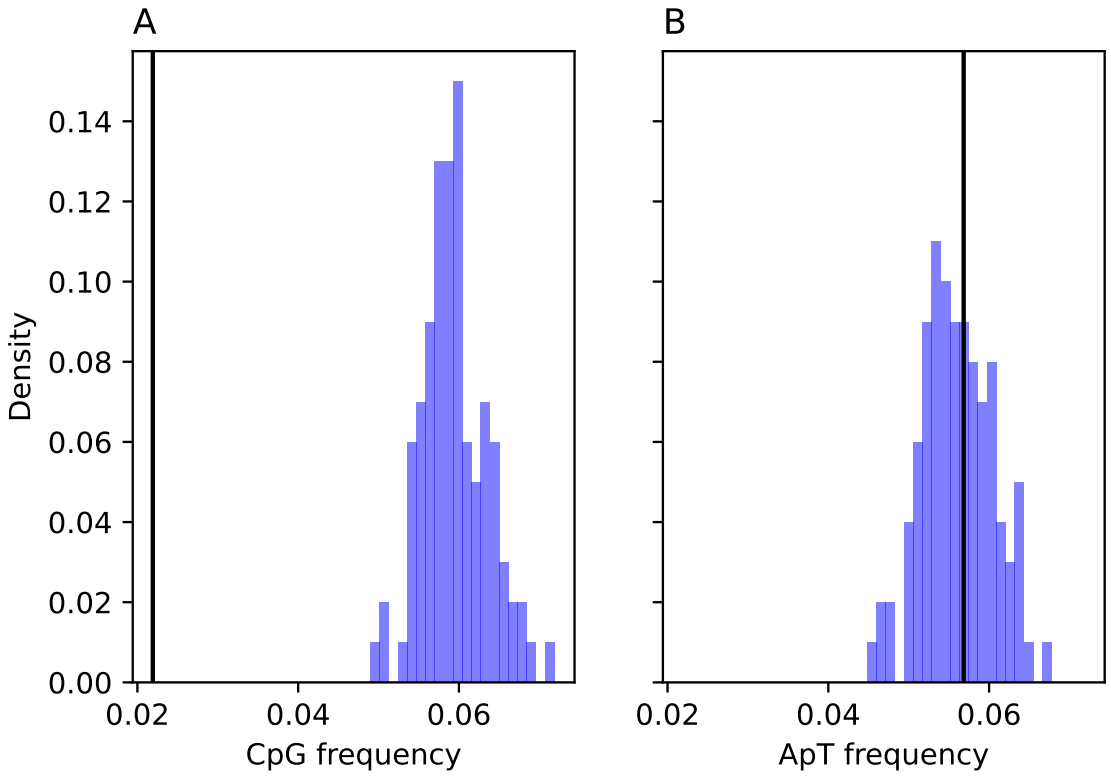

Figure 1: Comparison of CpG (A) and ApT (B) frequencies computed from the *MEP1A* mammalian protein-coding gene alignment, black vertical lines, and corresponding frequencies computed from posterior predictive alignments generated using the GTR+Γ substitution model, blue histograms. Note that the predicted CpG frequencies (A) overestimate the true value 100% of the time, while the true ApT frequency (B) falls within the distribution of predicted frequencies.

Tables

type of controls	models used to generate the synthetic alignments	positive tests (%)
negative	GTR+Γ	2
negative	MG-F1$\times$4+($\lambda = 1$)+($\omega = 0.2$)	0
negative	MG-F1$\times$4+($\lambda = 1$)+($\omega = 1.0$)	8
positive	MG-F1$\times$4+($\lambda = 4$)+($\omega = 0.2$)	100
positive	MG-F1$\times$4+($\lambda = 4$)+($\omega = 1.0$)	100
positive	MG-F1$\times$4+($\lambda = 8$)+($\omega = 0.2$)	100
positive	MG-F1$\times$4+($\lambda = 8$)+($\omega = 1.0$)	100

Table 1: Validation of the CpG test using posterior predictive sampling under the GTR+Γ substitution model with a α significance threshold of 5%. Synthetic alignments were generated using GTR+Γ and MG-F1$\times$4 codon substitution model using three CpG hypermutability values, $\lambda = \{1, 4, 8\}$ and global selection on amino acids using two ω values (i.e., 0.2 and 1).

References

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research 8(7):1499–1504, DOI 10.1093/nar/8.7.1499

Brown JM, Thomson RC (2018) Evaluating model performance in evolutionary biology. Annual Review of Ecology, Evolution, and Systematics 49(1):95–114, DOI 10.1146/annurev-ecolsys-110617-062249

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proceedings of the National Academy of Sciences 89(4):1358–1362, DOI 10.1073/pnas.89.4.1358

Davydov II, Salamin N, Robinson-Rechavi M (2019) Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. Molecular Biology and Evolution 36(6):1316–1332, DOI 10.1093/molbev/msz048

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7(1):1 – 26, DOI 10.1214/aos/1176344552

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall/CRC monographs on statistics and applied probability, Chapman and Hall, London

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17(6):368–376, DOI 10.1007/bf01734359

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6(4):733–760

Kosiol C, Vinař T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genetics 4(8):e1000,144, DOI 10.1371/journal.pgen.1000144

Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Systematic Biology 62(4):611–615, DOI 10.1093/sysbio/syt022

Laurin-Lemay S, Philippe H, Rodrigue N (2018a) Multiple factors confounding phylogenetic detection of selection on codon usage. Molecular Biology and Evolution 35(6):1463–1472, DOI 10.1093/molbev/msy047

Laurin-Lemay S, Rodrigue N, Lartillot N, Philippe H (2018b) Conditional approximate bayesian computation: A new approach for across-site dependency in high-dimensional mutation–selection models. Molecular Biology and Evolution 35(11):2819–2834, DOI 10.1093/molbev/msy173

Laurin-Lemay S, Dickson K, Rodrigue N (2022) Jump-chain simulation of markov substitution processes over phylogenies. Journal of Molecular Evolution 90(3-4):239–243, DOI 10.1007/s00239-022-10058-0

Ly-Trong N, Naser-Khdour S, Lanfear R, Minh BQ (2022) AliSim: A fast and versatile phylogenetic sequence simulator for the genomic era. Molecular Biology and Evolution 39(5), DOI 10.1093/molbev/msac092

Meyer X, Dib L, Silvestro D, Salamin N (2019) Simultaneous bayesian inference of phylogeny and molecular coevolution. Proceedings of the National Academy of Sciences 116(11):5027–5036, DOI 10.1073/pnas.1813836116

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Pond SLK (2015) Gene-wide identification of episodic selection. Molecular Biology and Evolution 32(5):1365–1371, DOI 10.1093/molbev/msv035

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003) Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution 20(10):1692–1704, DOI 10.1093/molbev/msg184

Rodrigue N, Lartillot N (2014) Site-heterogeneous mutation-selection models within the phylobayes-mpi package. Bioinformatics 30(7):1020–1021, DOI 10.1093/bioinformatics/btt729

Rodrigue N, Lartillot N, Philippe H (2008) Bayesian comparisons of codon substitution models. Genetics 180(3):1579–1591, DOI 10.1534/genetics.108.092254

Rodrigue N, Kleinman CL, Philippe H, Lartillot N (2009) Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Molecular Biology and Evolution 26(7):1663–1676, DOI 10.1093/molbev/msp078

Slodkowicz G, Goldman N (2020) Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. Proceedings of the National Academy of Sciences 117(11):5977–5986, DOI 10.1073/pnas.1916786117