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

Objective

CpG hypermutability is a process known to impact vertebrate evolution and is caused by spontaneous deamination of methylated cytosines within CpG contexts. CpG hypermutability has been shown to confound the detection of negative and positive selection on codon usage and amino acids. In this work, we propose a simple test based on the use of posterior predictive sampling to detect the presence of CpG hypermutability using the GTR+G substitution model.

Results

First, we validated the CpG test using simulations. We recovered between 0-8% of false positives and no false negatives. The simulations were performed using the jump-chain algorithm with a simple codon substitution model for more realistic simulation conditions and to allow for different degree of CpG hypermutability. All mammalian genes tested were positive with the posterior predictive test we developed to detect CpG hypermutability, which is consistent with previous results obtained using complex mechanistic phylogenetic codon substitution models designed to measure the level of CpG transition rate. This test could easily accompany any study aimed at detecting negative or positive selection, even for large-scale analyses, knowing that methylation of cytosines exists in all domains of life (i.e., bacteria, archaebacteria, and eukaryota).

Introduction

The phylogenetic method involves first the generation of data sets and then the analysis of those data sets with various substitution models. This is followed by a cycle of work in which molecular evolutionists compare measurements of population processes (i.e. mutation, selection, even demographic processes) included in the definitions of their models (e.g. **??**). One of the things most feared by molecular evolutionists is the discovery of confounding effects (e.g., **??**) which could invalidate the proposed mechanistic hypothesis testing approach to detect and measure levels of specific evolutionary processes. Confounders are evolutionary processes not accounted for by the substitution model. In the Bayesian framework, the presence of confounders will incorrectly affect the posterior values of one or more parameters of the substitution model, which was designed to parameterize another population process. Finally, false positives can result from the presence of confounders. For example, the hypermutability of CpGs due to the process of spontaneous deamination of methylated cytosines in CpG contexts may confound the detection of negative or positive selection on synonymous and non-synonymous mutations (e.g., **??**).

Once a confounding process has been identified, researchers should go back to the original data sets and try to control for that aspect, if possible, by removing the sites affected by the confounding process: (e.g., **?**). On the other hand, managing confounders by selecting sites for analysis may be particularly difficult or even impossible when dealing with confounders that affect substitution processes with dependencies between sites, such as the hypermutability of CpGs. Ultimately, the researcher should propose ways to incorporate the confounding process into a new substitution model, so that the process can be conditioned jointly with other evolutionary processes, or in a two-step procedure (e.g., topology is usually not co-conditioned with other mutation and selection

parameters in the field of molecular evolution).

Most substitution models assume that sites evolve independently, which is particularly convenient for calculating the likelihood of an alignment; we only need to multiply all the site likelihoods calculated from each site of the alignment (**?**). On the other hand, taking into account substitution processes involving dependencies between sites (e.g., **????**), as required to parameterize the hypermutability of CpGs, is not a trivial task. Dealing with such substitution processes makes the development of mathematics and code more complex and limited to a small circle of researchers. This has led researchers to develop hybrid methods that integrate standard posterior sampling methods (e.g., Markov Chain Monte Carlo) with so-called simulation-based methods (e.g., **?**) or to develop strategies that rely on data augmentation (e.g., **??**).

Posterior predictive sampling (**??**) is a central tool in the Bayesian toolbox. This tool is of primary use when validating new phylogenetic models (e.g., **??**). For example, knowing that CpGs are depleted by the high rate of spontaneous deamination due to methylation of most vertebrate cytosines within CpG contexts (**??**), we are interested in the ability of substitution models to predict dinucleotide (e.g. CpG) frequencies of mammalian protein-coding genes.

Following the textbook case of mammalian CpG hypermutability, we will test our ability to detect the presence of CpG hypermutability using one of the most widely used substitution models, the GTR+G substitution model, using Bayesian posterior predictive sampling. By definition, we expect the GTR+G 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. Similarly, the GTR+G substitution model will allow stop codons to be generated within predictive alignments. We first validated the posterior predictive test using simulations generated without and with CpG hy-

permutability, using the GTR+G substitution model and a codon substitution model implemented in the mutation selection framework. The codon substitution model used uses a GTR parameterization for the background mutation process plus a λ parameter for CpG hypermutability and a parameter to capture global negative selection on amino acids. We then apply the test to 137 mammalian protein-coding genes for which we have already measured the degree of CpG hypermutability using a complex codon substitution model (**?**). Finally, we discuss the ranking of other dinucleotides in terms of their ability to be predicted by the GTR+G substitution model using the same approach developed for testing for the presence of CpG hypermutability.

Materials and methods

Detecting CpG hypermutability using posterior predictive sampling

The CpG test consists of calculating the proportion of times, i.e., *p-value*, that the frequency of CpGs calculated from the real alignments is lower than the CpG frequency recovered from the predictive alignments, where the predictive alignments are generated from a set of parameter values, i, sampled from the posterior. The following equation details the CpG test:

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

where N is the predictive sample size and 1 is an indicator function that returns 1 if the CpG frequency calculated from the true alignment is less than the predictive CpG frequency and 0 otherwise.

Real data and tree topology

We retrieved the 137 mammalian codon alignments from **?** along with the tree topology.

Simulation study

We first analyzed 10 mammalian protein-coding genes selected for their wide range of GC content as used to validate the new approach developed in **?** using the GTR+G substitution model as well as the M0GTR codon substitution model, but with ω fixed at 1 (**?**) using Phylobayes-MPI (**?**) ** jai fait une petite bourde ici, en laissant omega fixed, mais c'est juste pour avoir des valeurs de paramètres, donc ca ne dérange pas **. We generated 100 predictive alignments (10 genes $\times$ 10 samples) using AliSim (**?**) for the GTR+G substitution model, which will later be analyzed with the same model to generate a new set of predictive alignments on which to apply the CpG test. We also generated 600 predictive alignments (10 genes $\times$ 3 values of λ $\times$ 2 values of ω $\times$ 10 posterior predictive samples) under the M0GTR substitution model using the jump chain simulation algorithm described in **?** to account for CpG hypermutability. All simulations were performed with a single ω value (i.e., 0.2 or 1). To evaluate the false positive and false negative rates of the test, we generated predictive alignments using $\lambda = 1$ and $\lambda = \{4.8\}$, respectively.

Each simulated predictive alignment was then analyzed with the GTR+G substitution model using Phylobayes-MPI (**?**) software, we generate posterior predictive alignments, 50 for each analysis, using the same software. The predictive alignments were generated by sampling 50 sets of parameter values from the posterior of each of the GTR+G analyses performed on synthetic alignments. Each of the experimental conditions (3 values of λ $\times$ 2 values of ω) is replicated 100 times. The proportion of false positives and false negatives in each set of ex-

perimental conditions was then calculated by applying the posterior predictive tests.

Empirical study

We analyzed the 137 mammalian codon sequence alignments with the GTR+G substitution model implemented in Phylobayes-MPI: (**?**). We sample 100 posterior predictive alignments, also using Phylobayes-MPI, for each mammalian gene of interest. We then apply the posterior predictive test by calculating the frequency of CpGs from the true and predictive alignments to detect the presence of CpG hypermutability. Similarly, for each dinucleotide context, we could apply the specific predictive test as designed for the CpG context and calculate the proportion of positive tests to identify dinucleotide contexts missed by the GTR+G substitution model.

Results

Simulation study

Less than 5% of the CpG tests were significant when performed on synthetic sequence alignments generated under the GTR+G 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 M0GTR codon substitution model (Table 1). On the other hand, under less realistic conditions, without negative selection on amino acids, $\omega = 1$, we recovered more false positives, i.e., 8% (Table 1), but still close to the alpha significance threshold of 5%. No false negatives were generated in the presence of CpG hypermutability, (i.e., Table 1: $\lambda = \{4, 8\}$), all CpG tests were positive, for both of the omega values used, i.e., $\omega = \{0.2, 1\}$.

Empirical study

Therefore, after validating the CpG test, we analyzed 137 mammalian protein-coding genes with the GTR+G substitution model and found 100% of the tested genes to be significant (p-$value$=0, α=5%), rejecting the null hypothesis that the GTR+G substitution model accounts for CpG hypermutability. In other words, CpG frequencies are significantly reduced in all real alignments compared to frequencies recovered from predicted alignments generated by the GTR+G substitution model. As an example, in Figure 1, panel A, we show the discrepancy between the CpG frequencies predicted by the GTR+G substitution model and the observed CpG frequency in one of the examined coding sequence alignments. In panel B of the same figure, we see that the predicted values of ApT are distributed on either side of the actual value calculated from the same gene examined in panel A.

Next, we applied the predictive test to all other dinucleotide contexts using the same substitution model, i.e., GTR+G. After CpG, TpA, and GpT contexts, it was found that the GTR+G substitution model significantly missed most of the tested genes (94.9% and 81%, respectively; Table S1), considering an alpha significance threshold of 5%.

Discussion

In this work, we designed a simple test, based on posterior predictive sampling, a tool from the Bayesian toolbox, to detect the presence of CpG hypermutability. We validated the test using a simulation study: with approximately the alpha significance threshold of false positives and no false negatives. We then applied the test to a set of mammalian protein-coding genes, for which we had already obtained measurements of CpG hypermutability using a complex codon sub-

stitution model implemented in the mechanistic mutation-selection framework (**?**).

We also showed that contexts other than CpG were systematically missed by the predictive samples generated by the GTR+G substitution model, namely TpA with 94.9% of genes detected as positive (Table S1) and GpT with 81% of positive tests (Table S1). The TpA context may be artifactually overpredicted by the GTR+G substitution model because it does not take into account the structure of the genetic code and allows the presence of stop codons in predictive protein-coding sequence alignments. Two out of three stop codons contain a TpA dinucleotide (i.e., TAA and TAG). The use of a codon substitution model implemented in the mechanistic mutation-selection framework, such as the one used here to validate the CpG test, i.e., M0GTR, should at least help to reduce the over-prediction of TpA contexts by disallowing the presence of stop codons within the predictive alignments generated when running the test, as the substitution model imposes infinite negative selection against mutations that land on stop codons. On the other hand, the etiology of TpA context depletion is highly controversial and may not be exclusively due to mutation or selection processes (e.g., **??**). Interestingly, GpT contexts have been shown to be hypermutable on shorter evolutionary timescales, i.e. within human populations (**?**), suggesting that the evolutionary process behind the depletion of GpT context is conserved over the mammalian tree. On the other hand, considering contexts other than CpG should also require extending the scope of the simulation study, since predicting the effect of one mutation process on another mutation process is particularly complicated. For example, CpG and TpA contexts share the same transition products (TpG and CpA).

However, knowing that CpG hypermutability could affect the measurement of negative and positive selection on synonymous and nonsynonymous mutations

(**??**), the test could be systematically applied when conducting large-scale evolutionary studies (e.g., **???**) or in future projects such as the Zoonomia Project (**?**) to promte the development of new mechanistic tests using codon substitution models that take into account a wider range of population processes (e.g., **??????**). We could also investigate the presence of hypermutability patterns from substitution mappings, similar to what has been developed by (**??**).

Limitations

The main limitation of this study is that we have not yet investigated how a non-mechanistic test, such as the one developed here based on comparison of posterior predictive alignments, might be affected by the presence of other population processes (i.e. mutation, selection, and demographic processes) not accounted for by the GTR+G substitution model, which could be addressed by simulation studies. Because the CpG test we developed here is non-mechanistic, it also does not allow quantification of the degree of CpG hypermutability, as a proper mechanistic test could do (e.g., **?**).

Availability of data and materials

Workflow and data is available.

Abbreviations

Figures

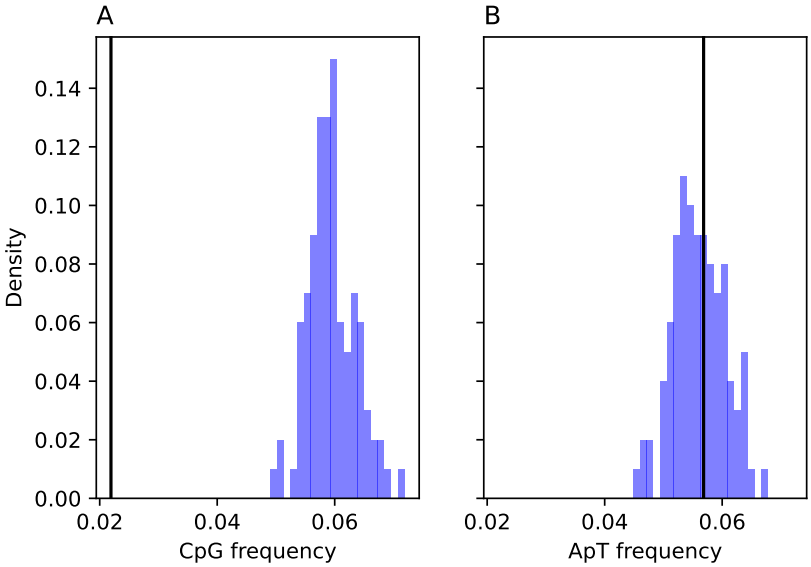

Figure 1: Comparison of CpG (A) and ApT (B) frequencies computed from the *MEP1A* mammalian protein-coding gene alignment, black vertical lines, and corresponding frequencies computed from posterior predictive alignments generated using the GTR+G substitution model, blue histograms. Note that the predicted 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+G	0
negative	M0GTR+($\lambda = 1$)+($\omega = 0.2$)	0
negative	M0GTR+($\lambda = 1$)+($\omega = 1.0$)	8
positive	M0GTR+($\lambda = 4$)+($\omega = 0.2$)	100
positive	M0GTR+($\lambda = 4$)+($\omega = 1.0$)	100
positive	M0GTR+($\lambda = 8$)+($\omega = 0.2$)	100
positive	M0GTR+($\lambda = 8$)+($\omega = 1.0$)	100

Table 1: Validation of CpG test using posterior predictive sampling under GTR+G substitution model with an α threshold of significance 5%. Synthetic alignments were generated using GTR+G and M0GTR codon substitution model using three CpG hypermutabilities values, $\lambda = \{1, 4, 8\}$ and a global selection on amino acids using two ω values (i.e., 0.2 and 1).

References

Acknowledgments

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

Funding

Author information

Authors and Affiliations

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.

Corresponding author

Correspondance to Simon Laurin-Lemay

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information

Table S1: Proportion of positive tests calculated for each dinucleotide context obtained using GTR+G substitution model over the 137 mammalian protein-coding genes studied.