CpG + CpNpG Analysis of Protein-Coding Sequences from Tomato

Asger Hobolth, Rasmus Nielsen,† Ying Wang,‡ Feinan Wu,‡ and Steven D. Tanksley‡*

*Bioinformatics Research Center, North Carolina State University; †Bioinformatics Centre, University of Copenhagen, Copenhagen, Denmark; and ‡Department of Plant Biology, Cornell University

We develop codon-based models for simultaneously inferring the mutational effects of CpG and CpNpG methylation in coding regions. In a data set of 369 tomato genes, we show that there is very little effect of CpNpG methylation but a strong effect of CpG methylation affecting almost all genes. We further show that the CpNpG and CpG effects are largely uncorrelated. Our results suggest different roles of CpG and CpNpG methylation, with CpNpG methylation possibly playing a specialized role in defense against transposons and RNA viruses.

Introduction

In plants, methylation of cytosine in CpG and CpNpG sites occur with much higher frequency than in other sequence patterns (Bender 2003). Methylation leads to an increased rate of C to T and G to A mutations in these sites (Ng and Bird 1999) and a corresponding genomic deficiency of CpG and CpNpG sites. Inference of patterns of methylation is of considerable interest because it relates to patterns of regulation, epigenetic phenomena, and chromosome structure. The pattern of methylation in coding regions is of particular interest because of the effect methylation has on expression and because unmethylated regions are thought to be enriched with genes. Some genomic sequencing strategies are, therefore, specifically targeting unmethylated regions in order to preferentially clone and sequence gene-rich regions.

In this paper, we develop statistical models for analysis of patterns of CpG and CpNpG deficiency in coding regions. Two factors make the development of such models more difficult than most standard models. Firstly, the pattern of codons usage and selection acting on nonsynonymous mutation must be taken into account while accounting for CpG and CpNpG hypermutation. Secondly, the assumption made in classical sequence evolution of independence among sites is violated by the CpG and CpNpG effects. We develop a new statistical method for analyzing CpG and CpNpG hypermutation from a single sequence based on context-dependent codon models. Formulating the model on the codon level allows us to take codon bias into account.

The proposed codon model is an extension of the model introduced by Jensen and Pedersen (2000). The Jensen and Pedersen (2000) model significantly improved the description of HIV sequence evolution by extending the Goldman and Yang (1994) codon model to take CpG hypermutation into account. Siepel and Haussler (2004) and Huttley (2004) also incorporated methylation into codon models and found that methylation has a significant effect on the evolution of protein-coding genes from mammals. Huttley (2004) not only considered CpG hypermutation but also included the dinucleotides CpA and CpT in the analysis. Sved and Bird (1990) and Lunter and Hein (2004) developed CpG mutation models on the nucleotide level and analyzed pseudogene evolution in human and noncoding evolution between human and mouse, respectively.

We develop a context-dependent codon model to investigate the relative effect of CpG and CpNpG mutation and the degree to which mutation rates in these two types of sequence motifs are correlated. Using this new method, we investigate the CpG and CpNpG methylation effects of 369 tomato genes. We show that there is a very strong deficiency of CpG sites. Thus, there are significantly fewer CpG sites in the genes than predicted from the codon usage. While there is a strong CpG deficiency, there seems to be little or no evidence for CpNpG hypermutation in the investigated genes. Additionally, we show that the effect of CpG and the effect of CpNpG hypermutation seem to be largely uncorrelated.

Materials and Methods

Context-Dependent Nucleotide Model

In order to understand the context-dependent "codon" model, it is useful first to consider a context-dependent "nucleotide" model. The evolution of a DNA sequence is often described as a stationary homogeneous reversible continuous time Markov process. In the site-independent model, the general time-reversible model has rate matrix (e.g., Yap and Speed 2004)

$$Q = \begin{bmatrix} \cdot & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & \cdot & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & \cdot & \zeta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \zeta\pi_C & \cdot \end{bmatrix},$$

where the off-diagonal entries, the instantaneous rates of substitutions, are all nonnegative, and the diagonal elements are such that each row sums to zero. We can write the rate matrix as $Q = S\,\mathrm{diag}(\pi)$, where

$$S = \begin{bmatrix} \cdot & \alpha & \beta & \gamma \\ \alpha & \cdot & \delta & \epsilon \\ \beta & \delta & \cdot & \zeta \\ \gamma & \epsilon & \zeta & \cdot \end{bmatrix}$$

is a symmetric matrix and $\mathrm{diag}(\pi)$ is the diagonal matrix with $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$ on the diagonal. We observe that the detailed balance condition $\mathrm{diag}(\pi)Q = Q^*\mathrm{diag}(\pi)$ is fulfilled, where * denotes vector transpose, and thus π is the stationary distribution.

Now suppose the flanking nucleotides of the evolving site are fixed during evolution and consider the case where

Key words: codon model, context dependency, CpG + CpNpG methylation.

E-mail: asger@daimi.au.dk.

Mol. Biol. Evol. 23(6):1318–1323. 2006
doi:10.1093/molbev/msk017
Advance Access publication April 6, 2006

the left nucleotide is an A and the right nucleotide is a C. In order to model the higher substitution rate away from CpGs, we divide each off-diagonal term in the row corresponding to substitution rates from C with a parameter $\lambda < 1$ so that the rate matrix becomes

$$\begin{bmatrix} \cdot & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & \cdot & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A/\lambda & \delta\pi_G/\lambda & \cdot & \zeta\pi_T/\lambda \\ \gamma\pi_A & \epsilon\pi_G & \zeta\pi_C & \cdot \end{bmatrix}.$$

This rate matrix can be written in terms of a symmetrix matrix and a diagonal matrix as

$$\begin{bmatrix} \cdot & \alpha & \beta/\lambda & \gamma \\ \alpha & \cdot & \delta/\lambda & \epsilon \\ \beta/\lambda & \delta/\lambda & \cdot & \zeta/\lambda \\ \gamma & \epsilon & \zeta/\lambda & \cdot \end{bmatrix} \mathrm{diag}(\pi_A, \pi_G, \lambda\pi_C, \pi_T),$$

and from detailed balance we conclude that the stationary distribution is proportional to $(\pi_A, \pi_G, \lambda\pi_C, \pi_T)$, which means that we observe fewer C's than in the site-independent model.

In the flanking situation where the left nucleotide is a C and the right nucleotide is a G, we also need to divide each off-diagonal term in the row corresponding to G by λ. We therefore obtain the following rate matrix and corresponding symmetric and diagonal matrices

$$\begin{bmatrix} \cdot & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A/\lambda & \cdot & \delta\pi_C/\lambda & \epsilon\pi_T/\lambda \\ \beta\pi_A/\lambda & \delta\pi_G/\lambda & \cdot & \zeta\pi_T/\lambda \\ \gamma\pi_A & \epsilon\pi_G & \zeta\pi_C & \cdot \end{bmatrix}$$
$$= \begin{bmatrix} \cdot & \alpha/\lambda & \beta/\lambda & \gamma \\ \alpha/\lambda & \cdot & \delta/\lambda & \epsilon/\lambda \\ \beta/\lambda & \delta/\lambda & \cdot & \zeta/\lambda \\ \gamma & \epsilon/\lambda & \zeta/\lambda & \cdot \end{bmatrix} \mathrm{diag}(\pi_A, \lambda\pi_G, \lambda\pi_C, \pi_T),$$

so that the stationary distribution is now proportional to $(\pi_A, \lambda\pi_G, \lambda\pi_C, \pi_T)$. In this case, we observe fewer C's and G's than in the site-independent model.

We now move from the fixed flanking situation to the general case. A change in the nucleotide sequence $x = (x_1, \ldots, x_n)$ consists of a change of one nucleotide only, and the rate matrix is no longer a 4×4 matrix but is a $4^n \times 4^n$ matrix. Consider the rate from sequence x to sequence $\tilde{x}$, where x and $\tilde{x}$ are the same except at position k. The new nucleotide is denoted as $\tilde{x}_k$. Following the ideas from above, the rate from x to $\tilde{x}$ is determined by two main components.

Firstly, there is the 4×4 substitution rate matrix Q, where the rates do not depend on the neighboring codons. This component corresponds to the model one would use had there been no interaction among nucleotides. We assume that the site-independent part of the model is reversible with stationary distribution π, such that detailed balance $\mathrm{diag}(\pi)Q = Q^*\mathrm{diag}(\pi)$ is fulfilled.

Secondly, there is the CpG component, determined by the parameter λ, that introduces dependence among nucleotides. If $\lambda < 1$, the component introduces higher substitution rates from CpG pairs. If $\lambda > 1$, the component introduces lower substitution rates, and if $\lambda = 1$, there is no CpG effect. Consider the nucleotide triplet (y_1, y_2, y_3) and suppose y_2 undergoes a change. If (y_1, y_2) or (y_2, y_3) are CG pairs and $\lambda < 1$ ($\lambda > 1$), the substitution rate for a change should increase (decrease); and if $\lambda = 1$, the substitution rate should remain unchanged. We therefore define the function

$$R(y_1, y_2, y_3) = (1/\lambda)^{1_{CG}(y_1,y_2) + 1_{CG}(y_2,y_3)},$$

which is $1/\lambda$, if y_2 is a member of a CG pair and 1 otherwise. The function $1_{CG}(y_1, y_2)$ is the indicator function for the event $(y_1, y_2) = (C, G)$.

The rate γ for a change from sequence x to sequence $\tilde{x}$ thereby depends on x_k, $\tilde{x}_k$, and the neighboring pairs x_{k-1} and x_{k+1} and is given by

$$\gamma(\tilde{x}_k; x_{k-1}, x_k, x_{k+1}) = Q(x_k, \tilde{x}_k)R(x_{k-1}, x_k, x_{k+1}).$$

The nice feature about this model is that the stationary distribution can be found. As can be proved from detailed balance on the sequence level, the stationary distribution is given by

$$P(x) = \frac{1}{Z(\lambda, \pi)}\lambda^{\sum_{k=0}^{n} 1_{CG}(x_k, x_{k+1})} \prod_{k=1}^{n} \pi_{x_k},$$

where $Z(\lambda, \pi)$ is a normalizing constant and x_0 and x_{n+1} are fixed. This expression is expected from the stationary distributions derived above for the fixed flanking situations.

We can use this expression for the stationary distribution to analyze the CpG effect. Indeed, we can estimate the parameters λ and $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$ from a single sequence using, for example, maximum likelihood, and if λ is significantly smaller than 1, we may conclude that CpG methylation has played a role during the evolution of the sequence.

Context-Dependent Codon Model

In this section, we formulate the model from the previous section on the codon level and include the CpNpG effect.

We write a codon sequence x of n codons as $x = (x_1, \ldots, x_n)$ with $x_k = (x_k^1, x_k^2, x_k^3)$, where the upper index u in x_k^u indicates the position within the kth codon, and we assume that the boundary codons of a sequence are a start codon (x_0 = ATG) and a stop codon ($x_{n+1} \in$ TAA, TGA, TGG). Now consider the rate from sequence x to a new sequence $\tilde{x}$, which equals x except at one position in codon x_k. The new codon is denoted $\tilde{x}_k$, and the rate from x to $\tilde{x}$ is determined by two main components.

Firstly, there is the codon substitution rate matrix Q, where the rates do not depend on the neighboring codons. This component corresponds to the site-independent part of the model. We assume that the model is reversible so that detailed balance $\pi_a Q(a, b) = \pi_b Q(b, a)$ is fulfilled. Here π_a is the frequency of codon a.

Secondly, there is the CpG + CpNpG component determined by the parameters λ_{CG} and λ_{CNG}. This component

introduces dependence among codons. If $\lambda_{CG} < 1$, the parameter introduces higher substitution rates from CG pairs; if $\lambda_{CG} > 1$, the parameter introduces lower substitution rates; and if $\lambda_{CG} = 1$, there is no methylation CG effect. As for the nucleotide model, we define the function

$$R_{CG}(y_1, y_2, y_3) = (1/\lambda_{CG})^{1_{CG}(y_1,y_2) + 1_{CG}(y_2,y_3)},$$

which is $1/\lambda_{CG}$, if y_2 is a member of a CG pair and 1 otherwise.

Similarly, consider the nucleotide pentet (y_1, …, y_5) and suppose y_3 undergoes a change. If (y_1, y_2, y_3) or (y_3, y_4, y_5) are CNG triplets, the rate should increase ($\lambda_{CNG} < 1$), decrease ($\lambda_{CNG} > 1$), or remain unchanged ($\lambda_{CNG} = 1$). We therefore define the function

$$R_{CNG}(y_1, y_2, y_3, y_4, y_5) = (1/\lambda_{CNG})^{1_{CNG}(y_1,y_2,y_3) + 1_{CNG}(y_3,y_4,y_5)}.$$

The rate γ for a change from sequence x to sequence $\tilde{x}$ thereby depends on x_k, $\tilde{x}_k$, and the neighboring pairs x^2_{k-1}, x^3_{k-1} and x^1_{k+1}, x^2_{k+1} and is given by

$$\begin{aligned}
&\gamma(\tilde{x}_k; x^2_{k-1}, x^3_{k-1}, x_k, x^1_{k+1}, x^2_{k+1}) \\
&\quad = Q(x_k, \tilde{x}_k) \\
&\qquad \times \begin{cases} R_{CG}(x^3_{k-1}, x^1_k, x^2_k) R_{CNG}(x^2_{k-1}, x^3_{k-1}, x_k), & \text{if } x^1_k \neq \tilde{x}^1_k \\ R_{CG}(x_k) R_{CNG}(x^3_{k-1}, x_k, x^1_{k+1}), & \text{if } x^2_k \neq \tilde{x}^2_k \\ R_{CG}(x^2_k, x^3_k, x^1_{k+1}) R_{CNG}(x_k, x^1_{k+1}, x^2_{k+1}), & \text{if } x^3_k \neq \tilde{x}^3_k. \end{cases}
\end{aligned}$$

A crucial feature of our model is again that the stationary distribution can be determined from detailed balance on the sequence level. The stationary distribution is given by

$$\begin{aligned}
P(x) &= \frac{1}{Z(\lambda_{CG}, \lambda_{CNG}, \pi)} \\
&\quad \times \lambda_{CG}^{\sum_{k=1}^{n} (1_{CG}(x^1_k, x^2_k) + 1_{CG}(x^2_k, x^3_k) + 1_{CG}(x^3_k, x^1_{k+1}))} \\
&\quad \times \lambda_{CNG}^{\sum_{k=1}^{n} (1_{CNG}(x_k) + 1_{CNG}(x^2_k, x^3_k, x^1_{k+1}) + 1_{CNG}(x^3_k, x^1_{k+1}, x^2_{k+1}))} \\
&\quad \times \prod_{k=1}^{n} \pi_{x_k},
\end{aligned} \tag{1}$$

where $Z(\lambda_{CG}, \lambda_{CNG}, \pi)$ is a normalizing constant. When $\lambda_{CG} = \lambda_{CNG} = 1$, we get $Z = 1$, and the stationary distribution for a sequence is the product of the codon frequencies along the sequence. Thus, the last term in equation (1) takes codon bias into account. The two terms in equation (1) that involves λ_{CG} and λ_{CXG} introduces a Markovian dependence structure along the sequence. If, for example, $\lambda_{CG} < 1$, it is less likely to observe the dinucleotide CG across a codon boundary than if $\lambda_{CG} = 1$.

From the stationary distribution (1), we obtain the log likelihood

$$\begin{aligned}
l(\lambda_{CG}, \lambda_{CNG}, \pi) = &- \log Z(\lambda_{CG}, \lambda_{CNG}, \pi) + N_{CG} \log \lambda_{CG} \\
&+ N_{CNG} \log \lambda_{CNG} + \sum_a n_a \log \pi_a,
\end{aligned}$$

where N_{CG} is the number of CG pairs in the sequence, N_{CNG} is the number of CNG triplets in the sequence, and n_a is the number of times codon a appears in the sequence. In the *Appendix*, we describe how the normalizing constant can be calculated. These results can be used directly for data analysis. The likelihood function for a particular data set can be calculated using the stationary distribution, and likelihood ratio tests can be conducted in the usual way. For example, a significant CpG effect can be tested by comparing minus two times the log-likelihood ratio between a model assuming $\lambda_{CNG} = 1$ and a model allowing λ_{CNG} to be a free parameter, to a $\chi^2(1)$ distribution.

Data

We considered 369 protein-coding genes from tomato. The genes were selected from a tomato expressed sequence tag–assembled unigene set in the Solanaceae Genome Network database (http://www.sgn.cornell.edu/content/sgn_data.pl#Solanumlycopersicum). Gene sequences were subjected to untranslated region trimming and manual correction of sequencing errors based on multiple sequence alignments of the corresponding *Arabidopsis* ortholog. The URL for the data is ftp://ftp.sgn.cornell.edu/COSII/Rasmus_s_cleantomatoseq.fasta.

Results

Tomato Gene Analysis

The tomato genes are analyzed using the stationary distribution of the context-dependent codon model described above. We assume common codon frequencies but gene-specific CpG + CpNpG effects. In order to find the maximum likelihood estimates, we used an iterative procedure where we updated the CpG and CpNpG parameters for each gene using the current estimate of the common codon frequencies and updated the common codon frequencies using the current estimates of the CpG and CpNpG parameters.

The left histogram in figure 1 shows the strong effect of CpG mutation in these genes. The mean value of $\hat{\lambda}_{CG}$ is 0.375 and the median is 0.326. The fact that $\hat{\lambda}_{CG} < 1$ in almost all cases suggest that the CpG effect is acting on all the genes. As can be seen from the right histogram in figure 1, in only 30 of the 369 genes was the test for $\lambda_{CG} = 1$ against the alternative that $\lambda_{CNG} \neq 1$ not rejected at the 5% significance level. We used the likelihood ratio test statistics $2[l(\hat{\lambda}_{CG}, \hat{\lambda}_{CNG}, \hat{\pi}) - l(1, \hat{\lambda}_{CNG}, \hat{\pi})]$ and a $\chi^2(1)$ distribution to calculate the P values.

The left histogram in figure 2 shows the maximum likelihood estimates of λ_{CNG}. It is evident that the mean value of λ_{CNG} is close to 1 (the mean is 1.02 and the median is 0.99), and the CpNpG effect is therefore less pronounced. Indeed, for each gene, we tested $\lambda_{CNG} = 1$ against the alternative that $\lambda_{CNG} \neq 1$, and the P values are almost uniformly distributed (right plot in fig. 2). We conclude that the CpNpG effect does not affect the stationary distribution of the sequences.

To confirm that these results are not an artifact of the statistical models, we conducted a simple analysis of the number of CpG and CpNpG sites spanning codon boundaries. We estimated codon frequencies from the observed codon counts of 369 protein-coding sequences from tomato. From the observed codon frequencies, we calculated

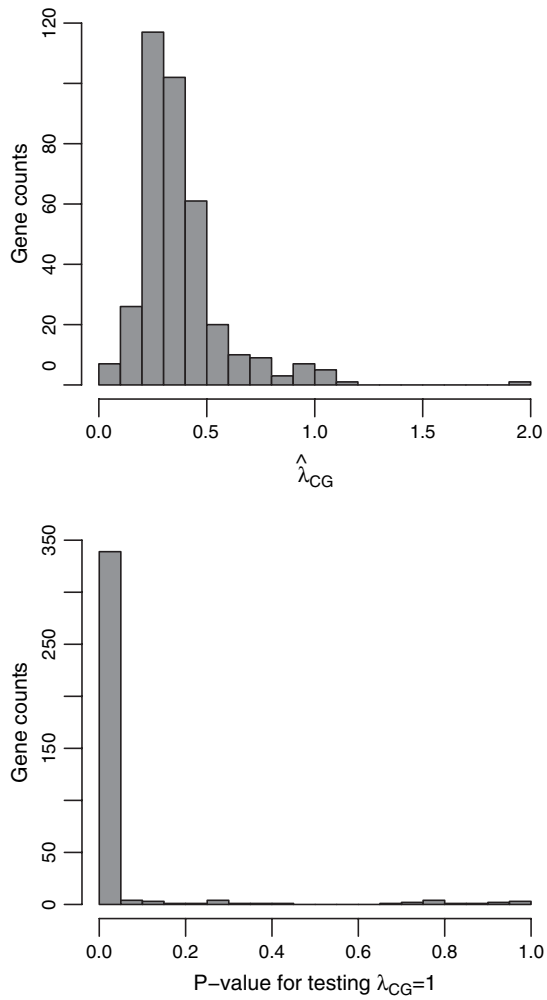

FIG. 1.—Left: histogram of $\hat{\lambda}_{CG}$. Right: histogram of *P* values for testing $\lambda_{CG} = 1$. The vast majority of the tomato genes have λ_{CG} significantly lower than 1, suggesting that the CpG effect is acting on almost all genes.

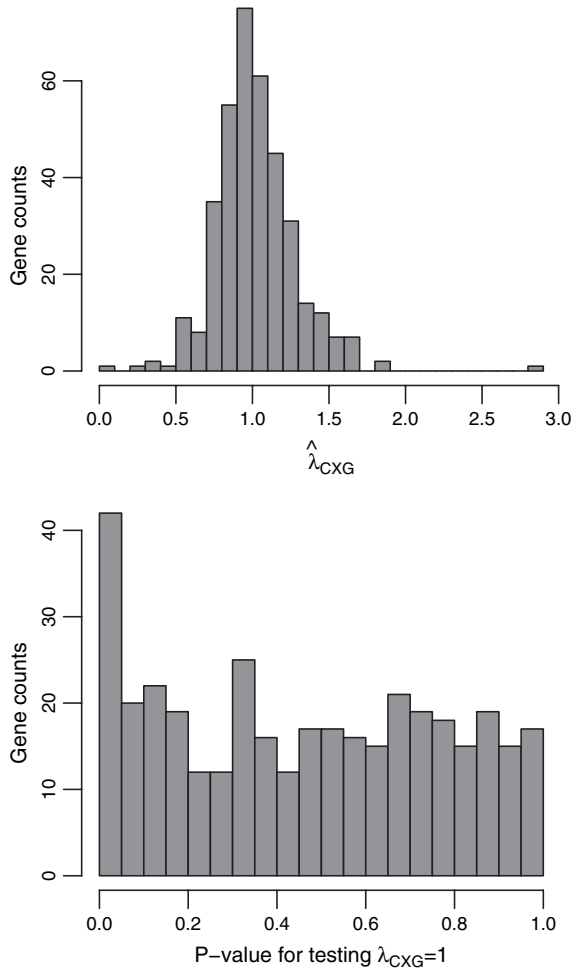

FIG. 2.—Left: histogram of $\hat{\lambda}_{CNG}$. Right: histogram of *P* values for testing $\lambda_{CNG} = 1$. Only a few tomato genes have λ_{CNG} significantly lower than 1, suggesting that the CpNpG effect is less pronounced.

the frequency of nucleotide C in third codon position (0.1714) and nucleotide G in first codon position (0.3332). The 369 coding sequences have 100,060 codon boundaries, and we thus expect $0.1714 \times 0.3332 \times 100060 = 5{,}715$ CpGs over codon boundaries (table 1). However, we only observe 2,674 CpGs over codon boundaries. A similar analysis can be performed for CpNpG patterns (see table 1). In this case, the boundary is either immediately after the C or just before the G. As is seen from table 1, we also observe fewer CpNpGs over codon boundaries, but the effect is not very pronounced.

These results confirm that the predominant mutational effect of methylation in the coding regions analyzed here is on CpG dinucleotides and not on CpNpG motifs.

Finally, we consider the joint distribution of $\hat{\lambda}_{CG}$ and $\hat{\lambda}_{CNG}$ as shown in figure 3. This plot shows no correlation between the CpG and CpNpG effects (the correlation coefficient between the CpG and CpNpG estimates is only −0.06).

In order to verify this conclusion in more detail, we considered the two issues of biasedness and standard errors (SEs) of the estimates. Maximum likelihood estimates are asymptotically unbiased, and the sample size is relatively large in our application. The vast majority of the genes have codon lengths in the range 100–400. Thus, we expect the maximum likelihood estimates to be approximately unbiased. We used profile maximum likelihood confidence intervals (CIs) to determine the SEs for λ_{CG} and λ_{CNG} for each gene. For λ_{CG}, the 95% CIs are not very variable between genes and in the range $(\hat{\lambda}_{CG} - 0.12, \hat{\lambda}_{CG} + 0.14)$. The λ_{CNG} CIs are also rather stable but somewhat larger. The intervals are typically in the range $(\hat{\lambda}_{CNG} - 0.25, \hat{\lambda}_{CNG} + 0.25)$. Thus, neither biasedness nor reliability of the estimates should be an issue when interpreting figure 3.

Table 1
Observed and Expected Counts of CpG and CpNpG Patterns Over Codon Boundaries

Pattern	xxC|Gxx	xCx|Gxx	xxC|xGx
Observed	2,674	7,702	2,858
Expected	5,715	7,780	2,875

NOTE.—Presumably due to methylation, there are fewer observed patterns than expected under the codon site independence assumption. The CpG methylation effect is very pronounced, while the CpNpG methylation effect is very weak.

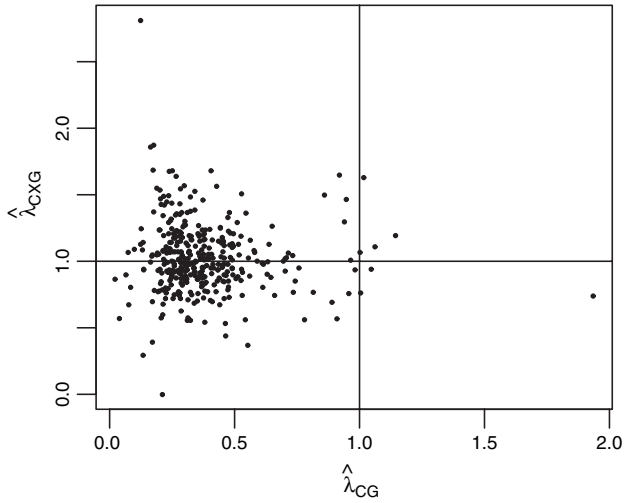

FIG. 3.—Scatter plot of $\hat{\lambda}_{CG}$ versus $\hat{\lambda}_{CNG}$. There is no correlation between the λ_{CG} and λ_{CNG} estimates.

Discussion

While comparative sequence analysis traditionally have been used to investigate patterns of substitution, there is also considerable information about these patterns in a single sequence. In traditional models of sequence evolution, which assume independence among sites, only three independent mutation parameters can be estimated from a single sequence. However, by increasing the state space of the model, context-dependent effects can be estimated. In particular, using context-dependent models, we were here able to estimate the rate of CpG and CpNpG hypermutation.

We found that mutation rates in CpG and CpNpG sites were uncorrelated in the tomato genome. This supports the idea that separate methyltransferases are responsible for CpG and CpNpG methylation activity (e.g., Pradhan and Adams 1995). Furthermore, we found an apparent absence of CpNpG hypermutation. This observation supports the notion that CpNpG methylation, in contrast to CpG methylation, almost never occurs in genic regions. CpNpG methylation may possibly play a role in silencing of genomic elements such as transposable and retrotransposable elements (e.g., Papa et al. 2001). CpNpG methylation is controlled by the CMT3 methyltransferase gene (e.g., Lindroth et al. 2001, Cao and Jacobsen 2002). CMT3 mutants display a wild-type morphology but exhibit both decreased CpNpG methylation and reactivated expression of endogenous retrotransposon sequences (Lindroth et al. 2001). In contrast, the pervasiveness of CpG avoidance in the tomato genes investigated here suggests that CpG methylation in plant genes does not have as radical an effect on expression as CpNpG methylation. The vast majority of all genes show a strong tendency toward CpG avoidance, suggesting that CpG methylation is pervasive in expressed genes.

Lindroth et al. (2001) observed an almost complete genomic loss of CpNpG methylation in CMT3 null mutants. However, the mutants were morphologically normal, even after five generations of inbreeding. They hypothesized that CpNpG and CpG methylation may act in a partially redundant fashion to silence most genes. Our results suggest instead different roles of CpNpG and CpG methylation. While CpG methylation may play a role in both normal regulation of gene expression and defense against viruses, the role of CpNpG methylation may be more specialized in the defense against transposons and RNA viruses. No phenotypic effects are, therefore, expected from reduced CpNpG methylation in the absence of viral or transposon activity.

Acknowledgments

We thank the associate editor and three anonymous reviewers for their constructive suggestions. A.H. and R.N. are supported by the Danish Research Council.

Appendix

Normalizing Constant for Stationary Distribution

In order to calculate the normalizing constant $Z(\lambda_{CG}, \lambda_{CNG}, \pi)$, we first write the stationary distribution (1) as

$$\begin{aligned} P(x) = \frac{1}{Z}\pi_{x_1}\lambda_{CG}^{1_{CG}(x_1^1,x_1^2)+1_{CG}(x_1^2,x_1^3)}\lambda_{CNG}^{1_{CNG}(x_1)} \\ \times\prod_{k=2}^{n}\lambda_{CG}^{1_{CG}(x_{k-1}^3,x_k^1)+1_{CG}(x_k^1,x_k^2)+1_{CG}(x_k^2,x_k^3)} \\ \times\lambda_{CNG}^{1_{CNG}(x_{k-1}^2,x_{k-1}^3,x_k^1)+1_{CNG}(x_{k-1}^3,x_k^1,x_k^2)+1_{CNG}(x_k)}\pi_{x_k} \\ \times\lambda_{CG}^{1_{CG}(x_n^3,x_{n+1}^1)}\lambda_{CNG}^{1_{CNG}(x_n^2,x_n^3,x_{n+1}^1)+1_{CNG}(x_n^3,x_{n+1}^2,x_{n+1}^3)} \\ = \frac{1}{Z}S(x_1)U(x_n)\prod_{k=2}^{n}T(x_{k-1},x_k), \end{aligned}$$

with obvious definitions of S, U, and T. We therefore have

$$\begin{aligned} Z &= \sum_{x=(x_1,\ldots,x_n)} S(x_1)U(x_n)\prod_{k=2}^{n}T(x_{k-1},x_k) \\ &= \sum_{a,b} S(a)U(b)T^{n-1}(a,b). \end{aligned}$$

We can evaluate this directly, but for large n, we can obtain a good approximation as follows. Consider a spectral expansion of T with eigenvalues $\mu_j, j=1,\ldots,61$ (there are 61 sense codons) in decreasing order. Let the right and left eigenvectors of T be r_j and l_j so that $Tr_j = \mu_j r_j$ and $l_j^* T = l_j^* \mu_j$, where * denotes vector transpose, and suppose the eigenvectors are normalized so that $l_j^* r_j = 1$. Then we have $T = \sum_j \mu_j r_j l_j^*$ and $T^{n-1} = \sum_j \mu_j^{n-1} r_j l_j^*$. Recalling that μ_1 is the largest eigenvalue, we get $T^{n-1} \approx \mu_1^{n-1} r_1 l_1^T$ and thereby

$$Z \approx \mu_1^{n-1}\left(\sum_a S(a)r_1(a)\right)\left(\sum_b U(b)l_1(b)\right).$$

Literature Cited

Bender, J. 2003. DNA methylation and epigenetics. Annu. Rev. Plant Biol. **55**:41–68.

Cao, X., and S. E. Jacobsen. 2002. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. Proc. Natl. Acad. Sci. USA **99**:16491–16498.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Huttley, G. A. 2004. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. Mol. Biol. Evol. **21**:1760–1768.

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

Lindroth, A. M., X. Cao, J. P. Jackson, D. Zilberman, C. M. McCallum, S. Henikoff, and S. E. Jacobsen. 2001. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science **15**:2077–2080.

Lunter, G. A., and J. Hein. 2004. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics **20S**: i216–i223.

Ng, H., and A. Bird. 1999. DNA methylation and chromatin modification. Curr. Opin. Genet. Dev. **9**:158.

Papa, C. M., N. M Springer, M. G. Muszynski, R. Meeley, and S. M. Kaeppler. 2001. Maize chromomethylase Zea methyltransferase2 is required for CpNpG methylation. Plant Cell **13**:1919–1928.

Pradhan, S., and R. L. P. Adams. 1995. Distinct CG and CNG DNA methyltransferases in Pisum sativum. Plant J. **7**: 471–481.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**:468–488.

Sved, J., and A. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. USA **87**:4692–4696.

Yap, V. B., and T. P. Speed. 2004. Modeling DNA base substitution in large genomic regions from two organisms. J. Mol. Evol. **58**:12–18.

Naoka Takezaki, Associate Editor

Accepted April 4, 2006