Sun *et al. Theoretical Biology and Medical Modelling* (2020) 17:3
https://doi.org/10.1186/s12976-020-00122-x

Theoretical Biology and Medical Modelling

RESEARCH Open Access

Methylation-driven model for analysis of dinucleotide evolution in genomes

Jian-Hong Sun[1,2†], Shi-Meng Ai[3†] and Shu-Qun Liu[1*]

Abstract

Background: CpGs, the major methylation sites in vertebrate genomes, exhibit a high mutation rate from the methylated form of CpG to TpG/CpA and, therefore, influence the evolution of genome composition. However, the quantitative effects of CpG to TpG/CpA mutations on the evolution of genome composition in terms of the dinucleotide frequencies/proportions remain poorly understood.

Results: Based on the neutral theory of molecular evolution, we propose a methylation-driven model (MDM) that allows predicting the changes in frequencies/proportions of the 16 dinucleotides and in the GC content of a genome given the known number of CpG to TpG/CpA mutations. The application of MDM to the 10 published vertebrate genomes shows that, for most of the 16 dinucleotides and the GC content, a good consistency is achieved between the predicted and observed trends of changes in the frequencies and content relative to the assumed initial values, and that the model performs better on the mammalian genomes than it does on the lower-vertebrate genomes. The model's performance depends on the genome composition characteristics, the assumed initial state of the genome, and the estimated parameters, one or more of which are responsible for the different application effects on the mammalian and lower-vertebrate genomes and for the large deviations of the predicted frequencies of a few dinucleotides from their observed frequencies.

Conclusions: Despite certain limitations of the current model, the successful application to the higher-vertebrate (mammalian) genomes witnesses its potential for facilitating studies aimed at understanding the role of methylation in driving the evolution of genome dinucleotide composition.

Keywords: Dinucleotide, Methylation-induced mutation, Genome composition, Genome evolution

Background

The k-mer abundance analysis is widely used in genomics research [1–10]. The term k-mer refers to all possible substrings (in the 5′–3′ direction) of length k in a DNA sequence and, therefore, the k-mer frequency is a good variable for characterizing the composition of a genome's DNA sequence. Generally, analysis of the 2-mer (i.e., dinucleotide) frequency will provide more abundant information on genome composition than a simple statistic of the 1-mer (i.e., single nucleotides A, C, G, and T) frequency.

It has been shown that in the vertebrate genomes, the CpG dinucleotide is present at a lower frequency than expected [11–14]. The reason for this is thought to be due to a high C-to-T mutation rate at the methylated CpG sites [14–18]. 5-Methylcytosine (5mc), a cytosine modified by addition of a methyl group on the fifth position of the cytidine ring, can spontaneously deaminate to form thymine (i.e., 5mc to T mutation). Unlike the cytosine (C) to uracil (U) mutation arising from the spontaneous deamination of cytosine, the 5mc to T mutation is rarely recognized and removed by DNA repair enzymes [13]. Therefore, the high level of methylation at

* Correspondence: shuqunliu@ynu.edu.cn
†Jian-Hong Sun and Shi-Meng Ai contributed equally to this work.
[1]State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan & School of Life Sciences, Yunnan University, Kunming 650091, China
Full list of author information is available at the end of the article

the CpG sites explains why the mutation rate of C to T is 10–50 times higher than that of C to other nucleotides [11, 14].

Although several statistical analyses have revealed a negative correlation between the CpG and TpG/CpA levels [19–21], to the best of our knowledge, a rigorous method to predict the effects of CpG dinucleotide depletion on the changes in frequencies of all 16 dinucleotides (i.e., ApA, ApC, ApG, ApT, CpA, CpC, CpG, CpT, GpA, GpC, GpG, GpT, TpA, TpC, TpG, and TpT) and in the GC content (GC%) of vertebrate genomes is still lacking. Inspired by the substitution models [22, 23] commonly used in phylogenetic analyses and based on the neutral theory of molecular evolution, in this paper we propose a mathematical model, called the methylation-driven model (MDM), to investigate the effects of the methylation-induced CpG decay on the evolution of the genome dinucleotide composition and GC content.

Results

For the 10 vertebrate genomes, the statistical results of the frequencies/proportions (%) of the 16 dinucleotides and the GC content are listed in Table 1, the corresponding expected values obtained by application of MDM to the initial genome state with 50% GC content and 6.25% proportion of each dinucleotide are listed in Table 2, and the application results for the other two initial genome states, i.e., with 40 and 60% GC contents, are shown in Supplementary Tables 1 and 2 (Additional file 1), respectively.

Comparison between the expected and observed trends of frequency/proportion changes relative to the initial proportions reveals that, when the initial genomes have a GC content of 50% and proportion for each dinucleotide of 6.25% (see Supplementary Table 3, Additional file 1), most of the 16 dinucleotides in most studied genomes (with the exception of TpA in all 10 genomes, GpA/TpC in cattle and sheep genomes, and ApG/CpT in zebrafish genome; see Tables 1 and 2) have a good consistency between the expected and observed changing trends. This indicates that, on the one hand, 50% GC content could be a rational assumption for the initial state of vertebrate genomes, and on the other hand, our model can achieve a good performance in predicting the changing trends in frequencies of most dinucleotides caused by the methylation-induced CpG to TpG/CpA mutations. It should be noted that, for the dinucleotide TpA, although its proportion/frequency should not be affected by cytidine methylation, the observed frequency in the eight mammalian genomes is either slightly higher or lower than the assumed initial proportion of 6.25% (ranging between 6.01 and 6.59%; see Table 1). Interestingly, in the zebrafish genome, the observed TpA frequency (8.06%) is significantly higher than the assumed initial proportion (6.25%), and in the chicken genome, the observed frequency (5.98%) is lower than those of the eight mammalian genomes.

For each of the 10 tested genomes, the comparison between the observed and expected (obtained with the initial GC content of 50%) frequencies/proportions of the 16 dinucleotides shows an acceptable conformance for most of the 16 dinucleotides (Supplementary Fig. 1, Additional file 1). In order to check whether our model could replicate the observed frequencies/proportions of the 16 dinucleotides, for each of the 10 vertebrate genomes, we have performed the paired t-test of the null hypothesis regarding the differences between the predicted (Table 2) and observed frequencies (Table 1) of the 16 dinucleotides. The results (see Supplementary Fig. 1, Additional file 1) show that all the 10 *P*-values are close to 1 (> 0.05), indicating that the null hypothesis cannot be rejected at the 95% confidence level and,

Table 1 Observed frequencies/proportions of the 16 dinucleotides and GC contents

	ApA/TpT	ApC/GpT	ApG/CpT	ApT	CpA/TpG	CpC/GpG	CpG	GpA/TpC	GpC	TpA	GC%
Proportion$_{obs}$ vs. Proportion$_{ini}$	↑	↓	↑↓	↑	↑	↓	↓	↑↓	↓	↓↑	↓
Bos taurus (cattle)	18.66%	10.20%	14.31%	7.47%	14.73%	10.74%	1.05%	**12.70%**	4.13%	**6.01%**	41.89%
Canis lupus familiaris (dog)	19.57%	9.69%	14.13%	7.75%	13.95%	10.84%	1.09%	12.35%	4.11%	**6.51%**	41.10%
Gallus gallus (chicken)	19.01%	10.40%	14.53%	7.14%	15.37%	9.61%	1.14%	11.89%	4.94%	**5.98%**	41.78%
Pan troglodytes (chimpanzee)	19.55%	10.08%	13.99%	7.71%	14.52%	10.43%	1.01%	11.87%	4.29%	**6.55%**	40.96%
Danio rerio (zebrafish)	22.12%	11.31%	**11.44%**	9.26%	14.64%	6.96%	1.79%	10.50%	3.92%	**8.06%**	36.62%
Homo sapiens (human)	19.63%	10.07%	13.99%	7.74%	14.50%	10.38%	0.98%	11.86%	4.26%	**6.59%**	40.99%
Mus musculus (house mouse)	18.06%	10.69%	14.74%	7.29%	14.95%	10.54%	0.85%	12.44%	4.13%	**6.31%**	41.93%
Papio anubis (olive baboon)	19.56%	10.16%	14.04%	7.62%	14.48%	10.39%	1.05%	11.94%	4.26%	**6.51%**	41.00%
Ovis aries (sheep)	18.65%	10.18%	14.35%	7.44%	14.70%	10.74%	1.08%	**12.68%**	4.17%	**6.01%**	41.95%
Sus scrofa (pig)	19.21%	10.07%	13.90%	7.46%	14.43%	11.09%	1.24%	11.90%	4.42%	**6.27%**	41.91%

Note: Only the autosomes of each genome were included in the statistical analyses; the symbols '↑' and '↓' represent an increase and decrease of the dinucleotide proportions observed (Proportion$_{obs}$) in genomes relative to the assumed initial proportions (Proportion$_{ini}$) obtained based on GC$_{ini}$% = 50% (see Supplementary Table 3, Additional file 1), respectively; the values highlighted in bold exhibit changing trends incompatible with those predicted by MDM (see Table 2)

Table 2 Expected/calculated proportions/frequencies of the 16 dinucleotides and GC contents obtained by MDM (**GC_{ini}**% = **50**%)

	ApA/TpT	ApC/GpT	ApG/CpT	ApT	CpA/TpG	CpC/GpG	CpG	GpA/TpC	GpC	TpA	GC%
$Proportion_{exp}$ vs. $Proportion_{ini}$	↑	↓	↑	↑	↑	↓	↓	↓	↓	↔	↓
Bos Taurus (cattle)	13.82%	12.12%	13.84%	7.71%	17.70%	11.16%	1.05%	11.18%	5.17%	6.25%	44.80%
Canis lupus familiaris (dog)	13.72%	12.22%	13.98%	7.62%	17.66%	11.02%	1.09%	11.28%	5.15%	6.25%	44.84%
Gallus gallus (chicken)	13.60%	11.87%	13.66%	7.99%	17.62%	11.34%	1.14%	11.40%	5.14%	6.25%	44.89%
Pan troglodytes (chimpanzee)	13.74%	12.29%	13.90%	7.66%	17.74%	11.10%	1.01%	11.26%	5.05%	6.25%	44.76%
Danio rerio (zebrafish)	13.62%	12.34%	13.44%	7.54%	16.96%	11.56%	1.79%	11.38%	5.12%	6.25%	45.54%
Homo sapiens (human)	13.67%	12.44%	13.98%	7.59%	17.76%	11.02%	0.98%	11.33%	4.97%	6.25%	44.98%
Mus musculus (house mouse)	13.84%	12.06%	13.90%	7.79%	17.90%	11.10%	0.85%	11.16%	5.14%	6.25%	44.60%
Papio anubis (olive baboon)	13.66%	12.42%	13.98%	7.58%	17.70%	11.02%	1.05%	11.34%	5.01%	6.25%	44.80%
Ovis aries (sheep)	13.78%	12.17%	13.84%	7.69%	17.68%	11.16%	1.08%	11.22%	5.14%	6.25%	44.83%
Sus scrofa (pig)	13.72%	12.22%	13.96%	7.63%	17.66%	11.04%	1.09%	11.28%	5.15%	6.25%	44.99%

Note: The values presented were obtained by application of MDM to the assumed initial genome state with GC_{ini}% = 50%; the symbols'↑', '↓' and '↔'represent an increase, decrease, and no-change of the expected/calculated dinucleotide proportions ($Proportion_{exp}$) relative to the assumed initial proportions ($Proportion_{ini}$; see Supplementary Table 3, Additional file 1), respectively

hence, the differences between the predicted and observed frequencies are statistically acceptable for each tested genome.

In order to evaluate the relative difference between the expected ($Proportion_{exp}$) and observed ($Proportion_{obs}$) dinucleotide frequencies/proportions, the value of ΔProportion, defined as the ratio of |($Proportion_{exp}$ - $Proportion_{obs}$)| to $Proportion_{obs}$, was calculated. Figure 1 shows the ΔProportion values of all 16 nucleotides (calculated using $Proportion_{exp}$ obtained from the assumed initial genome state with GC content of 50%) for the 10 vertebrate genomes investigated. For the eight mammalian genomes, most of the 16 dinucleotides have the ΔProportion values either lower than 10% (ApG/CpT, ApT, CpC/GpG, CpG, GpA/TpC, and TpA) or clustering around 20% (ApC/GpT, CpA/TpG, and GpC), indicating a high similarity between the expected and observed frequencies/proportions of them; only two dinucleotides (i.e., ApA/TpT) in few mammal genomes have the ΔProportion values greater than 30%, which indicate relatively large deviations of the expected frequencies from the observed ones. In contrast, the results for the lower-vertebrate (zebrafish) genome do not seem to be satisfactory because several dinucleotides (e.g., CpC/GpG and ApA/TpT) have the expected proportions/frequencies that radically deviate from the observed ones (**Δ**Proportion > 35%). Also worth noting is that the ΔProportion values of some dinucleotides in the lower-vertebrate genomes (e.g., ApA/TpT, ApG/CpT, ApT, CpC/GpG, and TpA in the zebrafish genome; ApT, CpC/GpG, and GpC in the chicken genome) do not cluster around those of the mammal genomes. Overall,

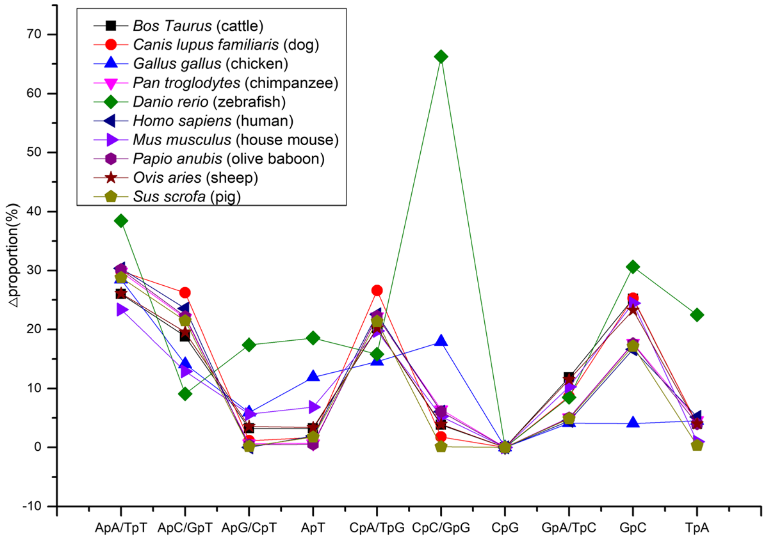

Fig. 1 Relative differences between the expected and observed proportions of the 16 dinucleotides

the above results indicate that, despite the different application effects on the higher-vertebrate (mammalian) and lower-vertebrate genomes, our model displays a better performance when applied to the mammalian genomes.

Changes in the GC content due to cytidine methylation are only related to the number of CpG to TpG/CpA mutations. The relative difference between the expected ($GC\%_{exp}$) and observed ($GC\%_{obs}$) GC contents was evaluated by calculating the ΔGC% value: $\Delta GC\% = |(GC\%_{exp} - GC\%_{obs})| / GC\%_{obs}$. Figure 2 shows ΔGC% values (calculated using $GC\%_{exp}$ obtained from the assumed initial state of a genome with 50% GC content) for the 10 genomes, out of which nine have ΔGC% values less than 10%, indicating a high agreement between the calculated and observed GC contents. The only exception is the zebrafish genome, for which the high ΔGC% value (24.4%) indicates a large deviation of the predicted GC content from the observed one.

Discussion

In this paper, we have proposed a mathematical model based on the neutral theory of molecular evolution to analyze the effects of the methylation-induced CpG to TpG/CpA mutations on the evolution of the genome dinucleotide composition and GC content**.** The model hypothesizes that the neutral mutations (i.e., non-CpG-to-TpG/CpA mutations) would have no effect on the evolution of genome composition. What needs to be highlighted is that the model established here mainly focuses on the methylation-driven evolution of the dinucleotide composition. The previously proposed substitution models, such as the JC69 [22], K80 [23], and TN93 [2] models, despite being widely used in molecular phylogenetic analyses and genome evolution studies, use the rates of the single-nucleotide (1-mer) substitutions as the main parameters. Moreover, these models cannot answer the question of why the observed CpG frequency in the vertebrate genomes is much lower than that expected from the GC content. Although the high rate of the methylation-induced CpG to TpG/CpA mutations can explain the globally reduced frequency of the CpG dinucleotide compared with its expected frequency, there has been a lack of theoretical models with which to predict the mutation effect on the genome composition in terms of 2-mers. Therefore, we propose the MDM through which the methylation-induced changes in frequencies/proportions of the 16 dinucleotides and the GC content can be predicted based on an assumed initial state of a genome.

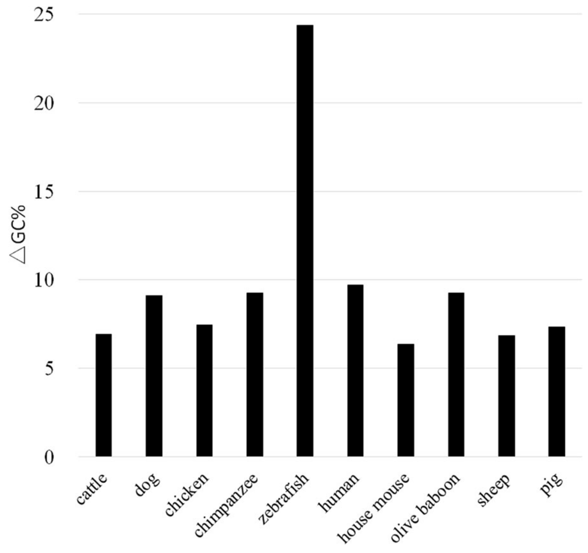

Fig. 2 Relative differences between the expected and observed GC contents

When modeling the genome evolution, it is inevitable to make assumptions regarding some parameters that cannot be directly obtained. As a matter of fact, the effectiveness of our model largely depends on the validity of the assumed parameters, which in turn depends on the composition characteristics of genomes of interest. For example, for a majority of the 16 dinucleotides and the GC content, their expected/calculated frequencies/content in the zebrafish genome are distinctly different from those in the mammalian genomes (Figs. 1 and 2), and this could be attributed to the large difference in the genome composition between zebrafish and mammals [24]. Specifically, the assumed initial state of the genome can have a large influence on the performance of our model. For example, when the dinucleotide proportions in the initial genome state were assigned based on the assumed GC contents of 40 and 60%, the expected/predicted values for more than half of the 16 dinucleotides and GC contents radically deviate from the observed ones in all the 10 tested genomes (Table 1 and Supplementary Tables 1 and 2, Additional file 1). Even in the case of the “reasonable” initial genome state with the assumed GC content of 50%, the expected proportions/frequencies for ApA/TpT differ from the observed ones by more than 25% in most of the 10 genomes (Fig. 1). We consider that it is the assumed initial proportions of ApA/TpT in the genome that leads to the large deviations from the observed values, while the model itself does not account for these.

It should be noted that there are still shortcomings in MDM regarding its application. Since the model is built based on the neutral mutation theory of molecular evolution, it assumes that the numbers of substitutions between any two single nucleotides are equal (e.g., the numbers of mutations from C to A and from A to C are the same) and, hence, all mutations except for CpG to TpG/CpA will have no effect on the evolution of the genome dinucleotide composition. In fact, mutation bias is prevalent in nature. In Kimura's two-parameter model [23], rates of transition (α) and transversion (β) substitutions are different and, furthermore, there is no evidence indicating that the CpG to TpG/CpA mutation is the sole factor influencing the rates of transitions and transversions. Ignoring the effects of non-CpG-to-TpG/CpA mutations on the evolution of the genome dinucleotide composition can be expected to produce deviations from the observed values. Moreover, there are limitations in the parameter estimation methods of the presented model, which assume that the rates of the methylation-induced CpG to TpG/CpA mutations are independent of the sequence context of CpG. In fact, the CpG to TpG/CpA mutations are non-neutral [25] and their substitution rates are sequence context dependent [26, 27].

Conclusions

In this work, we have proposed a mathematical model to investigate the effects of the methylation-induced CpG to TpG/CpA mutations on the evolution of genome composition in terms of the 2-mers and GC content. The application of our model to the 10 vertebrate genomes has achieved a good consistency between the predicted and observed trends of changes in the GC content and frequencies of most of the 16 dinucleotides with respect to their assumed initial values; moreover, for the 10 tested genomes, quantitative evaluations of the relative differences in the dinucleotide frequency and GC content between the expected and observed values show a better performance of our model when applied to the mammalian genomes than to the lower vertebrate genomes. Despite the capability of MDM to quantify the effects of the methylation-induced CpG decay on the evolution of the genome dinucleotide composition and GC content, there are still limitations to the current model because i) the rates of the methylation-induced CpG to TpG/CpA mutations are dependent (rather than independent, as assumed in the current model) on the sequence context of CpG sites and, ii) the proportions of the 16 dinucleotides in the initial state of different vertebrate genomes may not be simply identical but depend on the genome composition characteristics. As a result, future efforts in improving the model should be directed toward i) improving the parameter estimation method to make the estimated parameters reflect the context-dependent rates of methylation-induced CpG to TpG/CpA mutations and, ii) realizing the customization of the initial dinucleotide proportions. These two points could be addressed by assigning appropriate weighting coefficients to the parameters as estimated by the Trinucleotide-method and using the genome-composition-based bias factors to calibrate the proportions of the 16 dinucleotides in the initial state of a genome of interest, respectively.

Methods

According to the neutral mutation theory of molecular evolution [28], most evolutionary changes and most of the variability within species at the molecular level are not caused by natural selection, but by random genetic drift of selectively neutral mutations. Since neutral mutations are those that do not affect the survival or reproduction of an organism, they are expected to have a minor effect on the genome dinucleotide composition in the long-term evolutionary process. However, in vertebrate genomes, CpG hypermutability caused by the cytosine methylation clearly exceeds the random expectation. In vertebrate genomes, the observed frequency of CpG dinucleotides is much lower than that expected based on the GC content, implying that the methylation-induced CpG to TpG/CpA mutations exert a substantial effect on the evolution of the genome dinucleotide composition.

MDM proposed in this study focuses on the effects of the methylation-induced CpG to TpG/CpA mutations on the changes in frequencies/proportions of all 16 dinucleotides and in the GC content. The basic hypothesis of MDM is that the cytidine methylation is a key factor that causes the different rates of the transition (α) and transversion (β) substitutions, which under Kimura's model [23] are considered to drive genome evolution.

Model construction

We assume that in a DNA sequence, each CpG to TpG/CpA substitution is independent from its context. In the case of NpCpG (N = A, C, G, or T), the outcomes of the methylation-induced CpG to TpG mutation on the dinucleotide components can be dissected as follows: the number of CpG is reduced by 1, the number of TpG is increased by 1, and the changes in the numbers of other dinucleotides depend on the nucleotide type of N; for example, if N is A, the number of the dinucleotide ApC will be reduced by 1, while that of ApT will be increased by 1. Let P_A, P_C, P_G, and P_T represent the probabilities of N beating A, C, G and T, respectively, then $P_A + P_C + P_G + P_T = 1$. The changes in the numbers of all 16 dinucleotides in the context of NpCpG upon mutation can be described by the matrix D:

$$\mathrm{D} = (d_{ij}) = \begin{pmatrix} d_{AA} & d_{AC} & d_{AG} & d_{AT} \\ d_{CA} & d_{CC} & d_{CG} & d_{CT} \\ d_{GA} & d_{GC} & d_{GG} & d_{GT} \\ d_{TA} & d_{TC} & d_{TG} & d_{TT} \end{pmatrix} = \begin{pmatrix} 0 & -P_A & 0 & P_A \\ 0 & -P_C & -1 & P_C \\ 0 & -P_G & 0 & P_G \\ 0 & -P_T & 1 & P_T \end{pmatrix} \tag{1}$$

where d_{ij} represents the change in the probability of a dinucleotide composed of nucleotides i and j upon NpCpG to NpTpG mutation (i, j = A, C, G, or T), and a negative element, e.g., $d_{AC} = -P_A$, indicates that the number of the corresponding dinucleotide ApC is reduced, and vice versa. Similarly, changes in the numbers of 16 dinucleotides upon CpGpM to CpApM (M = A, C, G, or T) mutation can be represented by the matrix D'; where P'_A, P'_C, P'_G, and P'_T denote the probabilities of M beating A, C, G, and T, respectively, and $P'_A + P'_C + P'_G + P'_T = 1$.

$$\mathrm{D}' = \left(d'_{ij}\right) = \begin{pmatrix} d'_{AA} & d'_{AC} & d'_{AG} & d'_{AT} \\ d'_{CA} & d'_{CC} & d'_{CG} & d'_{CT} \\ d'_{GA} & d'_{GC} & d'_{GG} & d'_{GT} \\ d'_{TA} & d'_{TC} & d'_{TG} & d'_{TT} \end{pmatrix} = \begin{pmatrix} P'_A & P'_C & P'_G & P'_T \\ 1 & 0 & -1 & 0 \\ -P'_A & -P'_C & -P'_G & P'_T \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

The principle of complementary base pairing dictates that the number of NpCpG to NpTpG mutations on the forward strand is the same as that of CpGpM to CpApM mutations on the reverse strand; furthermore, this principle dictates that the CpG to TpG mutations caused by methylation on the reverse strand corresponds to the CpG to CpA mutations on the forward strand, and vice versa. As a result, both the CpG to TpG and CpG to CpA mutations observed on any one of the two strands are the consequence of cytidine methylation of the CpG dinucleotides, although the summation of their numbers doubles the number of the actually depleted CpG dinucleotides. If the number of CpG to TpG/CpA mutations is set to H, the changes in the numbers of the 16 dinucleotides can be calculated by the matrix Q:

$$\mathrm{Q} = \frac{\mathrm{H}}{2}\left(\mathrm{D} + \mathrm{D}'\right) = \frac{\mathrm{H}}{2}\begin{pmatrix} P'_A & P'_C - P_A & P'_G & P'_T + P_A \\ 1 & -P_C & -2 & P_C \\ -P'_A & -P'_C - P_G & -P'_G & P_G - P'_T \\ 0 & -P_T & 1 & P_T \end{pmatrix} \tag{3}$$

The values of the eight parameters P_A, P_C, P_G, P_T, P'_A, P'_C, P'_G, and P'_T vary between 0 and 1. Before estimating the parameters of the model, we cannot determine the positive or negative values of $(P'_C - P_A)$ and $(P_G - P'_T)$. However, it is clear from the matrix Q that for 14 of the 16 dinucleotides (with the exception of ApC (P'_C - P_A) and GpT (P_G - P'_T)), their changes in number can be directly inferred without requiring parameter estimation. Table 3 lists the changing trends in the numbers of the 14 dinucleotides upon CpG to TpG/CpA mutations inferred directly from the matrix Q.

Parameter estimation

In the matrix Q, the values for the parameters P_A, P_C, P_G, P_T, P'_A, P'_C, P'_G, and P'_T are unknown. To determine the change in the number of each dinucleotide as a function of H (i.e., the number of CpG to TpG/CpA mutations), each parameter value should be estimated. Based on the assumption that the rate of the methylation-induced CpG-to-TpG/CpA mutations is independent of the sequence context of CpG sites, we propose two methods for parameter estimation.

Trinucleotide-method: Using a simple ratio approach, the parameters P_A, P_C, P_G, and P_T can be calculated as the proportion of ApCpG, CpCpG, GpCpG, and TpCpG among all NpCpG trinucleotides, respectively. Accordingly, P'_A, P'_C, P'_G, and P'_T can be obtained through calculating the proportion of CpGpA, CpGpC, CpGpG, and CpGpT among all CpGpM trinucleotides, respectively. Since most CpG islands remain unmethylated in normal cells [29, 30], they are excluded from parameter estimation. In addition, the CpG sites located within the coding regions could also be excluded due to the high selection pressure on these regions. Nevertheless, since the proportion of CpGs in the coding regions out of the total CpGs is small, we expect that the inclusion of the coding-region CpGs would have a negligible effect on the estimation results. Through counting all the eight trinucleotides in the context of NpCpG and CpGpM, the parameters can be calculated by eqs. (4) and (5):

$$\left.\begin{aligned} P_A &= \frac{S_{\mathrm{ACG}}}{S_{\mathrm{ACG}} + S_{\mathrm{TCG}} + S_{\mathrm{CCG}} + S_{\mathrm{GCG}}} \\ P_C &= \frac{S_{\mathrm{CCG}}}{S_{\mathrm{ACG}} + S_{\mathrm{TCG}} + S_{\mathrm{CCG}} + S_{\mathrm{GCG}}} \\ P_G &= \frac{S_{\mathrm{GCG}}}{S_{\mathrm{ACG}} + S_{\mathrm{TCG}} + S_{\mathrm{CCG}} + S_{\mathrm{GCG}}} \\ P_T &= \frac{S_{\mathrm{TCG}}}{S_{\mathrm{ACG}} + S_{\mathrm{TCG}} + S_{\mathrm{CCG}} + S_{\mathrm{GCG}}} \end{aligned}\right\} \tag{4}$$

$$\left.\begin{aligned} P'_A &= \frac{S_{\mathrm{CGA}}}{S_{\mathrm{CGA}} + S_{\mathrm{CGT}} + S_{\mathrm{CGC}} + S_{\mathrm{CGG}}} \\ P'_C &= \frac{S_{\mathrm{CGC}}}{S_{\mathrm{CGA}} + S_{\mathrm{CGT}} + S_{\mathrm{CGC}} + S_{\mathrm{CGG}}} \\ P'_G &= \frac{S_{\mathrm{CGG}}}{S_{\mathrm{CGA}} + S_{\mathrm{CGT}} + S_{\mathrm{CGC}} + S_{\mathrm{CGG}}} \\ P'_T &= \frac{S_{\mathrm{CGT}}}{S_{\mathrm{CGA}} + S_{\mathrm{CGT}} + S_{\mathrm{CGC}} + S_{\mathrm{CGG}}} \end{aligned}\right\} \tag{5}$$

where S_{ACG}, S_{CCG}, S_{GCG}, and S_{TCG} are the numbers of ApCpG, CpCpG, GpCpG, and TpCpG, respectively, and

Table 3 CpG to TpG/CpA mutation-caused changing trends in the numbers of dinucleotides

	ApA/TpT	ApC/GpT	ApG/CpT	ApT	CpA/TpG	CpC/GpG	CpG	GpA/TpC	GpC	TpA
Changing trend	↑	undetermined	↑	↑	↑	↓	↓	↓	↓	↔

Note: The changing trends are inferred directly from the matrix Q, with the symbols'↑', '↓' and '↔' representing an increase, decrease, and no-change, respectively, in the numbers of corresponding dinucleotides; 'undetermined' denotes that the changing trends in the numbers of ApC/GpT cannot be determined from the matrix Q without parameter estimation

S_{CGA}, S_{CGC}, S_{CGG}, and S_{CGT} are the numbers of CpGpA, CpGpC, CpGpG, and CpGpT, respectively, in a genome sequence.

GC-method: Because of the sequence complementarity between the forward and reverse strands of DNA [31], the parameters have the following relationships: $P_A \approx P'_T$, $P_C \approx P'_G$, $P_G \approx P'_C$, and $P_T \approx P'_A$. Therefore, the eight parameters can be estimated based on the GC content in a genome. If the GC content is p, then the AT content is $(1 - p)$, and the following relations can be obtained:

$$\left.\begin{aligned} P_A &= \frac{1-\mathrm{p}}{2} \\ P_C &= \frac{\mathrm{p}}{2} \\ P_G &= \frac{\mathrm{p}}{2} \\ P_T &= \frac{1-\mathrm{p}}{2} \end{aligned}\right\} \tag{6}$$

$$\left.\begin{aligned} P'_A &\approx \frac{1-\mathrm{p}}{2} \\ P'_C &\approx \frac{\mathrm{p}}{2} \\ P'_G &\approx \frac{\mathrm{p}}{2} \\ P'_T &\approx \frac{1-\mathrm{p}}{2} \end{aligned}\right\} \tag{7}$$

For example, if the GC content is 40%, then $P_A = P_T = P'_T = P'_A = 0.3$ and $P_C = P_G = P'_C = P'_G = 0.2$.

It should be noted that in the practical application of MDM, the trinucleotide-method should be preferred over the GC-method. In the application example below, the parameters are estimated based on the current state of a genome and are assumed to be constant during the genome evolution. In fact, the cumulative mutations of CpG to TpG/CpA will inevitably lead to a reduction in the GC content. Therefore, if the GC-method is used, the four estimated parameters (i.e., P_C, P_G, P'_C, and P'_G) will be prone to errors due to the large changes in the GC content between the assumed initial state and current state of a genome.

Application example

The constructed MDM was applied to predict the effects of CpG to TpG/CpA mutations on the evolution of the dinucleotide composition in vertebrate genomes starting from three assumed initial states. The source code used is publicly available at https://github.com/sparkhonghe/MDM.

Ten vertebrate genomes, of which eight are from mammals (higher vertebrates) and two from non-mammals (lower vertebrates, i.e., *Gallus gallus* (chicken) and *Danio rerio* (zebrafish)) (see Supplementary Table 4, Additional file 1), were obtained from the genome database of the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov). To avoid artifacts arising from sex-specific effects, only the sequence from autosomes was included in the statistics of the frequencies/proportions (%) of the 16 dinucleotides and the GC content in these 10 genomes.

There are three basic assumptions in application of MDM: i) the loss of CpG dinucleotides in a vertebrate genome is exclusively caused by cytidine methylation; ii) the rate of the methylation-induced CpG to TpG/CpA mutations is independent of the sequence context of CpG sites; and iii) the total number of dinucleotides in the genome remains constant during evolution. In addition, the model's application also needs to make assumptions about the proportions/frequencies of the 16 dinucleotides in the initial state of a genome, which can be obtained based on the assumed initial GC content (GC_{ini}%). In this study, three assumed initial genome states, i.e., in which the proportions of the 16 dinucleotides were obtained (see eq. (8) below) according to the GC_{ini}% of 40% (see Supplementary Table 5, Additional file 1), 50% (Supplementary Table 3, Additional file 1), and 60% (Supplementary Table 6, Additional file 1), respectively, were tested. It should be noted that, under the same GC content, all the 10 genomes possess the same initial state in terms of the frequency/proportion of each dinucleotide.

Given the GC content, the frequencies/proportions of the 16 dinucleotides (NpM_{ini}%) in the assumed initial state of a genome can be calculated using the following equation:

$$\mathrm{NpM_{ini}}\% = (\mathrm{N_{num}} * \mathrm{M_{num}})/\mathrm{L}^2 \tag{8}$$

where M and N = A, C, G, or T, N_{num} and M_{num} represent the numbers of nucleotides N and M in the initial genome, respectively, and L is the genome length. As a result, N_{num}/L and M_{num}/L designate the frequencies of N and M, respectively. For example, assuming that the GC content is 50% and genome length is 100, then in the initial state of this genome, the number of each of the four nucleotides (A, C, G and T) is 25, and the frequency of each of the 16 dinucleotides (NpM_{ini}%) is 25 * 25/100^2 = 6.25%. Note that the eq. (8) can also be used for estimating the expected frequency of CpG in the current state of a genome if we take N_{num}/L and M_{num} / L as the observed frequencies of C and G (i.e., one-half of the observed GC content), respectively, in the genome.

As rationalized above, the GC content can vary during the evolution of a genome due to continuous CpG depletion. Therefore, we adopt the "Trinucleotide-method" to estimate the parameters P_A, P_C, P_G, P_T, P'_A, P'_C, P'_G, and P'_T in the matrix Q. The total number of the

Table 4 Parameters estimated by statistics of the trinucleotides NpCpG and CpGpM

	P_A	P_C	P_G	P_T	P'_A	P'_C	P'_G	P'_T
Bos Taurus (cattle)	0.2804	0.2590	0.2075	0.2531	0.2525	0.2072	0.2588	0.2815
Canis lupus familiaris (dog)	0.2654	0.2857	0.2128	0.2361	0.2360	0.2127	0.2856	0.2657
Gallus gallus (chicken)	0.3397	0.2284	0.2161	0.2158	0.2157	0.2166	0.2281	0.3396
Pan troglodytes (chimpanzee)	0.2683	0.2679	0.2285	0.2353	0.2349	0.2286	0.2675	0.2690
Danio rerio (zebrafish)	0.2887	0.2092	0.2523	0.2497	0.2498	0.2527	0.2096	0.2879
Homo sapiens (human)	0.2537	0.2818	0.2422	0.2223	0.2218	0.2420	0.2818	0.2544
Mus musculus (house mouse)	0.2856	0.2606	0.2053	0.2485	0.2483	0.2052	0.2605	0.2860
Papio anubis (olive baboon)	0.2553	0.2837	0.2384	0.2226	0.2224	0.2386	0.2837	0.2554
Ovis aries (sheep)	0.2778	0.2598	0.2150	0.2474	0.2466	0.2148	0.2595	0.2790
Sus scrofa (pig)	0.2663	0.2838	0.2137	0.2361	0.2356	0.2135	0.2840	0.2668

Note: Only the autosomes of each genome were included in the statistical analyses

depleted CpG dinucleotides (H) in a genome can be calculated using the following equations:

$$\mathrm{H} = (\mathrm{CpG\%_{ini}} - \mathrm{CpG\%_{obs}}) * \mathrm{N_{NpM}} \tag{9}$$

$$\mathrm{N_{NpM}} = \mathrm{L} - 1 \tag{10}$$

where $\mathrm{CpG\%_{ini}}$ and $\mathrm{CpG\%_{obs}}$ is the assumed initial proportion (see Supplementary Tables 3, 5, and 6, Additional file 1) and the observed frequency of CpG (Table 1) in the genome, respectively, $\mathrm{N_{NpM}}$ is the total number of dinucleotides in the genome, and L is the genome length after removing gaps (Supplementary Table 4, Additional file 1).

The CpG island, which is defined as a stretch of DNA sequence with length > 200 bp, GC content > 50%, and observed-to-expected CpG ratio > 0.6 [13, 32], was identified using the CpG Island Searcher [33]. After removing CpG islands, NpCpG and CpGpM trinucleotides in each of the 10 vertebrate genomes were counted using an in-house Java program (for results, see Supplementary Table 7, Additional file 1), and the eight parameters were then obtained with eqs. (4) and (5). Table 4 lists the estimated parameters for all the 10 vertebrate genomes.

Once the eight parameters (Table 4) were obtained and the number of CpG to TpG/CpA mutations (H) was determined (eq. (9)), the changes in the numbers of the 16 dinucleotides in a genome from the assumed initial state to the current state can be derived from the matrix Q (eq. (3)). Finally, for each of the 16 dinucleotides, its predicted/expected number in the current genome state was obtained by adding the predicted number of changes to the assumed initial number, followed by converting to the proportion of the total number of all 16 dinucleotides to facilitate comparison.

The expected GC content ($\mathrm{GC\%_{exp}}$) in a genome was obtained using the following equation:

$$\mathrm{GC\%_{exp}} = \mathrm{GC_{ini}\%} - \mathrm{H/L} \tag{11}$$

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12976-020-00122-x.

Additional file 1: Supplementary Table 1. Expected/calculated proportions/frequencies of the 16 dinucleotides and GC contents obtained by MDM ($\mathrm{GC_{ini}\%} = 40\%$). **Supplementary Table 2.** Expected/calculated proportions/frequencies of the 16 dinucleotides and GC contents obtained by MDM ($\mathrm{GC_{ini}\%} = 60\%$). **Supplementary Table 3.** Proportions/frequencies of the 16 dinucleotides in the assumed initial state of genomes with $\mathrm{GC_{ini}\%} = 50\%$. **Supplementary Table 4.** Information of the 10 vertebrate genomes. **Supplementary Table 5**. Proportions/frequencies of the 16 dinucleotides in the assumed initial state of genomes with $\mathrm{GC_{ini}\%} = 40\%$. **Supplementary Table 6.** Proportions/frequencies of the 16 dinucleotides in the assumed initial state of genomes with $\mathrm{GC_{ini}\%} = 60\%$. **Supplementary Table 7**. Numbers of the trinucleotides NpCpG and CpGpM in the 10 vertebrate genomes. **Supplementary Fig. 1**. Comparison between the observed and expected frequencies/proportions of the 16 dinucleotides. Note that the expected frequencies were obtained using $\mathrm{GC_{ini}\%} = 50\%$. *P*-value shown in the inserted box was obtained by performing the paired t-test on the observed and expected frequencies of the 16 dinucleotides for each genome.

Abbreviation

MDM: Methylation-driven model

Acknowledgements

Not applicable

Authors' contributions

JHS and SQL conceived and designed the study. JHS wrote the paper. JHS, SMA and SQL reviewed and edited the manuscript. All authors read and approved the manuscript.

Funding

This research has been funded by the National Natural Sciences Foundation of China (Nos. 31370715 and 31160181) and Programs for Excellent Young Talents (XT412003) and Donglu Scholar in Yunnan University.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan & School of Life Sciences, Yunnan University, Kunming 650091, China. [2]College of Engineering, Honghe University, Mengzi 661100, China. [3]Department of Applied Mathematics, Yunnan Agricultural University, Kunming 650201, China.

Received: 9 October 2019 Accepted: 10 March 2020
Published online: 08 April 2020

References

1. Tamura K. The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. Mol Biol Evol. 1992;9:814–25.
2. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10:512–26.
3. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011;29:987–91.
4. Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC Bioinformatics. 2016;17:38.
5. Fiannaca A, La Rosa M, Rizzo R, Urso A. A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network. Artif Intell Med. 2015;64:173–84.
6. Mohamed Hashim EK, Abdullah R. Rare k-mer DNA: identification of sequence motifs and prediction of CpG Island and promoter. J Theor Biol. 2015;387:88–100.
7. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011;21:2167–80.
8. Meher PK, Sahu TK, Rao AR. Identification of species based on DNA barcode using k-mer feature vector and random forest classifier. Gene. 2016;592:316–24.
9. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236.
10. Wang R, Xu Y, Liu B. Corrigendum: recombination spot identification based on gapped k-mers. Sci Rep. 2016;6:35331.
11. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A. 1993;90:11995–9.
12. Furano AV, Walser JC. Mutation rate of non-CpG DNA. In: eLS. Chichester: Wiley; 2009. https://doi.org/10.1002/9780470015902.a0021740.
13. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. J Mol Biol. 1987;196:261–82.
14. Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. Nat Genet. 2000;26:61–3.
15. Bird A. DNA methylation de novo. Science. 1999;286:2287–8.
16. Duret L, Galtier N. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. Mol Biol Evol. 2000;17:1620–5.
17. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. 2006;103:1412–7.
18. Scarano E, Iaccarino M, Grippo P, Parisi E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. Proc Natl Acad Sci U S A. 1967;57:1394–400.
19. Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 2004;333:143–9.
20. Upadhyay M, Samal J, Kandpal M, Vasaikar S, Biswas B, Gomes J, et al. CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J Virol. 2013;87:13816–24.
21. Xiang S, Liu Z, Zhang B, Zhou J, Zhu BD, Ji J, et al. Methylation status of individual CpG sites within Alu elements in the human genome and Alu hypomethylation in gastric carcinomas. BMC Cancer. 2010;10:44.
22. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic; 1969. p. 21–132.
23. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
24. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503.
25. Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet. 2008;4:e1000281.
26. Mugal CF, Ellegren H. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome Biol. 2011;12:R58.
27. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–55.
28. Kimura M. The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet. 1991;66:367–86.
29. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321:209–13.
30. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, et al. The DNA sequence of human chromosome 22. Nature. 1999;402:489–95.
31. Sueoka N. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of a = T and G = C. J Mol Evol. 1999;49:49–62.
32. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 2002;99:3740–5.
33. Takai D, Jones PA. The CpG island searcher: a new WWW resource. In Silico Biol. 2003;3:235–40.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.