Mammalian genome evolution is governed by multiple pacemakers

Sebastián Duchêne and Simon Y. W. Ho*

School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on October 14, 2014; revised on February 17, 2015; accepted on February 20, 2015

Abstract

Genomic evolution is shaped by a dynamic combination of mutation, selection and genetic drift. These processes lead to evolutionary rate variation across loci and among lineages. In turn, interactions between these two forms of rate variation can produce residual effects, whereby the pattern of among-lineage rate heterogeneity varies across loci. The nature of rate variation is encapsulated in the pacemaker models of genome evolution, which differ in the degree of importance assigned to residual effects: none (Universal Pacemaker), some (Multiple Pacemaker) or total (Degenerate Multiple Pacemaker). Here we use a phylogenetic method to partition the rate variation across loci, allowing comparison of these pacemaker models. Our analysis of 431 genes from 29 mammalian taxa reveals that rate variation across these genes can be explained by 13 pacemakers, consistent with the Multiple Pacemaker model. We find no evidence that these pacemakers correspond to gene function. Our results have important consequences for understanding the factors driving genomic evolution and for molecular-clock analyses.

Availability and implementation: ClockstaR-G is freely available for download from github (https://github.com/sebastianduchene/clockstarg).

Contact: simon.ho@sydney.edu.au

Supplementary information: Supplementary data are available at Bioinformatics online.
three factors at the genomic scale can be described using the pacemaker models of molecular evolution (Fig. 1; Ho 2014; Snir et al. 2012). The Universal Molecular Clock (MC) model, the evolutionary rate is constant among lineages but varies across genes. In the Universal Pacemaker (UPM) model, the evolutionary rate varies across genes and among lineages. However, all genes share the same pattern of among-lineage rate variation. In the Multiple Pacemaker (MPM) model, groups of genes share the same pattern of among-lineage rate variation. In the Degenerate Multiple Pacemaker (DMPM) model, each gene has a distinct pattern of among-lineage rate variation.

The model rate variation is governed by gene effects only. The Universal Pacemaker (UPM) model assumes that both lineage and gene effects are present, but that residual effects are negligible. Therefore, a single pattern of among-lineage rate variation is shared across all loci. The Multiple Pacemaker (MPM) model suggests that there is a limited number of pacemakers, such that there are groups of genes with the same pattern of among-lineage rate variation, but with different absolute rates. This implies that residual effects within pacemakers are very small compared with those between pacemakers. In the Degenerate Multiple Pacemaker (DMPM) model, residual effects are pervasive and there is a distinct pattern of among-lineage rate variation for each gene. This extreme case of gene-specific among-lineage rate variation is also known as ‘erratic’ evolution (Ayala et al., 1996; Rodríguez-Trelles et al., 2001).

Support for the UPM model has come from studies of various organisms, including archaea, bacteria, plants, fungi and Drosophila species (Snir et al., 2012, 2014; Wolf et al., 2013). In these analyses, the UPM model was preferred over the MC, the MPM, and DMPM models. These studies compared models using goodness-of-fit statistics, but these are limited because they do not allow testing of all of the possible scenarios under the MPM model. The MPM hypothesis is much more complex than the UPM or the DMPM because it comprises a large family of models: the $n$ loci in a data set can be assigned to one of $k$ pacemakers, where $k$ can take any value from 2 to $n - 1$. Given the large number of ways in which the loci can be assigned to pacemakers, testing the goodness-of-fit of every scenario under the MPM model is computationally intractable. This problem can be overcome by using clustering algorithms (Duchêne and Ho, 2014; Snir, 2014).

Here, we analyze a publicly available data set of 431 genes from 29 mammalian taxa to test the different pacemaker models. We tested the UPM model using a clustering algorithm implemented in a new version of our program, ClockstaR (Duchêne et al., 2014), which we have modified so that it can be used to analyze genome-scale data sets. We find that the patterns of evolutionary rate variation in the mammalian data set can be explained by 13 pacemakers, supporting the MPM model. Our results suggest that there is a degree of stability in genome evolution. However, we find no evidence of association between pacemakers and gene function.

2 Methods

We compiled a publicly available data set comprising 431 nuclear protein-coding genes from 33 mammalian taxa (Song et al., 2012). We removed four taxa (horse, New World bat, Old World bat and tree shrew) because their phylogenetic placement in most gene trees did not match that in the species tree, leaving 29 taxa. This is a crucial requirement of our method, which assumes the same topology across all gene trees. We did not analyze the data sets from previous studies of the pacemaker models (Snir et al., 2012, 2014; Wolf et al., 2013) because they do not meet this assumption. Moreover, they contain large proportions of missing data, which can mislead the estimates of tree distances that form an important component of our analysis.

To evaluate the MPM model, we estimated gene-trees under maximum-likelihood implemented in phangorn v1.99 (Schliep, 2011). For each gene we selected the nucleotide substitution model according to the Bayesian information criterion. We optimized the branch lengths by fixing the tree topology to the species tree inferred for these data (Song et al., 2012). We analyzed the gene trees using a modified version of the ClockstaR algorithm (Duchêne et al., 2014), ClockstaR-G.

The algorithm in ClockstaR-G involves scaling the length of each gene tree to 1. This step controls for the possibility that genes differ in their absolute evolutionary rates, and allows us to focus on the proportional differences among branch lengths (i.e. the pattern of among-lineage rate variation). The scaled branch lengths are used as individual dimensions in Euclidean space to calculate the distance between trees. This is possible because all genes are assumed to share the same topology, such that we always compare the same branches among genes. In this respect, trees with similar patterns of among-lineage rate variation are expected to have small pairwise distances between them. A pacemaker is a cluster of trees in this space. For example, under the MC and the UPM models (Fig. 1), there is a single cluster of gene trees with pairwise distances of 0. In the MPM model, there is more than one cluster of gene trees, and the gene-tree distances within each cluster are necessarily smaller than those between clusters. In the DMPM model, the gene trees are randomly distributed across Euclidean space, such that it is not possible to identify discrete clusters. Clustering algorithms can be used to identify the pacemakers. ClockstaR-G uses the Clustering for Large Applications (CLARA) algorithm, which is efficient for large data sets (Kaufman and Rousseeuw, 2005).

The optimal number of clusters can be identified via the Gap statistic (Tibshirani et al., 2001). We use CLARA to assign genes to each of the $k$ clusters and we calculate the average silhouette width
Mammalian genome pacemakers

3 Results and discussion

The MC model provided a poor fit to the concatenated alignment of mammalian genes. The log-likelihood of the unconstrained model (−8 400 383) was much higher than that of the strict-clock model (−8 463 250), such that the MC model was strongly rejected (P < 0.01).

Our test of the MPM model suggested that the optimal number of pacemakers for this data set is 13, with (Gap13) = 0.14, compared as the target variable, as implemented in the R packages Tree v1.0 (Ripley, 2014) and RandomForest v4.6 (Liaw and Wiener, 2002).

The classification-tree algorithm uses one gene feature at a time as a binary classifier, known as a decision node. The features are added recursively to the tree until all the genes are classified into pacemaker. Notably, a particular feature can be used at multiple nodes in the tree. To avoid over-fitting, the least informative nodes are pruned by using a k-fold cross-validation (Hastie et al., 2009).

The features can be ranked in terms of their importance by counting the number of times that they are used in the tree. The random forest is an ensemble learning method that uses a large number of classification trees (Breiman, 2001). In this algorithm, feature importance is determined by removing one feature at a time, and computing the increase in misclassification error. We chose these methods over other classification algorithms, such as logistic regression, because they do not make parametric assumptions and because the inclusion of categorical predictor variables is straightforward.

Fig. 2. Statistical fit of different numbers of pacemakers to 431 genes from 29 mammalian taxa and one replicate of 431 genes simulated under the Universal Pacemaker (UPM) model. Statistical fit is measured by the Gap statistic, Gapk. Results are shown for pacemakers (k) ranging from 2 to 100 and values of Gapk for k = 100 are shown elsewhere (Supplementary Fig. S2). The solid line represents the average Gapk, and the gray points represent the values estimated from 500 bootstrap replicates for each number of pacemakers. Large values of Gapk indicate high statistical fit. To select the optimal number of pacemakers, we consider the value of k with the highest average Gapk, and for which the 95% confidence interval does not overlap with those of Gapk−1 and Gapk+1. For the mammalian data set, the asterisk (*) denotes the optimal number of pacemakers. In the UPM simulations, the values for Gapk overlap among all successive values of k, so it is not possible to identify an optimal number of pacemakers.
Table 1. Features of genes assigned to 13 pacemakers in the mammalian genome, ordered by number of genes

<table>
<thead>
<tr>
<th>Pacemaker</th>
<th>Number of genes</th>
<th>Mean tree length (subs/site)</th>
<th>Mean isolation</th>
<th>Mean dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM1</td>
<td>257</td>
<td>1.63</td>
<td>1.77</td>
<td>2.98</td>
</tr>
<tr>
<td>PM2</td>
<td>60</td>
<td>1.38</td>
<td>1.79</td>
<td>3.03</td>
</tr>
<tr>
<td>PM3</td>
<td>31</td>
<td>1.67</td>
<td>1.77</td>
<td>3.45</td>
</tr>
<tr>
<td>PM4</td>
<td>22</td>
<td>1.83</td>
<td>1.33</td>
<td>3.51</td>
</tr>
<tr>
<td>PM5</td>
<td>22</td>
<td>1.52</td>
<td>1.26</td>
<td>3.37</td>
</tr>
<tr>
<td>PM6</td>
<td>14</td>
<td>1.77</td>
<td>1.37</td>
<td>3.37</td>
</tr>
<tr>
<td>PM7</td>
<td>5</td>
<td>1.47</td>
<td>1.07</td>
<td>3.91</td>
</tr>
<tr>
<td>PM8</td>
<td>5</td>
<td>1.37</td>
<td>0.82</td>
<td>2.69</td>
</tr>
<tr>
<td>PM9</td>
<td>5</td>
<td>1.30</td>
<td>1.03</td>
<td>3.50</td>
</tr>
<tr>
<td>PM10</td>
<td>4</td>
<td>1.40</td>
<td>1.02</td>
<td>3.21</td>
</tr>
<tr>
<td>PM11</td>
<td>3</td>
<td>1.47</td>
<td>1.05</td>
<td>3.33</td>
</tr>
<tr>
<td>PM12</td>
<td>2</td>
<td>1.83</td>
<td>0.72</td>
<td>1.78</td>
</tr>
<tr>
<td>PM13</td>
<td>1</td>
<td>0.94</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

would be the dominant source of rate variation. In contrast, genes that yield short trees might be subject to stronger selective constraints, such that they are governed by residual effects. Consequently, there would be fewer pacemakers for genes with long trees than for genes with short trees. This result stands in contrast with those of Du et al. (2013), who did not find any relationship between gene-specific evolutionary rates and the degree of among-lineage rate heterogeneity. Further data from a range of taxonomic groups will enable this hypothesis to be tested more comprehensively.

Wolf et al. (2013) proposed that there is stability in among-lineage rate variation throughout the genome. They showed that the UPM model explains a large proportion of the gene-specific rate variation among bacterial and archaeal lineages. Snir et al. (2014) also found support for the UPM model in Drosophila and yeast genomes. Our analyses of the mammalian data set favor the MPM model, but the number of pacemakers is small compared with the total number of genes (13 pacemakers for 431 genes). This lends support to the hypothesis of stable genome evolution because genome-wide rate variation can be clustered into a small number of pacemakers.

Our method has appeared to perform well in analyses of data from simulations and mammals, but might be less reliable when the level of within-cluster variation is high compared with that between clusters. Further investigation of its performance under a wider range of simulation scenarios, including its sensitivity to potential confounding factors, will be valuable. In addition, improved genome annotation and understanding of gene-specific functions will enable elucidation of the factors that determine the number and influence of genomic pacemakers. This opens the way for improving schemes for assigning multiple relaxed-clock models in phylogenetic analyses of evolutionary timescales (dos Reis et al., 2012; Duchêne and Ho, 2014; Ho and Duchêne, 2014). Furthermore, analyzing the genomes of a wider range of organisms will help to determine whether stability is a ubiquitous property of genome evolution. Our method provides a potential means of answering such questions about patterns of evolutionary rate variation on a genomic scale.

Funding
SD was supported by a Francisco José de Caldas Scholarship from the Colombian government and by a Sydney World Scholars Award from the University of Sydney. SYWH was supported by a Queen Elizabeth II Fellowship from the Australian Research Council (grant number DP10100383).

Conflict of Interest: none declared.

References