Bioinformatics, 36(17), 2020, 4662–4663
doi: 10.1093/bioinformatics/btaa585
Advance Access Publication Date: 23 June 2020
Applications Note

Systems biology

The Epigenetic Pacemaker: modeling epigenetic states under an evolutionary framework

Colin Farrell [1], **Sagi Snir**[2,*,†] **and Matteo Pellegrini**[3,*,†]

[1]Department of Human Genetics, University of California, Los Angeles, CA, USA, [2]Department of Evolutionary Biology, University of Haifa, Haifa, Israel and [3]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, CA, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: Pier Luigi Martelli

Received on March 29, 2020; revised on June 11, 2020; editorial decision on June 12, 2020; accepted on June 15, 2020

Abstract

Summary: Epigenetic rates of change, much as evolutionary mutation rate along a lineage, vary during lifetime. Accurate estimation of the epigenetic state has vast medical and biological implications. To account for these non-linear epigenetic changes with age, we recently developed a formalism inspired by the Pacemaker model of evolution that accounts for varying rates of mutations with time. Here, we present a python implementation of the Epigenetic Pacemaker (EPM), a conditional expectation maximization algorithm that estimates epigenetic landscapes and the state of individuals and may be used to study non-linear epigenetic aging.
Availability and Implementation: The EPM is available at https://pypi.org/project/EpigeneticPacemaker/ under the MIT license. The EPM is compatible with python version 3.6 and above.
Contact: ssagi@research.haifa.ac.il or matteop@mcdb.ucla.edu

1 Introduction

Methylation of cytosine plays an integral role in the regulation of gene expression and mammalian development (Jaffe *et al.*, 2016; Li et al., 1992; Okano *et al.*, 1999; Tate *et al.*, 1993). During the mammalian life cycle, age associated changes in DNA methylation proceed predictably and non-linearly with time (Snir *et al.*, 2019). The systematic changes of DNA methylation with age have led to the development of several epigenetic clocks (Hannum et al., 2013; Horvath, 2013). Most of these models assume that the change in methylation is linear with age, and as such are reminiscent of the molecular clock concept in molecular evolution. The predicted age from these models can be interpreted as a physiological or epigenetic age, and the residual error between the expected and predicted epigenetic age has been associated with several health outcomes (Horvath and Levine, 2015; Horvath et al., 2015; Perna et al., 2016). However, these approaches make a priori assumptions about the functional relationship between epigenetic changes and age (e.g. linearity) and hence may fail to adequately capture non-linear changes in methylation with age. This is important because there is substantial evidence to suggest that epigenetic changes are much more rapid early in life and progressively slow as we age across tissue type (Snir *et al.*, 2019).

To overcome the limitations of prior approaches we developed an evolutionary based approach—the Epigenetic Pacemaker (EPM)—for modeling epigenetic states as evolving entities (Snir *et al.*, 2016, 2018). The EPM borrows from the Universal Pacemaker formalism (UPM) (Snir *et al.*, 2012) under which the evolutionary rate of genes remains constant relative to one another but the absolute rate can change arbitrarily by factors affecting the evolving lineage. In contrast to the EPM, most previous epigenetic clocks resemble the molecular evolutionary concept of the Molecular Clock (Zuckerkandl *et al.*, 1965), where the evolutionary rate of genes remains constant with time. In the EPM, given a set of i methylation sites and j individuals, the observed methylation status, $\hat{m}_{ij}$, is given as $\hat{m}_{ij} = m_i^0 + r_i s_j + \in_{ij}$, where m_i^0 is the initial methylation value, r_i is the rate of methylation change, s_j is the epigenetic state, and $\in_{ij}$ is a normally distributed error term. Given an input matrix $\hat{M} = [\hat{m}_{ij}]$ the goal of the EPM is to find the optimal values of r_i, m_i^0 and s_j to minimize the error between the measured and predicted methylation values. Our approach is distinct from previous epigenetic clock methods that attempt to minimize the difference between observed and predicted age, thus implicitly constraining the functional form of that relationship.

The EPM optimization is accomplished through an implementation of a fast conditional expectation maximization algorithm that we have previously shown maximizes the model likelihood by minimizing the residual sum of squares error (Snir *et al.*, 2016). When fitting the EPM each methylation site is assigned an independent rate of change, and starting methylation value, and each individual is assigned an epigenetic state. We use chronological age as an initial guess for the epigenetic state, which is then updated through each iteration to minimize the error across the observed epigenetic

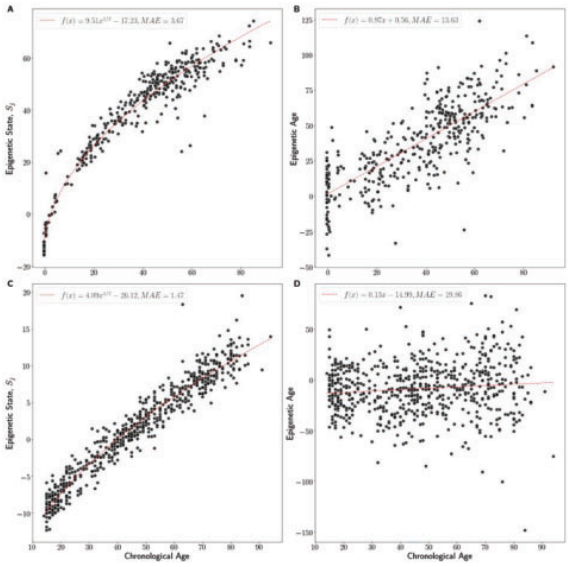

Fig. 1. Epigenetic state predictions for ($n = 405$) test samples compared to the chronological age of each sample with a line of best fit for the EPM (**A**) and linear regression (**B**) models. The non-linear trend observed in the EPM model better captures the observed aging trend and reduces observed error as measured by mean absolute error (MAE). (**C**) Epigenetic state predictions made for whole blood samples using the EPM and (**D**) linear model

landscape (i.e. the parameter set of our model). Because we model methylation and not age, the EPM relaxes the condition of linearity between a trait of interest (e.g. age) and the observed methylation values. This allows the EPM to model non-linear relationships between our state, s_i, and the trait, without needing to transform the trait of interest (as is done in certain epigenetic clocks).

2 Epigenetic Pacemaker

To highlight the utility of the EPM, we fit EPM and linear regression models using publicly available Illumina HumanMethylation450 (450k) microarray data (Sandoval *et al.*, 2011) generated from human brain tissue samples ($n = 657$, $0 - 96$ years) (Jaffe *et al.*, 2016). Briefly, we performed stratified sampling by age to select 270 brain tissue samples for site selection and model training. CpG sites were selected for model inclusion using the absolute value of the Pearson correlation coefficient between the training methylation values and chronological age, (PCC $\geq |0.85|$, $n = 254$). We then fit the EPM and regression models (Pedregosa, 2011; Sandoval *et al.*, 2011) using the selected sites and training methylation data. Age and epigenetic state predictions were made for the remaining brain tissue samples ($n = 405$) left out of model training. The EPM model shows the non-linear relationship between epigenetic state and chronological age (Fig. 1A) that is lost in the regression model (Fig. 1B). We then used the brain EPM and linear models to predict the epigenetic state of 450k data ($n = 732$, $14 - 96$ years) generated from whole blood tissue (Johansson *et al.*, 2013). Samples with missing methylation values for the CpG sites used in model generation were dropped, resulting in 634 analysis samples. The brain EPM model captures aging in the whole blood samples with minimal error (Fig. 1C), while the aging signal is largely lost in the linear model (Fig. 1D).

We have developed an optimized version of the EPM algorithm implemented as a python package (da Costa-Luis, 2019; Virtanen *et al.*, 2020; Walt *et al.*, 2011) that adopts Scikit-Learn (Pedregosa, 2011) style syntax for easy incorporation into current workflows with support for cross validation. The EPM is available through the python package repository, https://pypi.org/project/EpigeneticPacemaker/, under a MIT license. Full documentation, including tutorials, and source code can be found at https://epigeneticpacemaker.readthedocs.io and https://github.com/NuttyLogic/EpigeneticPacemaker, respectively.

Financial Support: This work was supported by the National Institutes of Health (T32CA201160 to C.F.).

Conflict of Interest: none declared.

References

Pedregosa,F. et al. (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn.* Res., **12**, 2825–2830.

da Costa-Luis,C.O. (2019) tqdm: a fast, extensible progress meter for Python and CLI. *J. Open Source Softw.*, **4**, 1277.

Hannum,G. *et al.* (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell*, **49**, 359–367.

Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, R115.

Horvath,S. and Levine,A.J. (2015) HIV-1 infection accelerates age according to the epigenetic clock. *J. Infect. Dis.*, **212**, 1563–1573.

Horvath,S. *et al.* (2015) Accelerated epigenetic aging in Down syndrome. *Aging Cell*, **14**, 491–495.

Jaffe,A.E. *et al.* (2016) Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.*, **19**, 40–47.

Johansson,A. *et al.* (2013) Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS One*, **8**, e67378.

Li,E. *et al.* (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

Okano,M. *et al.* (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.

Perna,L. *et al.* (2016) Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin. Epigenet.*, **8**, **64**.

Sandoval,J. *et al.* (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.

Snir,S. *et al.* (2012) Universal pacemaker of genome evolution. *PLoS Comput. Biol.*, **8**, e1002785.

Snir,S. *et al.* (2016) A statistical framework to identify deviation from time linearity in epigenetic aging. *PLoS Comput. Biol.*, **12**, e1005183.

Snir,S. *et al.* (2018) An epigenetic pacemaker is detected via a fast conditional expectation maximization algorithm. *Epigenomics*, **10**, 695–706.

Snir,S. *et al.* (2019) Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, **14**, 912–926.

Tate,P.H. *et al.* (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opin. Genet. Dev.*, **3**, 226–231.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, **17**, 261–272.

Walt,S. v d. *et al.* (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.

Zuckerkandl,E. *et al.* (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, 97–166.