Supplemental text

1. Defining transition rate matrices for epimutation data

For DNA mutations, there are a set of substitution models that are widely used in maximum likelihood and Bayesian phylogenic inference. These models also allow one to estimate the *genetic* distance between two DNA sequences. These distances form the basis for distance based phylogenetics methods, such as neighbor joining method [1] or UPGMA [2]. It is possible to modify classical substitution models for DNA methylation data (Table S1) and use them to derive the *epigenetic* distance between two methylomes (Section 2). Knowing this distance along with calibrated epimutation rates is a starting point for estimating the actual time since two methylomes have diverged from a recent common ancestor.

As discussed in the main text, an interesting application is the epigenetic analysis of clonal populations. Since segregation and recombination are absent in clones, new epimutations get immediately "fixed" in the form of new epigenotypes. To motivate this, consider the methylation data from a diploid species. We can use a continuous Markov model with a 3 by 3 rate matrix, Q, to represent the transitions among three epigenotypes UU (homozygous unmethylated, first row in matrix), UM (epi-heterozygous, second row) and MM (homozygous methylated, third row). In Table S1, we show a number of possible Q matrices for methylation data. These Q matrices are based on classical DNA-based substitution models. For haploid methylation data, each site has only two possible states: methylated (M) and unmethylated (U). The 2 by 2 transition rate matrices can be applied as well.

By presenting transitions between epigenotypes in the framework of a substitution model, we use the term "substitution" in a wide sense. It should be understood that such substitutions are themselves function of stochastic methylation gain and loss rates [3]. However, mathematically this conceptualization is tenable, and sufficient for us to bridge the observed transition among epigenotypes and divergence. To avoid unnecessary arguments in terminologies, we may define the divergence inferred with epigenotype data as "epigenetic distance".

2. Estimating epigenetic distances between two methylomes: a working example

As a working example, consider the following Q matrix (see also Table S1), which is based on similar assumptions as the K80 substitution model of Kimura [4].

(1)

$$Q = \begin{bmatrix} -(c+d) & c & d \\ c & -2c & c \\ d & c & -(c+d) \end{bmatrix}$$

Based on the matrix Q, the stationary distribution of three epigenotypes UU, UM and MM are: $\pi = [\pi_{UU}, \pi_{UM}, \pi_{MM}] = [\pi_1, \pi_2, \pi_3] = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. The total substitution rate for any CG site is,

$$\mu = -\sum_{i}^{3} \pi_i q_{ii} = \frac{2(2c+d)}{3}$$

and the epigenetic distance between two sequences by time t is

(2)

$$f = -\sum_i^3 \pi_i q_{ii} \cdot t = \frac{2(2c+d)}{3} \cdot t.$$

However, parameters c, d, and t are unknown. To be able to obtain a sample estimator of f, consider the following argument: When comparing two methylomes there are three kinds of possible transitions among epigenotypes (UU ↔ MM, UU ↔ UM, and MM ↔ UM). These can be readily detected from a differential methylation analysis. Let the observed proportions of these transitions be S_1, S_2 and S_3 . Based on rate matrix Q, their expected proportions at evolutionary time t can be obtained from the transition probability matrix $P(t)$, which has the form:

(3)

$$P(t) = e^{Qt} = \begin{bmatrix} \frac{1}{3} + \frac{1}{6}e_1 + \frac{1}{2}e_2 & \frac{1}{3} - \frac{1}{3}e_1 & \frac{1}{3} + \frac{1}{6}e_1 - \frac{1}{2}e_2 \\ \frac{1}{3} - \frac{1}{3}e_1 & \frac{1}{3} + \frac{2}{3}e_1 & \frac{1}{3} - \frac{1}{3}e_1 \\ \frac{1}{3} + \frac{1}{6}e_1 - \frac{1}{2}e_2 & \frac{1}{3} - \frac{1}{3}e_1 & \frac{1}{3} + \frac{1}{6}e_1 + \frac{1}{2}e_2 \end{bmatrix}, e_1 = e^{-3ct}, e_2 = e^{-(c+2d)t}$$

It can be shown that these expected proportions are:

$$E(S_1) = E(S_3) = \frac{2}{9} - \frac{2(e^{-3c})}{9}, E(S_2) = \frac{2}{9} + \frac{(e^{-3c})}{9} - \frac{\left(e^{-(c+2d)t}\right)}{3}$$

Now, let the ratio between d and c be k = d/c, and setting the above expectations equal to the observed proportions S_1, S_2 and S_3 , we obtain:

(4)

$$\widehat{ct} = -\frac{1}{3}\log\left(1 - \frac{9}{2}S_1\right)$$

(5)

$$\hat{k} = \frac{3\log\left(1 - \frac{3}{2}S_1 - 3S_2\right)}{2\log\left(1 - \frac{9}{2}S_1\right)} - \frac{1}{2}$$

We can easily verify equation (5) by setting $S_1 = S_2$. Then, the $\hat{k}$ will be 1.

Finally, substituting (4) and (5) these into equation (2), our sample estimator of the epigenetic distance is:

$$\hat{f} = \frac{4\widehat{ct}}{3} + \frac{2\hat{k}\widehat{ct}}{3}.$$

Closed form estimators of epigenetic genetic distance are not always available, particularly when the matrix Q is more complex. In such cases, estimates can be obtained via maximum likelihood methods [5-7]. The general form of the log-likelihood function is given:

(6)

$$l(\Theta) = \sum_{i}\sum_{j} n_{ij} log\ (\pi_i P_{ij}(t))$$

Where n_{ij}, is the number of sites occupied by epigenotype i and j in two sequences. $P_{ij}(t)$is transition probability from i to j, which is an element in transition probability $P(t)$. Maximization is with respect to parameter vector Θ, which contains of the unknown rate parameters as well as evolutionary time t.

References

1. Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), 406-425.

2. Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. Univ. Kansas, Sci. Bull., 38, 1409-1438.

3. Shahryary, Y. et al. (2020). AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. Genome biology, 21(1), 1-22.

4. Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution, 16(2), 111-120.

5. Sullivan, J. et al. (1995). Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Molecular Biology and Evolution, 12(6), 988-1001.

6. Yang, Z., and Kumar, S. (1996). Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Molecular Biology and Evolution, 13(5), 650-659.

7. Yang, Z. (2014). Molecular evolution: a statistical approach. Oxford University Press.