multiFaAcceleration: A program for the measurement of mutation velocity and acceleration from a four-species multiple alignment

Riley J. Mangan

July 30, 2021

1 Usage

multiFaAcceleration - Performs velocity and acceleration on a four way multiple alignment in multiFa format.

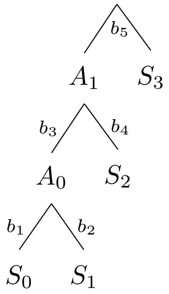
A four way multiple alignment must contain four species (index 0 to 3) in the topology that aln[0] is the most derived and species 1 to 3 are successive outgroups.

Three bed files are returned. The first produces the velocity score, the second returns the acceleration score, and the third returns the initial velocity score for each window of the genome for aln[0].

multiFaAcceleration chromName in.fa velocity.bed acceleration.bed initialVelocity.bed

2 Branch Length Calculation

Consider a phylogenetic tree with extant species $S_0 : S_3$, extinct ancestors $A_0 : A_1$, and branch lengths $b_1 : b_5$ with the following topology.



Note that in the above tree, b5 represents the entire distance between A_1 and S_3. Consider that we can measure the pairwise mutation distance between any two extant species on this tree, represented by the Greek letter π. It follows that the pairwise distance between two extant species is equal to the sum of branch lengths separating those species on the phylogenetic tree shown above. Thus, we are able to produce the following system of linear equations.

$$\begin{aligned} b_1 + b_2 &= \pi(S_0, S_1) \\ b_1 + b_3 + b_4 &= \pi(S_0, S_2) \\ b_2 + b_3 + b_4 &= \pi(S_1, S_2) \\ b_1 + b_3 + b_5 &= \pi(S_0, S_3) \\ b_2 + b_3 + b_5 &= \pi(S_1, S_3) \\ b_4 + b_5 &= \pi(S_2, S_3) \end{aligned} \tag{1}$$

If our interest is to study the genome evolution of S_0, we can define the mutation distance as b_1, the distance between that extant species and its most recent common ancestor with S_1. We can then define the initial mutation distance as b_3, the distance along the previous branch between A_0 and its ancestor A_1.

Below is the solved expression for b_1.

$$\boxed{b_1 = \frac{\pi(S_0, S_1) + \pi(S_0, S_2) - \pi(S_1, S_2)}{2}} \tag{2}$$

This result can be verified using the above system of equations and a bit of algebra.

$$b_1 = \frac{(b_1 + b_2) + (b_1 + b_3 + b_4) - (b_2 + b_3 + b_4)}{2}$$

$$b_1 = \frac{2b_1}{2}$$

The equation for b_3 is as follows.

$$\boxed{b_3 = \frac{\pi(S_1, S_2) + \pi(S_0, S_3) + \pi(S_2, S_3) - \pi(S_0, S_1)}{2} - \pi(S_2, S_3)} \tag{3}$$

We can also verify this expression using the same system of equations.

$$b_3 = \frac{(b_2 + b_3 + b_4) + (b_1 + b_3 + b_5) + (b_4 + b_5) - (b_1 + b_2)}{2} - (b_4 + b_5)$$

$$b_3 = \frac{2b_3 + 2b_4 + 2b_5}{2} - b_4 - b_5$$

3 Algorithm

For a given four-way alignment in multiFa format, $gonomics : multiFaAcceleration$ calculates b_1 and b_3 using pairwise mutation distance (defined as the number of SNPs and INDELs, where each INDEL counts as one mutation regardless of length) for each window of a user-specified window size. Windows may be every possible window of the genome, or may be restricted to a particular subset of the genome using the option $-searchSpaceBed$, which enables the input of a bed file which specifies the regions that should be considered. The option $-searchSpaceProportion$ enables the user to consider all windows in which at least a user-specified proportion of bases are within the searchSpace.

We define $\mathbf{v}$ as the normalized mutation velocity, or the normalized rate of mutation over the branch b_1. To calculate $\mathbf{v}$, we calculate the average b_1 length $\overline{b_1}$ across all windows. For each window:

$$\mathbf{v} = \frac{b_1}{\overline{b_1}}$$

Similarly, the normalized initial rate of mutation, or the normalized rate of mutation over the branch b_3, can be calculated as:

$$\mathbf{v}_0 = \frac{b_3}{\overline{b_3}}$$

Where $\overline{b_3}$ is the average value of b_3 over all windows.

$\mathbf{v}$ and $\mathbf{v}_0$ have intuitive numerical interpretations. If $\mathbf{v} = 1$ for a particular window, the mutation rate in the branch b_1 is equal to the chromosome-wide average mutation rate. $\mathbf{v} = 2$ would be found in a region evolving twice as quickly, and $\mathbf{v} = 0.5$ in a region evolving at half the average rate. The same interpretations apply for $\mathbf{v}_0$, the rate of evolution along the branch b_3.

Finally, we define the quantity $\mathbf{a}$, for acceleration, as the normalized change in mutation rate between branches branches $b1$ and $b3$:

$$\mathbf{a} = \mathbf{v} - \mathbf{v}_0$$

The quantity $\mathbf{a}$ is equal to zero when the mutation rate along b_1 is equal to the mutation rate along b_2. As both $\mathbf{v}$ and $\mathbf{v}_0$ are normalized, this holds true even if b_1 and b_3 are not equal in absolute length, which will be the case when the extant species S_0, S_1, and S_2 are not separated by equal amounts of evolutionary time. Positive values for $\mathbf{a}$ indicate accelerated regions, and negative values suggest regions under negative acceleration, in which a region evolved at a slower rate along b_1 than b_3.