

Some Definitions

- n_t : The number of time points (cardinality) in the set $\{(f_{ai}, f_{bi}) | f_{detected} < f < f_{fixed}\}$.
- f_{ai} : The frequency of mutation a at time point i
- f_i : The mean of two mutations at time point i
- σ_i^2 : The variance of the paired time series at time i
- \bar{d} : The average of the differences between both time series
- σ_p^2 : Variance for the pair of mutational time series.

The Math

The script calculates the relative similarity between all pairs of mutational time series with n_t time points with frequencies in the range $f_{detected} \leq f \leq f_{fixed}$. These are consistent with n independent draws from a normal distribution, assuming a variance of

$$1. \sigma_i^2 = \frac{1}{n_t} f_i (1 - f_i)$$

where f_i is the mean frequency of both mutations at each time point. The variance σ_{tot} for the pair of mutational time series is calculated as the mean of the variance at all time points:

$$2. \sigma_p^2 = \frac{1}{n_t} \sum_i^{n_t} \sigma_i^2 = \frac{1}{n_t} \sum_i^{n_t} f_i (1 - f_i)$$

Since we are interested in the similarity between both mutational time series, we will compare the differences in values between both time series. Let $d_i = |f_{ai} - f_{bi}|$ for all time points i yielding a mean of

$$3. \bar{d} = \frac{1}{n_t} \sum_i^{n_t} d_i = \frac{1}{n_t} \sum_i^{n_t} |f_{ai} - f_{bi}|$$

Finally, given σ_p and \bar{d} we can calculate the probability that this pair of mutations belong to the same genotype using the cumulative probability distribution of the normal distribution:

$$4. \quad p_{pair} = 1 - \int_{-\infty}^{\bar{d}/\sigma_p} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \equiv 1 - erf\left[\frac{\bar{d}}{\sqrt{2\sigma_p^2}}\right]$$

Notes

For each pair we discarded a time point if both f_{ai} and f_{bi} failed the condition $f_{detected} < f < f_{fixed}$, where $f_{detected} = 0.03$ and $f_{fixed} = 0.97$. Note the these filters check "greater than" and "less than", and NOT "equals to". This won't affect the results that much, but should be clarified since it affects repeatability.