Chapter 1

Nesting successive genotypes

1.1 Relevant Parameters

These parameters control how the lineage of genotypes is inferred:

–**additive 0.05** The additivecutoff value controls when the sum of a pair of genotypes is considered consistently greater than the fixed cutoff value. The default value is controlled by the `detection` parameter.

–**subtractive 0.05** The subtractivecutoff value controls when a potential ancestor is considered greater than an unnested genotype and *vis versa*.

–**covariance 0.01** Defines how each of the three possible covariance scenarios are defined (see below).

1.2 Evidence of lineage

After grouping mutational trajectories into genotypes, the scripts infer the lineage of each genotype based on the frequency measurements. Genotypes are nested according to a few basic rules:

1.2.1 Summation

Checks whether the unnested genotype and potential ancestor consistently sum to greater than the fixed cutoff limit (typically 0.97 to 1). This is weak evidence that one of the genotypes is a background and the other arises in the background of the first.

1.2.2 Genotype Size

This checks whether one of the genotypes is consistently larger than the other. A genotype is considered consistently greater than another if the difference between the two frequencies exceeds a value of 0.15 (based on original scripts) at least once or exceeds 0.03 (based on the uncertainty option) at least twice. **Note: This is currently combined with the subtractive check when scoring each genotype pair.**

1.2.3 Covariance

The covariance between two series of random variables provides a measure of the correlation between both series. A genotype which arises in the background of another genotype must be correlated with the parent genotype. A negative covariance between two genotypes indicates that they are competing genotypes with one outcompeting the other. The covariance of two series X and Y is defined as

$$Cov(X,Y) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{1.1}$$

1.2.4 Expected vs. Observed Jaccard Distance

The jaccard distance is a measurement of the dissimilarity of two sample sets. Since each frequency measurement describes the abundance of a mutation at a sampled timepoint, we can use the jaccard distance to compare the abundance of two genotypes over the course of an experiment. Since each genotype represents a set of abundance measurements, the jaccard distance between two genotypes X and Y can be calculated as follows:

$$J(X,Y) = 1 - \frac{|X \cup Y|}{|X| + |Y| + |X \cap Y|} \equiv \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} \tag{1.2}$$

Since each genotype is a set of frequency measurements, the cardinality of each set can be computed as

$$|X| = \sum_{i=0}^{n} X_i \tag{1.3}$$

Based on this, the union and intersection can be defined as

$$|X \cap Y| = \sum_{i=0}^{n} min(X_i, Y_i)$$
$$|X \cup Y| = |X| + |Y| - |X \cap Y|$$

When Y arises in X, all elements of Y are also contained in X, $|X \cup Y| = |X|$ and $|X \cap Y| = |Y|$, reducing the above equation to

$$J_{expected}(X,Y) = \frac{|X| - |Y|}{|X|} \tag{1.4}$$

If the computed jaccard distance is equal to the expected jaccard distance, genotype Y arises in genotype X.

1.2.5 Example

For the genotypes listed below, how should the genotypes be nested?

Genotype	0	17	25	44	66	75	90
genotype-6	0	0	0	0.273	0.781	1.000	1.000
genotype-7	0	0	0	0.403	0.489	0.057	0.080
genotype-3	0	0	0	0	0.211	0.811	0.813

The jaccard distance between genotype-6 and genotype-7, assuming genotype-6 is the background is calculated as follows:

$$
\begin{aligned}
|X| &= 0.273 + 0.781 + 1.000 + 1.000 = 3.054 \\
|Y| &= 0.403 + 0.489 + 0.057 + 0.080 = 1.029 \\
|X \cap Y| &= 0.273 + 0.489 + 0.057 + 0.080 = 0.899 \\
|X \cup Y| &\equiv |X| + |Y| - |X \cap Y| = 3.054 + 1.029 - 0.899 = 3.184
\end{aligned}
$$

Now let's calculate the jaccard distance:

$$J(X, Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} = \frac{3.184 - 0.899}{3.184} = 0.718$$

Test against the expected distance, computed as:

$$J_{expected}(X, Y) = \frac{|X| - |Y|}{|X|} = \frac{3.054 - 1.029}{3.054} = 0.663$$

Since $0.718 \neq 0.663$, We cannot say that Y arises in the background of X.

1.2.6 Example

Let's test if genotype-3 arises in the background of genotype-6:

$$
\begin{aligned}
|X| &= 0.273 + 0.781 + 1.000 + 1.000 = 3.054 \\
|Y| &= 0.211 + 0.811 + 0.813 = 1.835 \\
|X \cap Y| &= 0.000 + 0.211 + 0.811 + 0.813 = 1.835 \\
|X \cup Y| &\equiv |X| + |Y| - |X \cap Y| = 3.054 + 1.835 - 1.835 = 3.184
\end{aligned}
$$

The jaccard distance is then

$$J(X, Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|} = \frac{3.184 - 1.835}{3.184} = 0.425$$

Test against the ideal distance:

$$J_{expected}(X, Y) = \frac{|X| - |Y|}{|X|} = \frac{3.054 - 1.835}{3.054} = 0.425$$

Since $J = J_{expected}$ there is evidence that genotype-3 arises in the background of genotype-6.

1.3 Scoring

Each of the above checks is used to calculate a score between each unnested genotype and a potential ancestor. The parent genotype is then chosen as the potential ancestor with the greatest score when compared against the given genotype.

additive score [**0,1**] Tests whether the combined unnested genotype and potential ancestor consistenty sum to greater than 1. The only scenario where this should occur is if the unnested genotype arises in the background of potential ancestor.

- **1** The average sum of the unnested genotype and the potential ancestor is greater than 1.
- **0** The average sum of the unnested genotype and the potential ancestor is equal or below 1.

subtractive score [**-1,1**] The subtractive score adds weak evidence for/against the hypothesis that the potential ancestor is a true ancestor of the unnested genotype.

- **1** The unnested genotype is always detected at a lower frequency than the potential ancestor.
- **0** The unnested genotype is occasionally detected at a greater frequency than the potential ancestor.
- **-1** The unnested genotype is always detected at a greater frequency than the potential ancestor.

jaccard score [**-2,2**] The unnested genotype is compared against both the potential ancestor and a hypothetical genotype consisting of the remaining population not represented by the potential ancestor. This hypothetical genotype is assigned a frequency of 0 for all timepoints where potential ancestor is fixed and a frequency of $1 - f_i$ for all timepoints that the potential ancestor is not fixed.

- **2** The unnested genotype is only a subset of the potential ancestor.
- **1** The unnested genotype is a subset of both the potential ancestor and the hypothetical genotype, but overlaps with the potential ancestor twice as much as the hypothetical genotype.
- **0** The unnested genotype is a subset of both the potential ancestor and the hypothetical genotype, but shares a similar or smaller overlap with the potential ancestor as it does with the hypothetical genotype.
- **-2** The unnested genotype is a subset of only the hypothetical genotype, and is not a subset of the potential ancestor

covariance score [**-2,2**] Tests whether the two genotypes are competing against each other or mutually evolving together. The covariance $Cov(X, Y)$ between the two genotypes is compared against a cutoff value d (see the program options for more information) to test which scenario this pair of genotypes falls under. **Note: This is only checked if the sum of the previous scores is positive.**

- **2** $Cov(X, Y) > d$
- **0** $-d \leq Cov(X, Y) \leq d$
- **-2** $Cov(X, Y) < -d$

1.3.1 Example