

# Simulating Concerted Evolution

---

Thomas Mailund

June 27, 2005

We describe a method for simulating duplicated genes under concerted evolution. The method takes into account both inter and intra gene variance and models both inter gene and intra gene conversion.

## INTRODUCTION

---

Teshima and Innan, in [1], describes simulation results for duplicated genes under concerted evolution. Their simulation method evolves a pair of initially identical genes; mutations can accumulate in the two genes—under an infinite sites mutation model—causing divergence between the genes, and gene conversions can occur between the two genes, reducing the divergence again. Their method, however, does not take into account the diversity on the individual gene level found in the population where the genes evolve, and how gene conversion between homologue genes (intra gene conversion) interfere with the gene conversions between the two paralogue genes (inter gene conversion).

In the following we extend the method from [1] to include a population in which the two genes evolve thus extending the model to include both variation between homologue and paralogue genes.

## PARAMETERS

---

The following parameters influence the process:

*Effective population size:* Each generation consists of  $2N$  haplotype individuals. In the first phase (see below), all haplotypes contain one gene; in the second phase, only descendants of the duplication contain two genes, while the remaining only have a single copy of the gene; and in the third phase all haplotypes contain two genes.

*Mutation rate:* The per gene per generation mutation rate,  $\mu$ . Each gene has, per generation, probability  $\mu$  of acquiring a mutation, thus individuals with both genes has probability  $2\mu$  of acquiring a mutation. The mutation model is the infinite sites model, thus all mutations are assumed to be unique and are randomly placed on the gene.

*Terminator mutation rate:* The per gene per generation rate,  $m$ , of ‘terminator’ mutations—mutations so severe that regions not sharing such mutations cannot join in gene conversions.

*Pairing probability:* The probability of two genes pairing in meiosis,  $p(g, g')$ —in [1] the notation  $S_1(-)$  is used. This is the probability of two specific genes pairing in meiosis, and is dependent on the divergence between the two genes. If the genes do not pair, no gene conversion is possible.

A possible form of  $p$ , taken from [1], is:

$$p(g, g') = \begin{cases} 1 & \text{when } \pi(g, g') < t \\ 0 & \text{when } \pi(g, g') \geq t \end{cases}$$

where  $\pi(g, g')$  is the pair-wise distance between  $g$  and  $g'$  (number of mutations for which they differ), and  $t$  is some threshold value.

*Gene conversion rate:* Each paired genes per generation can initiate a gene conversion with probability  $g$ . The direction of the conversion—which gene is the donor and which is the receiver—is assumed to be random.

*Expected tract length:* The expected length of the conversion,  $1/Q$ . If a pair of genes initiate a gene conversion, the location of the conversion is selected at random, with a length drawn from the exponential distribution with intensity  $Q$ :  $\text{length} \sim \text{Exp}[Q]$ .

*Conversion success probability:* If a gene conversion is initiated over regions  $r$  and  $r'$ , in genes  $g$  and  $g'$ , respectively, it is rejected if one of the regions contain a terminal mutation not found in the other. Otherwise, the conversion is accepted with probability  $\alpha(r, r')$ —in [1] the notation  $S_2(-)$  is used—otherwise the conversion is rejected. The acceptance probability depends on the divergence of the regions; notice that while the pairing probability,  $p(g, g')$  depends on the entire genes, the acceptance probability,  $\alpha(r, r')$  only depends on the regions involved in the conversion.

Possible forms of  $\alpha$ , taken from [1], are:

$$\alpha(r, r') = \begin{cases} 1 & \text{when } \pi(r, r') < t \\ 0 & \text{when } \pi(r, r') \geq t \end{cases}$$

where  $\pi(r, r')$  is, again, the pair-wise distance and  $t$  is some threshold value, or

$$\alpha(r, r') = \begin{cases} 1 - \frac{\pi(r, r')}{t} & \text{when } \pi(r, r') < t \\ 0 & \text{when } \pi(r, r') \geq t \end{cases}$$

where the acceptance probability decreases linearly as the pairwise distance grows, up to some maximal threshold value  $t$ .

## SIMULATION SETUP

---

The simulation method is a simple discrete, fixed size, generations method: each generation consists of  $2N$  haplotypes that are sampled at random to form the next generation.

In this model we divide the simulation into three phases: The first phase is a burn-in phase that is used to introduce diversity into the population before the duplication event; the second phase is initiated by the duplication event and terminates when the duplication is either lost or fixed in the population; the third phase, only run if the duplication is fixed, simulates the concerted evolution.

### General Simulation Steps

Each iteration of the simulation builds a new generation from the previous generation in the following way:

1.  $2N$  haplotypes are selected, randomly with replacement, from the  $2N$  haplotypes of the previous generation.
2. Mutations are applied:
  1. Each selected haplotype mutates with probability  $\mu$  per gene. Mutations are placed uniformly random on the gene.
  2. Each selected haplotype obtains a terminator mutation with probability  $m$  per gene. Mutations are placed uniformly random on the gene.
3. The haplotypes are now paired into  $N$  diplotypes and the genes tried paired: if both haplotypes contain only a single gene, those are tried paired; if both contain two genes, the first is tried paired with the first and the second with the second; and if one contain one gene and the other two, the one gene is tried paired with a random of the two. With probability  $s(g, g')$  the pairing is successful, and a gene conversion is feasible. Otherwise, no gene conversion is possible between the pair.
4. If a gene conversion is feasible:
  1. With probability  $g$  per pair, a gene conversion is initiated. The starting point and direction (left or right) is chosen at random, and the length is selected from  $\text{Exp}[Q]$ .
  2. Let  $r$  be the region selected in this way from gene  $g$ , and let  $r'$  be the corresponding region in  $g'$ . If either  $r$  or  $r'$  contain a terminator mutation not found in the other, the conversion is rejected. Otherwise, the gene conversion is accepted with probability  $a(r, r')$  and rejected otherwise.

## Simulation Phases

The first phase is a burn in period used to introduce variation into the population before the real simulation. It starts with a population of  $2N$  identical haplotypes, each with a single gene and runs the iterations described above until an equilibrium is reached. Different criteria can be used to decide when an equilibrium is reached, e.g. when all haplotypes in the current generation are descendants of the same original haplotype (and thus the lack of variation in the original generation is invisible, ignoring for sake of argument the pieces of the haplotypes that have other ancestors due to gene conversions).

The second phase is started with a gene duplication in one of the haplotypes. Descendants of the duplication will have two genes, while all other haplotypes will have a single gene. The second phase is run until the duplication haplotype has become fixed or has been lost.

The third phase is run if the duplication is fixed in phase two. It runs from the fixation until some stop criteria is reached. In [1], two criteria were used: the simulation is run for either  $5 \cdot 10^9$  generations, or until the distance between the two genes reach 20%. In [1], the length of the simulated genes are 1kb and the stop criteria is  $\pi(g, g') \geq 200$ . In this model, since we have a population of each gene, we cannot use this measure of 20% divergence; if we let  $g_1^i, g_2^i$  define the first and second gene, respectively, of haplotype  $i$ , we can define the intergenic distance, or paralogue distance  $\pi_p$  as the average distance between the first and second gene:

$$\pi_p = \frac{1}{4N^2} \sum_{i=1}^{2N} \sum_{j=1}^{2N} \pi(g_1^i, g_2^j)$$

and use, as stop criteria,  $\pi_p \geq 200$ .

## SAMPLING QUANTITIES

---

In [1], the quantity of interest, sampled during the simulation, was the distance between the two genes. The corresponding quantity in our setup is the paralogue distance,  $\pi_p$ , define above. Another quantity of interest, not present in [1] but in our method, is the intragenic distance, or homologue distance,  $\pi_h$ , defined as the average distance on the same gene (first or second):

$$\pi_h = \frac{1}{N(2N-1)} \left( \sum_{i=1}^{2N} \sum_{j=i+1}^{2N} \pi(g_1^i, g_1^j) + \sum_{i=1}^{2N} \sum_{j=i+1}^{2N} \pi(g_2^i, g_2^j) \right)$$

It is also interesting to see how the concerted evolution affects estimators of  $\theta = 4N\mu$ . To examine this, we could sample  $n$  genes—for this purpose it probably only makes sense to sample from either the first *or* the second, but not mix the samples—and calculate estimators from this sample of  $n$  haplotypes. Possible estimators are:

*Tajima's  $\theta$  estimator  $\hat{\pi}$* : The average pairwise distance:

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i,j} \pi(g_i, g_j)$$

*Watterson's estimator  $\hat{\theta}_W$* :

$$\hat{\theta}_W = \frac{S_n}{a_n}$$

where  $S_n$  is the number of segregating sites in the sample, and

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}.$$

*Tajima's D*: The normalised difference of the two estimators above—expected to be near zero under a normal coalescence model and thus interesting in this extended model.

$$D = \frac{\hat{\pi} - \hat{\theta}_W}{\sqrt{e_1 S_n + e_2 S_n (S_n - 1)}}$$

where

$$e_1 = \frac{n+1}{3a_n(n-1)} - \frac{1}{a_n^2}$$

$$e_2 = \frac{1}{a_n^2 + b_n} \left( \frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{na_n} + \frac{b_n}{a_n^2} \right)$$

and

$$b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

The effect of sampling from the population of one of the genes, where this gene occasionally obtains polymorphism from the other gene, we expect to result in an increased effective population size, and thus an increase in estimates for  $\theta$ . If, on the other hand, we sample full haplotypes and build 'super' genes by concatenating the pairs of genes, and estimate  $\theta$  from this (dividing by 2 to compensate for the doubling in gene length and thus  $\mu$ ), we would expect an under estimate due to the lessened diversity due to concerted evolution.

## REFERENCES

---

1. K.M. Teshima and H. Innan, *The effect of gene conversion on the divergence between duplicated genes*, *Genetics* **166** (2004), 1553–1560.