

Introducing mutations to Wright-Fisher

Thomas Mailund

September 2, 2007

We can introduce mutations to the Wright-Fisher model. For simplicity, we assume that both allele A and allele a are equally likely to mutate into the other, and that the only possible mutations are from A to a and a to A . We further assume that all mutations occur in the gamete pool: alleles are first put in the gamete pool in proportions given by the population frequencies; the gametes then potentially undergo mutation, changing the frequencies in the gamete pool; and finally the next generation is drawn from the gamete pool with the new frequencies.

Let μ be the probability of a gamete mutating. If p_A is the frequency of A s in the gamete pool before mutations, then the frequency of A s in the gamete pool after mutations is $p'_A := (1 - \mu)p_A + \mu(1 - p_A)$. (Why?) So drawing next generation from the gamete pool means that the number of A s in the next generation is distributed as $b(2N, p'_A) = b(2N, (1 - \mu)p_A + \mu(1 - p_A))$.

Exercises

Simulations with mutations: Update your Wright-Fisher simulator to include mutations. Simulate and plot some runs. What do you see? What happens when you change μ ? Simulate 100 traces and look at the histograms of distributions of various generations.

Sampling number of A s: If we run the WF model long enough, select a random generation, and then count the number of A s, that would be like sampling the distribution of A s, we would get a feeling for how the number of A s are distributed in a general population (as a function of population size, N and mutation rate, μ). We would, of course, need to sample a lot of these before we get a clear idea about the distribution, but it is a start. If we sample several A counts from the same process, they will be correlated, so we are not in the preferred world of independent samples, but it might be good enough (later in the course we will see when we can expect “good enough” – there are some theoretical results to go by here).

Try sampling, say, every 100 generation for 100,000 generations. Compare the distribution you get here with the histograms from before.

Convergence of the distribution: Use the matrix representation of the WF model to calculate the distribution after, say, 50 or 100 steps. Compare it to the histograms you looked at earlier. What do you observe? How important is the initial distribution to the distribution after 100 steps? How does that differ when $\mu = 0$ compared to $\mu > 0$?

If π_{100} is the distribution after, say, 100 steps, and $\pi_{101} = \pi_{100}P$ (where P is the transition matrix), then how are π_{100} and π_{101} related? If you consider the (vector) distance between π_i and π_{i+1} for increasing i , what do you observe? Try calculating it and plot how the distance changes.

You can use something like this to calculate the vector distance:

```
sqrt(sum((p.100 - p.101)^2))
```

We call a distribution π that satisfy $\pi = \pi P$ for a *stationary distribution* of P . Do you see why?