The frequency of a segregating duplicate gene

Martin O'Hely

ARC Centre of Excellence for Mathematics and Statistics of Complex Systems
University of Queensland

AUSTRALIAN RESEARCH COUNCIL
Centre of Excellence for Mathematics and Statistics of Complex Systems

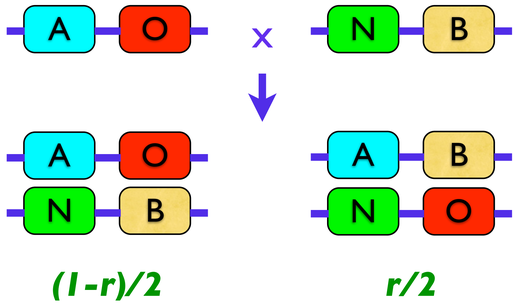
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Duplicate genes

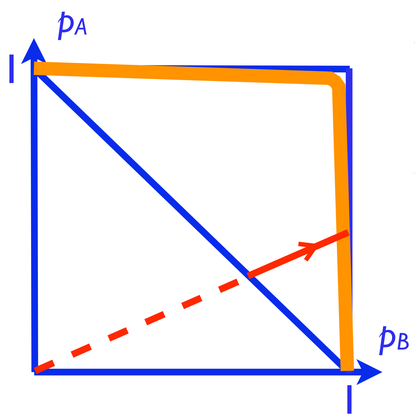
Sometimes, the process of reproduction results in genes which are not only mis-copied (mutations) but which are copied into the wrong place in the genome. This produces a *duplicate* gene. Genome projects have revealed multitudes of duplicated genes across a wide variety of species. The existence of duplicate genes is thought to be critical in the process of genetic innovation: duplicates are free to explore new variations, perhaps finding advantageous new forms, without depriving the organism of any essential functions performed by the original gene. Knowing the population genetics of a duplicate, i.e. the way in which a duplicate can spread through a population, is an essential step towards a quantitative understanding of how this evolution can proceed.

Loosely-linked models

Take a haploid model which allows for recombination: that is, with probability *r* an offspring takes genes at original and duplicate loci from different parents.



A and B represent functional alleles at the original and duplicate loci; N and O are mutated, non-functional versions of the gene (these mutations are assumed to occur with probability *u* at each reproduction event; an NO individual is not viable). Eventually, all individuals will be AO (*duplicate loss*) or all individuals will be NB (*map change*). Although the state of the process is described by p_A and p_B, simulations show that the process remains largely in a mutation-selection balance described by a hyperbola in (p_A,p_B)-space. Diffusion modelling implies the hyperbola is approached along lines through the origin. As a result the process will be analyzed via a one-dimensional diffusion. The diffusion variable will be taken as $g = p_B/(p_A+p_B)$.



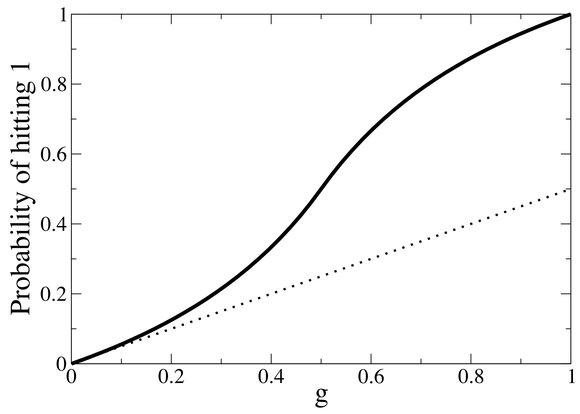
WATTERSON (1983) used a similar approach but with a "fast" diffusion around and a "slow" diffusion along the hyperbola.

The one-dimensional diffusion

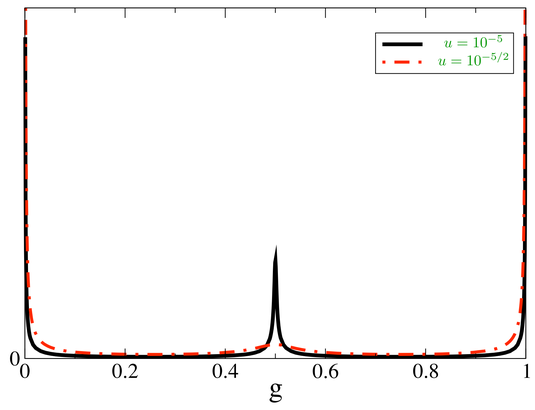
The one-dimensional diffusion along the hyperbola has infinitesimal mean and variance

$$\frac{2-\epsilon}{2}(2g-1)g(1-g)$$
$$\frac{g(1-g)}{2}\left(1-2(2-\epsilon)g(1-g)+\sqrt{1-2(2-\epsilon)g(1-g)}\right)$$

Here ε is a composite parameter roughly equal to the mutation rate. More meaningfully, this diffusion has scale function and speed densities as illustrated below:



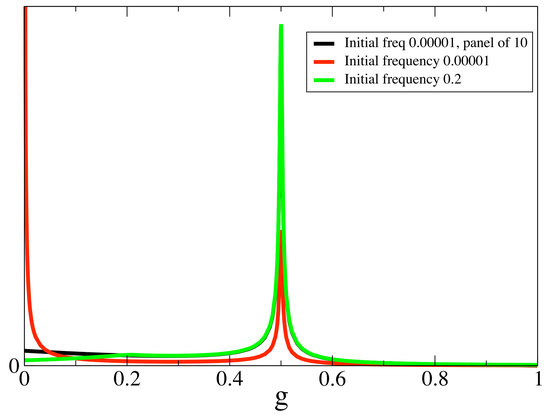
Scale function: for small initial duplicate frequencies g_0, the probability of map change is $g_0/2$.



Speed density: there is the usual slow-down near absorbing points, but also near equi-frequency, indicating that a duplication process not on the verge of map change or loss of the duplicate is likely to be very close to equi-frequency.

Observed frequencies

As the speed density does not correspond to a probability distribution, it doesn't provide direct information about what frequency an active or *segregating* duplication is likely to be observed at. For this kind of information there is the pseudo-transient distribution (EWENS 1963).



This is a probability distribution for the time spent at a certain frequency before loss or map change. Note that:

- there is a local peak at the initial frequency
- for small initial frequencies the spike in the middle gets smaller: most of the duplicates seen are newly-arisen and being quickly lost

However, a duplicate that is quickly lost is unlikely to be identified in a population without considerable sampling effort. This is *ascertainment*: one might find duplicates by examining a small number of individuals (called a *panel*), but then determine their frequencies by examining a larger sample or even the whole population. Doing this, we see that the spike at g=0.5 increases again — although there is still a slightly greater tendency to pick up low frequency duplicates.

The intuitive reason for the spike is that evolution proceeds slowly near $p_A = p_B = 1$ since there is little genetic diversity: most change comes from mutation which is assumed to be rare.

References

EWENS, W.J. (1963). The diffusion equation and a pseudo-distribution in genetics. *Journal of the Royal Statistical Society. Series B (Methodological)* **25** 405-412.

WATTERSON, G. A. (1983). On the time for gene silencing at duplicate loci. *Genetics*. **105** 745-766.

ARC AUSTRALIAN RESEARCH COUNCIL