The handle http://hdl.handle.net/1887/18976 holds various files of this Leiden University dissertation.

Author: Ghosh, Atiyo
Title: Calculating hazard rates of introgression with branching processes
Date: 2012-05-22
CHAPTER 2: COMBINING MODELS WITH EXPERIMENTS

1. Introduction

The models presented within this thesis have been developed as a part of a wider project which uses the carrot (*Daucus carota*) as a case study for the development of a methodology to quantify introgression risk realistically. The carrot is primarily an outcrossing species, and there have been many documented occurrences of hybridisation between wild and cultivated carrots (Wijnheijmer et al. 1989 Hauser and Bjorn 2001, for example), so it is a good candidate for such a study. The development of models has been concurrent with the execution of field trials by collaborators working on the same project. At the time of writing, field trials are ongoing, but have reached a stage where the models in the thesis can be combined with preliminary results from experiments.

The data for the calculations come from a field trial conducted during the winter of 2011 in Lisse, The Netherlands. A commercially available variety known as Flakkese was used as a cultivar, which was crossed and backcrossed with wild carrots collected from Stevenshof, a suburb of Leiden, The Netherlands. For full details of crossing experiments, and field trials, see Grebenstein (in prep.). A summary of this data for wild plants, F1 hybrids, and BC1 backcrossed individuals can be found in Table 1. The data set at the time of writing is still incomplete, with data for umbel sizes available as opposed to full seed sets, small sample sizes, and some missing data. I use them merely as an illustration of how the hazard rate is calculated from real data.

<table>
<thead>
<tr>
<th>Plant type</th>
<th>Average primary umbel diameter (cm)</th>
<th>Average primary umbel area ($\text{cm}^2$)</th>
<th>Survival probability to flowering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stevenshof</td>
<td>5.73</td>
<td>27.52</td>
<td>0.94</td>
</tr>
<tr>
<td>F1</td>
<td>6.72</td>
<td>40.70</td>
<td>1</td>
</tr>
<tr>
<td>BC1</td>
<td>6.25</td>
<td>33.61</td>
<td>0.86</td>
</tr>
</tbody>
</table>

The average umbel area was calculated from the separately measured umbel diameters, under the assumption that each umbel is circular.

2. Calculation of the hazard rate

All plants in the field trial either flowered or died after one year, i.e. no flowers remained in a vegetative state. Thus, we use the model in Chapter 1 with the survival probabilities of non-flowering one year old plants set to zero (i.e. $p_1 = p_3 = p_5 = 0$). Please refer to Chapter 1 for full derivations and details of assumptions.

To estimate the seed set from the data on umbel diameter, first umbel area was calculated, assuming that each umbel was circular. From the umbel area, a
measure for the seed set was calculated, assuming that each plant had the same density of seeds per unit of umbel area, denoted by the constant, $k$.

2.1. Seed establishment probability. Adapting previous results to annuals leads to the following seed establishment probability of a single seed:

$$\hat{p}_0 = \frac{1}{m_1 r_1},$$  

(1)

where $r_1$ is the flowering probability of a wild plant, and $m_1$ is the expected number of seeds that a flowering wild plant produces. Looking at the values in Table 1 and substituting into Eq. (1), we find $\hat{p}_0 = \frac{1}{27.52 \times 0.94 \times k} = \frac{1}{25.87k}$.

2.2. Extinction probability of a lineage initiated by a backcrossed individual. The extinction probability of the lineage initiated by a single backcrossed becomes the following when made applicable for annuals:

$$q = r_5 G_5 (\hat{p}_0 q + (1 - \hat{p}_0)) + 1 - r_5.$$  

(2)

where $r_5$ is the flowering probability of a single backcrossed individual. $G_5(s)$ is the probability generating function (p.g.f.) of the seed numbers produced by a single BC1 plant. Using a Poisson p.g.f. (i.e. $G_5(s) = e^{-m_5 (1-s)}$, where $m_5$ is the expected number of seeds produced by a flowering BC1 plant) yields the following expression:

$$q = r_5 e^{-\hat{p}_0 m_5 (1-q)} + 1 - r_5.$$  

(3)

From Table 1, we have $m_5 = 33.61k$. Putting this value into the above equation, and also substituting the calculated value of $\hat{p}_0$, the following expression is reached.:  

$$q = 0.86e^{-\frac{33.61}{25.87} (1-q)} + 0.14.$$  

(4)

Note that this expression is independent of $k$. Also, the solution of $q = 1$ satisfies this equation, but the extinction probability is the smallest root of this equation, which can be calculated numerically to give $q = 0.83$ (to two decimal places).

2.3. The asymptotic hazard rate. In this case the asymptotic hazard rate equals

$$\hat{H}(q) = 1 - G_0 (\hat{p}_0 (r_3 G_3 (\hat{p}_0 q + 1 - \hat{p}_0) + 1 - r_3) + 1 - \hat{p}_0)$$  

(5)

where $G_0(s)$ is the p.g.f. of the number of hybrid seeds produced in the wild population per generation, $G_3(s)$ is the p.g.f. of the number of seeds produced by a flowering F1 hybrid, and $r_3$ is the flowering probability of an F1 hybrid. It is easiest to break up the calculation of Eq. (5) into two parts. First, define and calculate a part of the argument of $G_0(s)$ as follows:

$$c = r_3 G_3 (\hat{p}_0 q + 1 - \hat{p}_0) + 1 - r_3$$  

$$= r_3 e^{-\hat{p}_0 m_3 (1-q)} + 1 - r_3$$  

(6)

where I have assumed that the seed production of a flowering F1 plant is Poisson distributed in writing down the second line. From Table 1, we have that $r_3 = 1$ and $m_3 = 40.70k$. This gives a value of $c = 0.76$ (to two decimal places). Note that $c$ is independent of $k$. 

Combining Models with Experiments 31

Assuming that the number of hybrid seeds produced in the wild population is Poisson-distributed with mean \( m_0 \), we find the following expression for the asymptotic hazard rate:

\[
\hat{H}(q) = 1 - e^{-\hat{p}_0 m_0 (1-c)} \\
= 1 - e^{-m_{hyb}(1-c)}
\]  \( (7) \)

The term \( \hat{p}_0 m_0 \) in the exponent is the product of the seed-establishment probability and the expected number of hybrid seeds formed per generation. Thus, \( \hat{p}_0 m_0 \) can be interpreted as the expected number of hybrid plants produced per generation, denoted by \( m_{hyb} \). This can vary, depending on several factors, e.g. distance the wild population to the crop field, and size of the wild population.

3. Results

Figure 1 shows the asymptotic hazard rate plotted against hybridisation rate. At small hybridisation rates, as in Fig. 1 (b), the asymptotic hazard rate increases nearly linearly, and we can use the first order Taylor approximation:

\[ H \approx m_{hyb} (1 - c). \]  \( (8) \)

At larger hybridisation rates, the hazard rate approaches one. This can be seen straight from Eq. (7), where the exponential term becomes zero as \( m_{hyb} \) becomes large.

Figure 1. The asymptotic hazard rate plotted against \( m_{hyb} \) at two different scales. Parameters are as shown in Table 1.

Figure 2 shows the effect that changing the average F1 and BC1 umbel areas has on the asymptotic hazard rate. The asymptotic hazard rate is more sensitive to changes in BC1 fitness than F1 fitness.
Figure 2. The asymptotic hazard rate plotted againsts average F1 (subplot (a)) and BC1 (subplot (b)) primary umbel areas. A value of $m_{hyb} = 0.1$ is used for both plots. Other relevant parameters are as shown in Table 1.

4. DISCUSSION

In this chapter, I have applied the methodology from Chapter 1 to estimate the hazard rate of introgression of crop into genes using measured data from a field trial of *Daucus carota*. Data on hybridisation rates is still required to accurately quantify introgression risk, and this data is currently being gathered. On the basis of the calculations presented here, we can predict that an average of 0.1 hybrid plants produced per generation in the wild population will lead to a hazard rate of 0.024. This value implies that the expected time until introgression (initiation of a permanent lineage) occurs is 40 years.

The sensitivity analysis shows that the asymptotic hazard rate is more sensitive to BC1 fitness than F1 fitness. The reason for this can be explained by an assumption made in the Chapter 1, namely that BC2 and subsequent backcross generations have the same life-history parameters as BC1 plants. Consequently, the life-history parameters of BC1 plants affects all subsequent generations and the model is especially sensitive to changes in BC1 fitness. If data about BC2 and further backcross generations is gathered empirically, then this can be incorporated into the model, and procedures for doing so are given in Chapter 3.

To calculate the hazard rate from the currently available data, several assumptions had to be made, in addition to the assumptions already enumerated in Chapter 1. The data is only currently available in the form primary umbel diameters, whereas data is required in the form of seed sets. Consequently, a direct proportionality was assumed between umbel area and seed set. Furthermore, I assumed the same constant of proportionality to hold for wild, F1 and BC1 plants. This might not be the case, e.g. area in larger umbels might be due to empty space and not seeds. I also assumed that the numbers of seeds produced by plants are
Poisson-distributed. This is a convenient choice, since it results in the final result of the hazard rate to be independent of the constant of proportionality, $k$. An alternative would be to use p.g.f.s as implied by umbel areas, but this would allow an arbitrary choice of $k$ in the final result, so would be unsatisfactory in that regard.

In a full implementation of the model, $\hat{p}_0$ would be calculated from the life-history parameters of wild plants, and $m_0$ would be measured. A hurdle in the implementation in this chapter is that there currently no measurement for $m_0$, and $\hat{p}_0$ is dependent on $k$. It was possible to combine these two parameters and interpret the combination as the expected number of hybrid plants produced per generation. Results could be seen in terms of this hybridisation rate. In a full implementation of the model, using seed sets instead of umbel area would result in no factor of $k$ appearing in $\hat{p}_0$, and data would be available for $m_0$, so these assumptions could be avoided. With richer and more complete data sets, more accurate estimates for introgression risk can be made. Work towards this is currently underway.

References

Grebenstein, C., in prep.