Simera 2 Algorithm

1 Input

The algorithm requires the following data and variables.

- The number of rounds of PCR to simulate.
- The number of potential chimeras to generate.
- A list of sequences with their associated abundances. These represent the true DNA sequences that are supposed to be amplified.
- The sequences representing the chosen PCR primers.
- A value for the parameter, λ. This is the probability of extension failure at any nucleotide position on any sequence undergoing amplification. It is assumed that λ is constant and, because it is a probability, $0 < \lambda < 1$.

2 Chimera Formation Step

A specified number of chimeras are generated and stored, along with the probability of formation for each chimera, using the following procedure. Chimeras are generated alternately from forward extension using the forward primer and reverse extension using the reverse primer.

- Two sequences, A and B, are selected at random with probability of selection relative to their abundance.
- A break point is selected on sequence A using a random variable drawn from the geometric distribution with parameter λ, as shown. If the break point is longer than the sequence then this step is repeated until a valid break point is found.

$$\text{Break Point} \sim \mathit{Geometric}(\lambda)$$

The geometric distribution is chosen because it models the number of successes before a failure - e.g. a break point - occurs.

- The fragment of sequence A to the left of the break point is kept. For reverse extension, the fragment to the right of the break point is kept instead.
- A number of nucleotides equal to the length of the primer, at the end of the fragment, are compared with all positions on sequence B.
- The position for which there are the fewest differences between the end of the fragment and sequence B is chosen as shown.

```
Fragment      GCTTGTCTCAAAGATTAAGCCATGCATGTCT→
                       |||||||| |||||||| ||||
Sequence B    -CATGCTAAAAAAGATTAGGCCATGCACGTCTGTCATTTATTAGAACAAAACCAATTG-
```

Here there are only two differences between the last 20 base pairs of the fragment and the optimal part of sequence B. Primer length is assumed to be 20 base pairs for this example.

- The chimera is formed using the fragment of sequence A and the remainder of sequence B.

```
Chimera       GCTTGTCTCAAAGATTAAGCCATGCATGTCTGTCATTTATTAGAACAAAACCAATTG-
```

- The probability of a fragment of length l forming is

$$\text{Prob} = \lambda(1-\lambda)^k$$

where $k = l - p$ and p is the length of the primer used to form the fragment. Therefore k is the same value as the number of successful nucleotide extensions prior to failure.

- This can then be used to calculate the probability of the chimera forming:

$$\text{Prob} = \lambda(1-\lambda)^k a_i a_j \mathrm{e}^{-m}$$

where a_i and a_j are the relative abundances of the two sequences, expressed as a fraction of the total abundance, and m is the number of differences, or mismatches, between the active part of the fragment and sequence B.

Note that the e^{-m} term (a value between 0 and 1) acts as the 'weight' associated with the fragment and the sequence. It is assumed that a chimera is less likely to form if this weight is low, i.e. the number of mismatches is high.

- The weight, $W = e^{-m}$, is calculated for all chimeras generated and the mean of these is recorded. The mean primer weight for good sequences is also recorded.

3 PCR Step

This step is repeated for the specified number of rounds of PCR. The initial fragment abundance is set to zero and the sequence and fragment abundances are updated accordingly at the start of each round.

- The maximum abundance of sequences at the end of the round is calculated to be double the total abundance at the start of the round.

$$a_{tot} = a_{tot} \times 2$$

- This final abundance is first reduced by the number of PCR failures which is determined using a binomial random variable.

$$\text{Fail rate} = 1 - \frac{\text{primer abundance} + \text{frag abundance}}{\text{prim abundance} + \text{frag abundance} + \text{seq abundance}}$$

$$\text{Failures} \sim Binomial(\text{Fail rate}, a_{tot})$$

$$a_{tot} = a_{tot} - \text{Failures}.$$

The failures in this step represent the sequences for which PCR does not commence.

- The final abundance is further reduced by the number of fragments which are formed using the parameter, λ for each sequence.

$$\text{Prob} = 1 - [1 - \lambda^{(\text{seq length} - \text{primer length})}]$$

$$\text{Fragments} \sim Binomial(\text{Prob}, a_{seq})$$

where a_{seq} is the abundance of the current sequence.

$$a_{tot} = a_{tot} - \text{Fragments}.$$

The failures in this step represent the sequences for which PCR commences but does not complete.

- The number of sequences and chimeras to be generated this round is found.

$$\beta = \frac{\text{mean primer weight} \times \text{primer abundance}}{\text{mean prim weight} \times \text{prim abund} + \text{mean frag weight} \times \text{frag abund}}$$

$$\text{sequences} \sim Binomial(\beta, a_{tot})$$

$$\text{chimeras} = a_{tot} - \text{sequences}.$$

Primers and fragments are selected for use in amplification based on their relative abundances and weighting, or suitability. Primer usage results in standard amplification and fragment usage results in chimeras.

- The individual abundances of the existing sequences are increased such that the total abundance is now equal to the value found in the previous step. Sequences are amplified randomly with probability proportional to their abundances at the start of the round using a multivariate hypergeometric random variable vector. New sequence abundances may not be more than double their old value.

$$\underline{\mathbf{a}}^* = \underline{\mathbf{a}} \times 2$$

$$\underline{\mathbf{a}}_{new} \sim Hypergeometric(\text{sequences}, \underline{\mathbf{a}}^*)$$

where $\underline{\mathbf{a}}$ is the vector of sequence abundances at the start of the round and $\underline{\mathbf{a}}_{new}$ is the vector of sequence abundances at the end of the round.

The vector of sequences is doubled for use in the generation of the hypergeometric r.v. because this represents the maximum new abundance if amplification is perfect. The resultant vector of abundances is forced to be the correct size and no sequence can have a greater abundance than its potential maximum.

- The calculated number of chimeras are chosen with probability of selection proportional to the probability of formation associated with each individual chimera.

$$\underline{\mathbf{c}} \sim Multinomial(\text{chimeras}, \underline{\mathbf{p}})$$

where $\underline{\mathbf{p}}$ is the vector containing the probability of formation for each potential chimera and $\underline{\mathbf{c}}$ is a vector of abundances of the newly generated chimeras.

Note that, in most cases, the majority of entries in $\underline{\mathbf{c}}$ will be zero.

- New chimeras are added to the list of sequences.