Genome Doubling Test Notes

1 Using a discrete Branching Process to decide whether or not genome doubling occurred given a copy number profile

We consider two types of processes by which a genomic profile can evolve from a normal copy number state, an branching process that acts independently on each arm, and a branching process that acts on the entire genome.

Let X_n be the total number of copies of some chromosomal arm after n iterations of this process with $X_0 = 1$.

Consider the following rule: in each "time period" each copy of each arm is given a probability α of being deleted, β of not changing and γ of being duplicated ($\alpha+\beta+\gamma = 1$). Then we can write down the transition probabilities for this chain by

$$P(X_{n+1} = i|X_n = k) = P(\sum_{j=1}^{k} W_j = i)$$

where W_j is equal to 0 with probability α, 1 with probability β and 2 with probability γ. If we suppose that this process acts independently on each chromosomal arm we have $23 \cdot 2 \cdot 2 = 92$ independent observations.

The probability generating function for a discrete probability distribution $p(X_n = k), X_n, k, n \in \mathbb{Z}$ is given by

$$\phi_{X_n}(s) = \mathbb{E}[s^{X_n}] = \sum_{k=0}^{\infty} p(X_n = k)s^k$$

We have $\phi_{X_0}(s) = s$ since $X_0 = 1$ with probability one. Likewise $\phi_{X_1}(s) = \alpha + \beta \cdot s + \gamma \cdot s^2$ since you have a probability α of that initial one copy being lost, a probability β of nothing occurring and a probability γ of a duplication. It is well known that for such a process that

$$\phi_{X_n}(s) = f(\phi_{X_{n-1}}(s)) = \phi_{X_{n-1}}(f(s))$$

where in our specific case f is $f(s) = \alpha + \beta \cdot s + \gamma \cdot s^2$.

Proof:

$$\begin{aligned}
\phi_{X_n}(s) &= \mathbb{E}[s^{X_n}] \\
&= \sum_{k=0}^{\infty} \mathbb{E}[s^{X_n} | X_{n-1} = k] P(X_{n-1} = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}[s^{W_1 + \cdots + W_k} | X_{n-1} = k] P(X_{n-1} = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}[s^{W_1}] \cdots \mathbb{E}[s^{W_k}] P(X_{n-1} = k) \\
&= \sum_{k=0}^{\infty} f(s)^k P(X_{n-1} = k) \\
&= \phi_{X_{n-1}}(f(s))
\end{aligned} \tag{1}$$

The second process we consider is that of genome doubling. In a genome doubling event the copy number of all chromosomal arms double at the same time. Let $X_{n,GD}$ denote a RV which following n periods of the branching process doubled in value. It is easy to see that the generating function of this process can be recursively reconstructed like so:

$$\phi_{X_{n,GD}}(s) = \phi_{X_n}(s^2)$$

Proof. The substitution of $s \to s^2$ will move the coefficients of s^k to the coefficients of s^{2k} and the probabilities p_k will become p_{2k}.

To model a profile with GD we allow for N discrete periods of the branching process, then a GD event, then a further M events of the branching process. We can use the same recursive procedure to model the branching process both before and after the genome doubling event. For brevity if GD does not occur we say $M = -1$ and hence write down our RV's like so $X_{N,M}$.

For example, the generating function for $N = 1$ and $M = 1$ and $\alpha = \gamma$ is given by $f(f(s)^2)$ which evaluates to

$$\begin{aligned}
&(\alpha^5 + \alpha^2 \cdot \beta + \alpha) + (4 \cdot \alpha^4 \cdot \beta + 2 \cdot \alpha \cdot \beta^2) \cdot s + \\
&(2 \cdot (2 \cdot \alpha^2 \cdot \beta^2 + (2 \cdot \alpha^2 + \beta^2) \cdot \alpha^2) \cdot \alpha + (2 \cdot \alpha^2 + \beta^2) \cdot \beta) \cdot s^2 + \\
&(2 \cdot \alpha \cdot \beta^2 + 4 \cdot (\alpha^3 \cdot \beta + (2 \cdot \alpha^2 + \beta^2) \cdot \alpha \cdot \beta) \cdot \alpha) \cdot s^3 + \\
&(\alpha^2 \cdot \beta + (2 \cdot \alpha^4 + 8 \cdot \alpha^2 \cdot \beta^2 + (2 \cdot \alpha^2 + \beta^2)^2) \cdot \alpha) \cdot s^4 + \\
&(4 \cdot (\alpha^3 \cdot \beta + (2 \cdot \alpha^2 + \beta^2) \cdot \alpha \cdot \beta) \cdot \alpha) \cdot s^5 + \\
&(2 \cdot (2 \cdot \alpha^2 \cdot \beta^2 + (2 \cdot \alpha^2 + \beta^2) \cdot \alpha^2) \cdot \alpha) \cdot s^6 + \\
&(4 \cdot \alpha^4 \cdot \beta) \cdot s^7 + (\alpha^5) \cdot s^8
\end{aligned} \tag{2}$$

These generating functions grow exponentially and are prohibitively large for N and M much greater than about 6. An efficient program written in sage that is reliant dynamic programming to reduce computational time and memory

consumption is available under the name "GD_v1.0.sage". This program needs only be run once.

The provided output file with the results begins with a small R program which describes how the generating functions can be used to maximise the likelihood of the data given N, α, γ in the case of no GD and $N, \gamma, \beta, \alpha, M$ in the case of GD. By the comparison of the likelihoods a decision can be made as to which model suits the data the best using AIC to account for the fact that one of the models has more parameters.

1.1 A small caveat: constructing the joint distribution from the marginals

It may be reasonable to suppose that a genome cannot lose both the maternal and paternal copies of a given arm. This implies that the maternal and paternal material are not independent of each other.

Let the Boolean argument $E \in \{True, False\}$ describe whether or not our meta process can go extinct and let $X_{N,M,E}$ be the RV of interest. Then the generating function of $X_{N,M,True}$ is simply that of $X_{N,M}$ above.

Since GD does not affect the probability of extinction we need only consider the recursive procedure of the branching process.

Claim: We can modify the recursive procedure $\phi_{X_n}(s) = \phi_{X_{n-1}}(f(s))$ used to obtain the GF for the $X_{N,M,True}$ by instead writing down that

$$\phi_{X_n}(s) = \phi_{X_{n-1}}(f(s)) + (s-1)\phi_{X_{n-1}}(0)$$

with $\phi_{X_0}(s) = s$ as before to obtain the GF for $X_{N,M,False}$.

The claim is that this recursive definition of a probability generating function defines a RV which is the branching process introduced above except that the last copy[1] has probability γ of duplicating and probability of $\alpha + \beta$ of staying at copy one and probability 0 of being deleted rather than the respective probabilities of γ, β and α as before.

Proof. The modification of this recursive formula takes the coefficients which describe the probability of X_n going to zero and moves them to the state one.

Let $p_j(\alpha, \gamma, N, M, E)$ be the probability of finding the RV $X_{N,M,E}$ at state j. Let $p_{jk}(\alpha, \gamma, N, M)$ be the probability of finding the maternal copy number at j, the paternal copy number at value k. Then if $j = k = 0$, $p_{jk}(\alpha, \gamma, N, M) = 0$ and otherwise if $j < k$

$$p_{jk}(\alpha, \gamma, N, M) = p_j(\alpha, \gamma, N, M, j \geq 0) \cdot p_k(\alpha, \gamma, N, M, k > 0)$$

Independent loci are identically distributed with this joint distribution.

[1] arbitrarily consider some ordering of events at each time period