Shuffling Chromosomes

December 7, 2012

Original work by Rick Durrett, summary by Diane Holcomb

Reference: Durrett, Rick *Suffling Chromosomes*, Journal of Theoretical Probability, Vol. 16, No. 3, July 2003.

The motivation for the paper is the study of the reorganization of genes within a chromosome. For example a comparative study of chromosome 2 of *D. repleta* and chromosome arm 3R of *D. melanogaster* shows similar genes with some rearrangement.

D.*rep*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
D.*mel*	12	7	4	2	3	21	20	18	1	13	9	16	6	14	26	25	24	15	10	11	8	5	23	22	19	17

The changes to this chromosome region may (or this paper suggests: did) have come about as the result of many inversions of a section of chromosome. In other words remove a section of chromosome, flip it and reattach it, now with the ends reversed.

One model of this process studied in the paper is called the n-**Reversal Chain**. Consider n markers on a chromosome (in any possible order), to these add two further markers 0 and $n+1$. For example when $n = 5$ we may have $0-2-4-3-1-5-6$. We assume that the probability of inversion in any given generation is small and so formulate the inversion process in continuous time. We fix the labels 0 and $n+1$, then at events that occur with the distribution of a rate 1 Poisson process we uniformly choose two edges connecting the sites. We then reverse the sequence of sites that are bracketed between these two edges. For example if we chose edges $2-4$ and $5-6$ after the reversal we would have the arrangement $0-2-5-1-3-4-6$. Note that this shuffling mechanism fixes the added sites 0 and $n+1$ as desired, moreover continually shuffling will result in a uniform distribution on the the permutations of n for the interior sites.

The result I will discuss concerns the amount of time it takes for this system to reach equilibrium, that is, when will the distribution of the interior sites reach the uniform distribution.

Theorem 1. *Consider the state of the system at time $t = cn\log n$ starting with all markers in order. If $c < 1/2$ then the total variation distance to the uniform distribution ν goes to 1 as $n \to \infty$. If $c > 2$ then the distance goes to* 0.

Remark. The n-Reversal Chain is not the only model for chromosome shuffling considered in the paper. Two other models are considered which place different weight on the possible switches (i.e. some switches are more likely then others, for example shorter switches). There are similar results on these models that are proved using an extension of the techniques used in the proof of Theorem 1.

The proof of Theorem 1 is really proving a lower bound for $c < 1/2$ and then proving an upper bound for $c > 2$. The proof of the lower bound will explicitly be given by proving that the amount of time for and n-reversal chain to reach equilibrium is at least $\frac{n+1}{2}\log(n+1)$. For the proof of the upper bound see the original paper.

Proof. We start by giving the following two definitions. We say that an edge is *conserved* is the difference of the 2 endpoints is ±1, and we say and edge is *undisturbed* if it has not been involved with a reversal before time t (that is not selected, it is still possible that the edge has reversed at a consequence of edges on either side of it being selected). The proof is now is done in three pieces.

Lemma 1. *The expected number of conserved edges in equilibrium is 2.*

Proof. The first as last edges are conserved with probability $1/n$ and all other edges are preserved with probability $2/n$ each. □

Lemma 2. *Let U be the number of undisturbed edges and let $t(\epsilon) = (1-\epsilon)\frac{n+1}{2}\log(n+1)$, then $EU = (n+1)^\epsilon$ and $VarU/EU \to 1$ as $n \to \infty$.*

Proof. Let $u_i = 1$ if the ith edge is disturbed and 0 otherwise. Since edge i is disturbed at rate $2/(n+1)$ we have

$$P(u_i = 1) = \exp\left[-\frac{2}{n+1}t(\epsilon)\right] = (n+1)^{-1+\epsilon}$$

Then $EU = (n+1)P(u_i = 1) = (n+1)^\epsilon$.

Now for the result on the variance we not that if $i \neq j$ the rate at which at least one edge is disturbed is $\frac{4}{n+1} - \frac{2}{(n+1)n}$ (rate of choosing 1 of the 2 - rate at which they could be chosen together). This gives us

$$\mathrm{Cov}(u_i, u_j) = P(u_i = 1)P(u_j = 1)\left[\exp\left(\frac{2t(\epsilon)}{n(n+1)}\right) - 1\right]$$

Then summing over i, j we get

$$\mathrm{Var}U = nP(u_i = 1)(1 - P(u_i = 1)) + P(u_i = 1)^2\frac{n(n+1)}{2}\left[\exp\left(\frac{2t(\epsilon)}{n(n+1)}\right) - 1\right]$$

Notice that $EU = (n+1)P(u_i = 1)$ and so we can cancel these terms leaving us to show that

$$P(u_i = 1)\frac{n}{2}\left[\exp\left(\frac{2t(\epsilon)}{n(n+1)}\right) - 1\right] \to 0$$

which is a straight forward computation. □

Lemma 3. *If n is large then the total variation $\|p_{t(\epsilon)} - \nu\|_{TV} \geq 1 - 9/EU$.*

Proof. Let A_ϵ be the set of configurations with at most $EU_{t(\epsilon)}/2$ conserved edges. Recall that the expected number of conserved edges is 2 and so Markov's ine quality gives us $\nu(A_\epsilon) \leq 2/EU$. Now using lemma 3 together with Chebyshev's inequality we get

$$\left(\frac{EU}{2}\right)^2 p_{t(\epsilon)}(A_\epsilon) \leq \mathrm{Var}U$$

which for large n is bounded by $5/4EU$. It follows that

$$\|p_t - \nu\|_{TV} = \sup_A |p_t(A) - \nu(A)| \geq |p_t(A_\epsilon) - \nu(A_\epsilon)| \geq 1 - \frac{9}{EU}$$

□

Noting that $EU = (n+1)^\epsilon$ for $t(\epsilon) = (1-\epsilon)\frac{n+1}{2}\log(n+1)$, we have that $c < 1/2$ exactly when $\epsilon > 0$, therefore the distance will go to 1 as $n \to \infty$.

□