Rick Durrett: Shuffling Chromosomes

MATH285K - Spring 2010 *Presenter: Joshua Hernandez*

1 Introduction

Comparative chromosome mapping of closely related species reveals a history of chromosomal inversions, in which whole segments of chromosome are reversed end to end. We would like to estimate the time until gene order is completely scrambled.

2 Definitions

2.1 The n-Reversal Model

Let the n-*reversal* $\rho_{i,j}$ the permutation on $\{1, \ldots, n\}$ that sends k to $i + j - k$ for $i \leq k \leq j$, and fixes k otherwise. Let $E_n = \{\rho_{i,j}\}_{i \leq j=1}^{n}$.

An n-*chain* is a continuous-time Markov chain η_t on S_n defined as follows:

Definition 1 (n-Reversal Chain) *Set $\sigma_0 = id_{S_n}$ and let $\sigma_1, \sigma_2, \ldots$ be a sequence chosen i.i.d. uniformly at random from E_n. Let N_t be a rate-1 Poisson process. The adapted process*

$$\eta_t = \sigma_{N_t} \circ \sigma_{N_t - 1} \circ \cdots \circ \sigma_2 \circ \sigma_1 \circ \sigma_0.$$

*is called an n-*reversal chain. *If instead of n-reversals we take the set $T_n \subseteq S_n$ of transpositions, we have a process called the n-*transposition chain.

Let p_t be the probability measure associated with the η_t, and let ν be the uniform measure on S_n. Clearly $p_t \to \nu$ as $t \to \infty$. We would like to know how quickly this convergence happens. Our first result is

Theorem 1 (Convergence of the n-Chain) *Consider the state of the system at time $t = cn \ln n$.*
If $c < \frac{1}{2}$ then the total variation distance to the uniform distribution n goes to 1 as $n \to \infty$. If $c > 2$ then the distance goes to 0.

The proof of the lower bound is more or less elementary (but rather tricky). The proof of the upper bound relies on a result from [2], which compares the convergence of p_t to that of $\tilde{p}_t$, the measure associated to the n-transposition chain:

Lemma 1 (Comparison of Rates of Convergence) *Let $\tilde{p}$ and p be symmetric probabilities on a finite group G. Let E be a symmetric set of generators. Suppose that the support of p contains E. Then*

$$\|p_t - \nu\|_{TV} \leq \|\tilde{p}_{t/A} - \nu\|_{TV} \quad \textit{with} \quad A = \max_{z \in E} \frac{1}{p(z)} \sum_{y \in G} |y| N(z, y) \tilde{p}(y).$$

Here, $|y|$ is the length of the shortest sequence in $E = E_n$ whose product is y, and $N(z, y)$ is the number of times $z \in E$ appears in this sequence. The measures p and $\tilde{p}$ are the uniform measures on the set of n-reversals and n-transpositions, respectively. Considering the relative sizes of these sets and the representations of transpositions as products of (at most 2) n-reversals, we get a value of $A = 4\frac{n+1}{n-1}$. Using the known bound on convergence of the n-transposition chain, we get our upper bound.

The proof of Theorem 1 uses Dirichlet forms (simply-defined functionals on p and $\tilde{p}$) to estimate the eigenvalues of the associated semigroups, which in turn can be used to estimate rates of convergence.

2.2 p-Reversal Model

Comparing the rank-correlation (a measure of uniformity) with the number of conserved adjacencies (assuming the n-reversal model, an estimate of the number of reversals) of shuffled *Drosophila* chromosomes, it appears that marker order is converging far too slowly to the uniform distribution. One explanation is that the reversal of short segments is preferred to long segments.

Let the *circular reversal* $r_{i,j}$ be the permutation on $\{1, \ldots, n\}$ sending k to $[i + j - k]_n$ if either $i \leq k \leq \min i + j, n$ or $1 \leq k \leq [i + j]_n \leq i$, and fixes k otherwise. Let F_n be the set $\{r_{i,j}\}_{i,j=1}^{n}$, and let $p(r_{i,j}) = \frac{1}{n} p_j$, where $\sum_j p_j = 1$.

Definition 2 (p-Reversal Chain) *Let $\sigma_0 = id_{S_n}$, and let $\sigma_1, \sigma_2, \ldots,$ be a sequence in F_n, picked i.i.d. with distribution p. Substituting these σ_i in the definition of the n-reversal chain, we get the* p-reversal chain.

One good candidate weighting function is $p_j = \frac{1}{n}\theta^{j-1}(1 - \theta)$ for some $\theta \in (0, 1)$, however the main result of this paper concerns the simpler *L-reversal model*, for which a limit L is chosen on the length of segment reversals, and p_j is simply $1/L$ for $1 \leq j \leq L$ and 0 otherwise.

3 Main Result

Theorem 2 *If $L/n \to 0$, the time required for the L-transposition to reach equilibrium is at most*

$$\frac{8n^3}{(L+1)^2} \ln n \quad \textit{and at least} \quad \frac{3n^3}{8\pi^2 L^2} \ln .$$

This is much slower convergence than in the n-reversal model, here a marker propagates more slowly from one end of a "chromosome" to the other. the The same rank-correlation analysis is applied to this model with better results. The upper bound relies on another result from [2], which handles the propagation of markers along short segments

Lemma 2 *Let $\mathcal{G}$ be a connected graph on $\{1, 2, ..., n\}$ with edge set E. Define a probability on the symmetric group S_n by $p(id) = 1/n$, $p(i,j) = \frac{n-1}{|A|n}$ for $(i,j) \in A$ and $p(\sigma) = 0$ otherwise. For each $x, y \in \mathcal{G}$ let $\gamma_{x,y}$ be a path from x to y in $\mathcal{G}$. Let γ be the length of the longest path, and let*

$$b = \max_{e \in E} |\{(x,y) : e \in \gamma_{x,y}\}|$$

be the maximum number of times an edge appears in this collection of paths. Let

$$k = \left(\frac{8(|E|\gamma b)}{n-1} + n \right) 2(\log n + c)$$

There is a universal constant $\alpha > 0$ so that

$$\|p_k - n\|_{TV} \leq \alpha e^{-c}$$

This gives an upper bound on the convergence of the L-transposition chain, and Theorem 1, in turn gives the bound on convergence of the L-reversal chain.

References

[1] Durrett, R. (2001). Shuffling Chromosomes. *Journal of Theoretical Probability*. **16**, 725-750

[2] Diaconis, P., and Saloff-Coste, L. (1993). Comparison techniques for random walks on finite groups. *Ann. Probab.* **21**, 2131-2156.