J. Appl. Prob. **38,** 324–334 (2001)
Printed in Israel

THE EQUILIBRIUM DISTRIBUTION FOR A GENERALIZED SANKOFF–FERRETTI MODEL ACCURATELY PREDICTS CHROMOSOME SIZE DISTRIBUTIONS IN A WIDE VARIETY OF SPECIES

ARKENDRA DE,* *Purdue University*
MICHAEL FERGUSON,** *Cornell University*
SUZANNE SINDI,*** *California State University, Fullerton*
RICHARD DURRETT,**** *Cornell University*

Abstract

Sankoff and Ferretti (1996) introduced several models of the evolution of chromosome size by reciprocal translocations, where for simplicity they ignored the existence of centromeres. However, when they compared the models to data on six organisms they found that their short chromosomes were too short, and their long chromosomes were too long. Here, we consider a generalization of their proportional model with explicit chromosome centromeres and introduce fitness functions based on recombination probabilities and on the length of the longest chromosome arm. We find a simple formula for the stationary distribution for our model which fits the data on chromosome lengths in many, but not all, species.

Keywords: Reciprocal translocation; chromosome size model; Markov chain

AMS 2000 Subject Classification: Primary 60J05
Secondary 92D10

1. Introduction

There are many models for the evolution of DNA sequences due to the effects of local events such as nucleotide substitution, insertions, and deletions, but at this time there are few models that address the evolution of whole genomes. Humans and mice diverged from a common ancestor about 100 million years ago. Comparison of their genomes (see Nadeau and Taylor (1984), Nadeau and Sankoff (1998), and O'Brien *et al.* (1999)) shows that their evolution has involved three types of large-scale changes:

(i) Chromosome fissions and fusions: a chromosome splits into two, or two chromosomes combine into one.

(ii) Translocations: part of the ends of two chromosomes break off and reattach to the other chromosome.

(iii) Inversions: one chromosome breaks in two places and when it reassembles the central piece has changed its orientation.

Received 31 October 2000.
* Current address: 1704 White Ash Drive, Carmel, IN 46033, USA.
** Current address: 10512 Pilla Terra Court, Laurel, MD 20723, USA.
*** Current address: 1387 Burgundy Way, Placentia, CA 92870, USA.
**** Postal address: Department of Mathematics, Cornell University, Ithaca, NY 14853, USA.
Email address: rtd1@cornell.edu

Pioneering work of Sankoff and Ferretti (1996) considered a stochastic model for the evolution of the size of k chromosomes due to random reciprocal translocations. At each step, two of the chromosomes are chosen 'at random' and each is broken at a place that is uniformly distributed along the length. They then pair the left half of one with the right half of the other; for example,

$$\begin{array}{ll} xxxxxxxx|xxxx & xxxxxxxx|oooooo \\ oooooooo|oooooo & oooooooo|xxxx \end{array}$$

Sankoff and Ferretti explored two different meanings for the phrase 'at random': (i) all chromosomes are equally likely to be chosen (the *uniform model*), and (ii) each chromosome is chosen with a probability proportional to its length (the *proportional model*).

If there are only two chromosomes, the two schemes are identical, since in either case both chromosomes are always chosen. Sankoff and Ferretti were able to show that if the chromosome lengths are normalized to sum to 1, then in equilibrium the longer chromosome has length Y, where Y has probability density function

$$\mathrm{P}(Y = y) = 12y(1 - y) \quad \text{for } y \in [\tfrac{1}{2}, 1]. \tag{1}$$

From (1) and calculus it follows that the expected value $\mathrm{E}Y = \frac{11}{16}$. In other words, on average the longer chromosome will be about 70% of the entire genome.

Sankoff and Ferretti also considered in some detail the case of three chromosomes. Unfortunately, they defined the state of the system to be the lengths in decreasing order $\ell_1 > \ell_2 > \ell_3$. With this scheme, a translocation may change the rank of the lengths of any number of the chromosomes, so the equation for the stationary distribution became very complicated, and they were not able to find a simple formula for the stationary distribution. Without a simple formula, Sankoff and Ferretti were forced to 'compute' the stationary distribution by simulation.

Sankoff and Ferretti restricted their attention to the autosomes (non-sex chromosomes) since the X and Y chromosomes do not undergo reciprocal translocations with the autosomes (Ohno (1967)). Comparing their results to data from *Muntlacus muntjak* (barking deer) with $k = 3$ autosomes, the pea with $k = 7$, *Zea mays* (corn) with $k = 10$, *Oriza sativa* (rice) with $k = 12$, wheat with $k = 21$, and humans with $k = 22$, they found in all cases that the shortest chromosomes produced by their model were too short compared to those observed and the longest ones in the model were too long. Faced with this problem, Sankoff and Ferretti modified the model to impose a lower bound on the size of chromosomes that could be created. This modification resulted in a good fit for four of their examples, but for humans and wheat the long chromosomes were still too long.

2. Our models and their stationary distributions

Here, we will take a slightly different approach to modeling chromosome lengths. Consider k chromosomes with centromeres. Letting ℓ_{2j-1} and ℓ_{2j} be the lengths of the two arms of the jth chromosome, we arrive at a model with $2k$ chromosome arms. To keep the Markov chain theory simple, we will suppose that the length of the ith chromosome arm is a positive integer ℓ_i. More sophisticated readers can work directly with the corresponding chain with state space the set of vectors $(x_1, \ldots, x_{2k})$ of nonnegative real numbers that add up to 1. However, since reality corresponds to a discrete model with total genome size T in the millions or billions of base pairs, we will first develop a theory that works for any fixed finite genome length T and then investigate the simplifications in the solution that occur when $T \to \infty$.

The set of possible states of our model will be those length vectors $(\ell_1, \ldots, \ell_{2k})$ with $\ell_1 + \cdots + \ell_{2k} = T$, where T is the total genome size. In contrast to Sankoff and Ferretti, we will not write the lengths in order. To be picturesque, we will consider the ith chromosome being made up of LEGO® bricks. On any given move we may take from 1 to ℓ_i bricks from the ith stack. Each chromosome thus has ℓ_i 'cut points' and the genome has a total of T 'cut points'. On each transition we will pick two of these T possible cut points at random. We then pair the left half of one with the right half of the other. This procedure may pick the same chromosome arm twice, and hence not make a perceptible difference in chromosome length. In addition, our two cut points may be on different arms of the same chromosome. When this occurs, the total length of the chromosome does not change but the centromere may move. Biologists call this event a *pericentric inversion*.

Following Sankoff and Ferretti, we will first examine the case of two chromosome arms. That is, we have only one chromosome and all that is at stake is the position of the centromere on that chromosome. Specializing further, we will consider the case of two chromosome arms with a total of $T = 9$ LEGO bricks. Since our rules guarantee that each arm always has at least one brick, the state space is $(8, 1), (7, 2), \ldots, (1, 8)$. By considering the $\binom{9}{2} = 36$ possible choices of two distinct breakpoints, we can easily compute that 36 times the transition probability for our chain is

$$
\begin{array}{c} \\ (8,1) \\ (7,2) \\ (6,3) \\ (5,4) \\ (4,5) \\ (3,6) \\ (2,7) \\ (1,8) \end{array}
\begin{array}{c}
\begin{array}{cccccccc} (8,1) & (7,2) & (6,3) & (5,4) & (4,5) & (3,6) & (2,7) & (1,8) \end{array} \\
\left[\begin{array}{cccccccc}
28+1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 22+2 & 2 & 2 & 2 & 2 & 2 & 1 \\
1 & 2 & 18+3 & 3 & 3 & 3 & 2 & 1 \\
1 & 2 & 3 & 16+4 & 4 & 3 & 2 & 1 \\
1 & 2 & 3 & 4 & 16+4 & 3 & 2 & 1 \\
1 & 2 & 3 & 3 & 3 & 18+3 & 2 & 1 \\
1 & 2 & 2 & 2 & 2 & 2 & 22+2 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 28+1
\end{array}\right]
\end{array}.
$$

Here the diagonal elements are written as a sum of the number of choices that pick the same arm twice and the number that pick different arms and make no change, in order to make the patterns in the entries more apparent.

This transition probability is symmetric:

$$p(u, v) = p(v, u). \tag{2}$$

From (2) it follows that in the special case of nine bricks and two chromosome arms, the stationary distribution is the uniform distribution $\pi(u) = \frac{1}{8}$ for each of the 8 states. It is easy to see that the j-arm model has the same symmetry property. On a given step we will either (a) choose the same arm twice and nothing will happen, or (b) choose two different arms. In the second case we can assume without loss of generality that the two arms chosen are the first and the second. This having been done, the third to jth arms play no role, so we are reduced to the case of two chromosome arms.

Theorem 1. *The stationary distribution for the j-arm model with total genome size T is uniform over the set of all $(\ell_1, \ldots, \ell_j)$ with $\ell_1 + \cdots + \ell_j = T$.*

The human genome corresponds to the case with $T = 3 \times 10^9$ base pairs, so it is sensible to consider the rescaled lengths $x_i = \ell_i / T$ and let $T \to \infty$ to get a process with state space

the set of all $(x_1, \ldots, x_j)$ with $x_i \geq 0$ and $x_1 + \cdots + x_j = 1$. Let $\mathrm{P}(X_i = x)$ be the probability density function for the length of the ith chromosome arm. Symmetry implies that $\mathrm{P}(X_i = x) = \mathrm{P}(X_1 = x)$. It follows from Theorem 1 that

$$\mathrm{P}(X_1 = x) = C \,\mathrm{vol}(x_1 = x, x_2 + \cdots + x_{j-1} \leq 1 - x) = C'(1 - x)^{j-2} \tag{3}$$

by scaling, since we are taking the volume of a $(j-2)$-dimensional set. In order for (3) to define a probability distribution, we must have $C' = j - 1$. A little calculus shows that each chromosome arm has average size $1/j$, which must be true by symmetry.

To connect our results for the LEGO brick chain with a discrete version of the proportional model of Sankoff and Ferretti, let $h(u)$ be the probability that in our scheme we pick the same arm twice when the collection of lengths is $u = (\ell_1, \ldots, \ell_j)$. If we follow the rules of Sankoff and Ferretti and pick two different arms with probabilities proportional to their lengths, we arrive at a Markov chain with transition probability

$$s(u, v) = \begin{cases} \dfrac{p(u, v)}{1 - h(u)} & \text{if } u \neq v, \\ \dfrac{p(u, v) - h(u)}{1 - h(u)} & \text{if } u = v. \end{cases} \tag{4}$$

When $T = 9$, the transition probability matrix is

$$\begin{array}{c} \\ (8,1) \\ (7,2) \\ (6,3) \\ (5,4) \\ (4,5) \\ (3,6) \\ (2,7) \\ (1,8) \end{array} \begin{array}{c} \begin{array}{cccccccc} (8,1) & (7,2) & (6,3) & (5,4) & (4,5) & (3,6) & (2,7) & (1,8) \end{array} \\ \begin{bmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{1}{14} \\ \frac{1}{18} & \frac{2}{18} & \frac{3}{18} & \frac{3}{18} & \frac{3}{18} & \frac{3}{18} & \frac{2}{18} & \frac{1}{18} \\ \frac{1}{20} & \frac{2}{20} & \frac{3}{20} & \frac{4}{20} & \frac{4}{20} & \frac{3}{20} & \frac{2}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{2}{20} & \frac{3}{20} & \frac{4}{20} & \frac{4}{20} & \frac{3}{20} & \frac{2}{20} & \frac{1}{20} \\ \frac{1}{18} & \frac{2}{18} & \frac{3}{18} & \frac{3}{18} & \frac{3}{18} & \frac{3}{18} & \frac{2}{18} & \frac{1}{18} \\ \frac{1}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{2}{14} & \frac{1}{14} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix} \end{array}.$$

This matrix is not symmetric, owing to the different normalizations of the rows, so it is harder to guess the stationary distribution. However, from the definition of s given in (4) it is easy to see that if $u \neq v$ then

$$\{1 - h(u)\}\, s(u, v) = p(u, v) = p(v, u) = \{1 - h(v)\}\, s(v, u). \tag{5}$$

Summing (5) over u and using $\sum_u s(v, u) = 1$ gives

$$\sum_u \{1 - h(u)\} s(u, v) = 1 - h(v),$$

so the transition probability $s(u, v)$ has stationary distribution $\sigma(u) = C_T(1 - h(u))$, where C_T is a constant that depends on the total genome size T, and is chosen to make the sum of the $\sigma(u)$ equal to one.

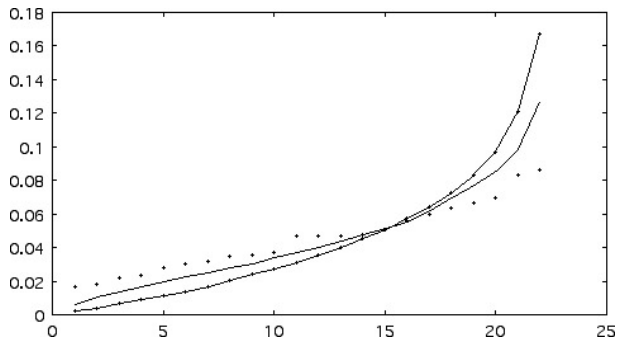

FIGURE 1: Chromosome lengths of humans (dots) compared to the predictions of our model (smooth line) and those of Sankoff and Ferretti (line with tick marks).

Considering scaled chromosome arm lengths $x_i = \ell_i / T$ and letting $T \to \infty$, we see that for Sankoff and Ferretti's proportional model $h(x) = x_1^2 + \cdots + x_j^2$, so the stationary distribution is

$$C(1 - x_1^2 - x_2^2 - \cdots - x_j^2) \quad \text{when } x_i \geq 0 \text{ and } \textstyle\sum_i x_i = 1.$$

In the case of two chromosomes, if we let $x_1 = x$ and $x_2 = 1 - x$, then the length of the first chromosome has density

$$\mathrm{P}(X_1 = x) = C'x(1 - x) \quad \text{for } 0 < x < 1. \tag{6}$$

In order for (6) to be a probability density, we must have $C' = 6$. To make the connection with Sankoff and Ferretti's result, note that if $X^1 = \max\{X_1, X_2\}$ is the length of the longer chromosome, then $\mathrm{P}(X^1 = x) = 2\mathrm{P}(X_1 = x)$ for $x \in [\frac{1}{2}, 1]$, in agreement with (1).

The calculations above show that Sankoff and Ferretti's proportional model is a time change of ours. In other words, if we ignore the transitions in our model that choose the same arm twice, then the result is the model of Sankoff and Ferretti. To understand the differences between the stationary distributions, we turn to simulation. With the human genome in mind, we choose $k = 22$. Figure 1 shows the sizes of the 22 human chromosomes as dots. The curves give the average chromosome lengths in the Sankoff and Ferretti proportional model (line with tick marks) and in our model with 44 chromosome arms (smooth line). In both cases, the chromosomes that the models produce have been sorted in order of increasing size and normalized so that their total length is 1. The curve gives the average of 100 simulations.

The line with tick marks which starts lower on the left and ends higher on the right is the result of the Sankoff and Ferretti model. The smooth line, which is somewhat closer to the data, is from our model. The reason for the difference between the models becomes clear if one simulates a random breakage model in which the unit interval is fractured into 22 pieces which are sorted in order of increasing size. The average of 100 simulations in this case is almost indistiguishable from the averages for the Sankoff and Ferretti model. Since our stationary distribution results from breaking the unit interval into 44 pieces and then adding randomly chosen pairs, chromosome sizes will be more uniform.

3. Adding relative fitness to the dynamics

As Figure 1 shows, the stationary distribution of our new model is more uniform than that of Sankoff and Ferretti, but still has short chromosomes that are too short and long chromosomes that are too long. As mentioned above, Sankoff and Ferretti fixed this problem by imposing an absolute lower limit on chromosome sizes. Here we will start with a softer version of their constraint by introducing a fitness function $\phi(\boldsymbol{u})$ that describes the relative fitness of an individual with chromosome arm lengths $\boldsymbol{u} = (\ell_1, \ldots, \ell_j)$.

To develop a concrete formula for $\phi(\boldsymbol{u})$ we will take as inspiration the fact that recombination is an important mechanism for pairing homologous chromosomes during meiosis. Since the rate of crossing over per base pair is not constant, we now switch to writing chromosome lengths in terms of centiMorgans (cM), a measure of chromosomal distance that corresponds to a 1% probability of recombination per generation. With lengths written in Morgans (= 100 cM) crossovers are a rate 1 Poisson process. We define our fitness $\phi(\boldsymbol{u})$ to be the probability that each pair of homologous chromosome arms experiences at least one recombination event:

$$\phi(\boldsymbol{u}) = \prod_{i=1}^{j} (1 - \mathrm{e}^{-\ell_i}). \tag{7}$$

Having decided on a form for our fitness function, the next step is to decide on how to modify the dynamics. Inspired by the well-known Metropolis algorithm, which can be used to look for the maximum of a function ϕ, we introduce the modified chain with transition probability

$$q(\boldsymbol{u}, \boldsymbol{v}) = \begin{cases} p(\boldsymbol{u}, \boldsymbol{v}) & \text{if } \phi(\boldsymbol{v}) \geq \phi(\boldsymbol{u}), \\ p(\boldsymbol{u}, \boldsymbol{v}) \dfrac{\phi(\boldsymbol{v})}{\phi(\boldsymbol{u})} & \text{if } \phi(\boldsymbol{v}) \leq \phi(\boldsymbol{u}). \end{cases} \tag{8}$$

In words, if the new state proposed by the transition probability has greater fitness, we always accept the proposed move. Otherwise, we accept the transition with probability $\phi(\boldsymbol{v})/\phi(\boldsymbol{u})$.

We will refer to this new Markov chain as the model with recombination fitness. The next result describes its stationary distribution.

Theorem 2. *The transition probability defined in (7) and (8) has*

$$\phi(\boldsymbol{u}) q(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{v}) q(\boldsymbol{v}, \boldsymbol{u}). \tag{9}$$

Consequently, $q(\boldsymbol{u}, \boldsymbol{v})$ has stationary distribution $C_T \phi(\boldsymbol{u})$, where C_T is a constant chosen to make the sum of the probabilities equal to 1.

Proof. We can suppose without loss of generality that $\phi(\boldsymbol{v}) \geq \phi(\boldsymbol{u})$, so $q(\boldsymbol{u}, \boldsymbol{v}) = p(\boldsymbol{u}, \boldsymbol{v})$ and $q(\boldsymbol{v}, \boldsymbol{u}) = p(\boldsymbol{v}, \boldsymbol{u}) \phi(\boldsymbol{u}) / \phi(\boldsymbol{v})$. Using the last two equalities, we have

$$\phi(\boldsymbol{u}) q(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u}) p(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{v}) \frac{\phi(\boldsymbol{u})}{\phi(\boldsymbol{v})} p(\boldsymbol{v}, \boldsymbol{u}) = \phi(\boldsymbol{v}) q(\boldsymbol{v}, \boldsymbol{u}),$$

where the second equality follows from the fact that $p(\boldsymbol{u}, \boldsymbol{v}) = p(\boldsymbol{v}, \boldsymbol{u})$. Having checked (9), the second conclusion follows immediately; see Durrett (1999, p. 61).

4. Fitting our model to data

We first found data on the world wide web (see below for URLs) on chromosome lengths measured in centiMorgans for eight different species. We then compared the predicted chromome lengths from our model with recombination fitness to the observed data. Figure 2 shows four successful fits. On each graph the line gives the observed sizes of chromosomes sorted in increasing order. The bars give 90% confidence intervals for the sizes of chromosomes in our stochastic model. In all four cases, the data stays between the 90% confidence intervals. Figure 3 shows four bad examples. In the case of mice, the observed chromosome lengths are much more uniform than predicted by the model. This deviation is made more surprising by the observation that for mice the centromere is always very near an end of the chromosome. That is, if we use a model with 38 chromosome arms, then 19 of these arms (one on each chromosome) are very short.

In the case of sheep the line of observed chromosome sizes are too small for chromosomes 19–23 and then suddenly too high on chromosomes 24 and 25. One possible explanation for the long chromosomes is that fusions have recently occurred and the system has not returned to its stationary distribution. In the case of wheat, the observed chromosome arm lengths are more uniform than predicted by our model. One possible explanation for this is the fact that since its divergence from rice, the wheat genome has undergone hexaploidization, which increased its haploid chromosome number from 7 to 21; see e.g. Gale and Devos (1998, p. 1973). After events like this one, it takes some time for chromosome sizes to return to equilibrium. In comparison, the rice genome is 25 times smaller than that of wheat and has undergone relatively few large-scale duplications (see e.g. Moore *et al.* (1995) or Wilson *et al.* (1999)). It has chromosome lengths that are not as uniform as those of wheat, but still lie outside the 90% confidence intervals at 4 of the 12 points.

5. Imposing a maximum length on chromosome arms

To improve the fit in the bad cases, we can make another change in the model. Sankoff and Ferretti modified their model by forbidding chromosomes that are too short. Here we will forbid chromosomes that are too long. In the notation of definition (8), we will use as fitness function

$$\phi(\boldsymbol{u}) = \prod_{i=1}^{j} (1 - \mathrm{e}^{-\ell_i}) \mathbf{1}_{(l_i \le b)}. \tag{10}$$

Here $\mathbf{1}_{(l_i \le b)}$ is 1 if $l_i \le b$ and is 0 otherwise, and b represents the bound on the maximum length of a chromosome arm, which is adjusted for fitting purposes. We will call the Markov chain that results from using our new ϕ in (8) the truncated model.

In support of this fitness function, we can cite Schubert and Oud (1997) whose experiments with the bean *Vicia fabia* suggested that 'for the normal development of an organism, the longest chromosome arm must not exceed half of the spindle axis at telophase'. In simpler language, if the end of a chromosome is in the wrong half of the cell while division occurs, then it becomes entangled with the cell wall with disastrous consequences for the genetic material. See Schubert and Oud (1997, p. 518) for some pictures of this disaster. Results presented in Table 1 of Schubert and Oud (1997, p. 517) show that when the longest chromosome arm is at most 21.3% of the whole genome, plants are normal, but when the longest arm is 21.7% or more, the offspring are mostly sterile. We take this as evidence that the sharp cutoff imposed in (10) is a reasonable approximation.

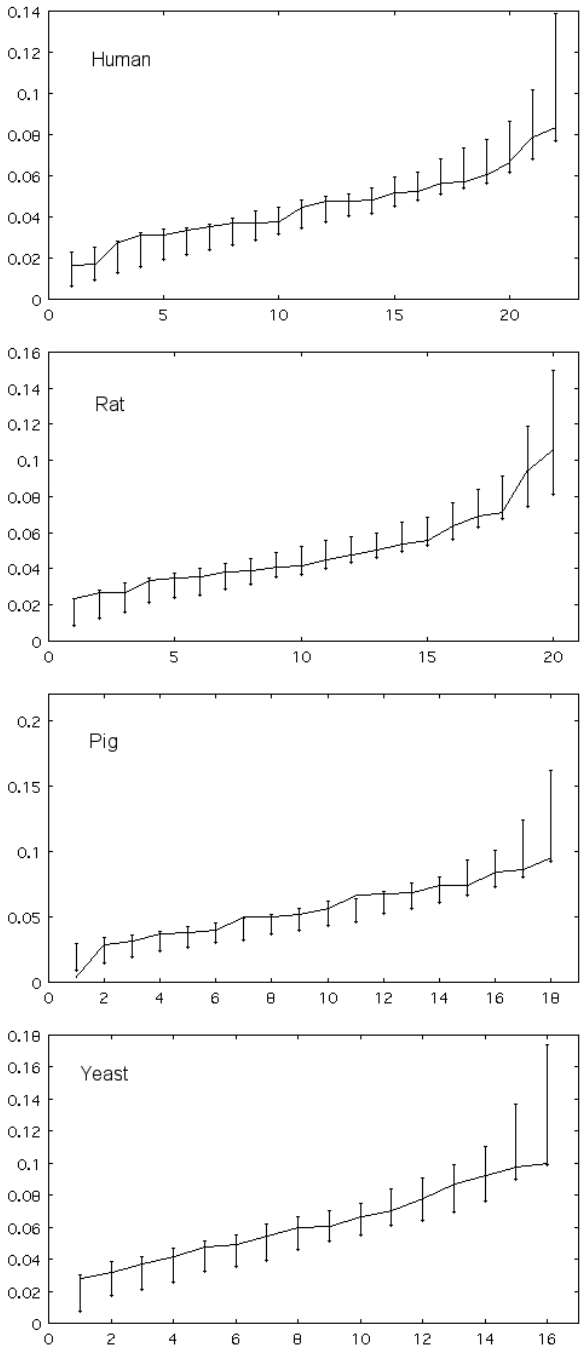

FIGURE 2: Fit of the model with recombination fitness to data from humans, rats, pigs, and yeast.

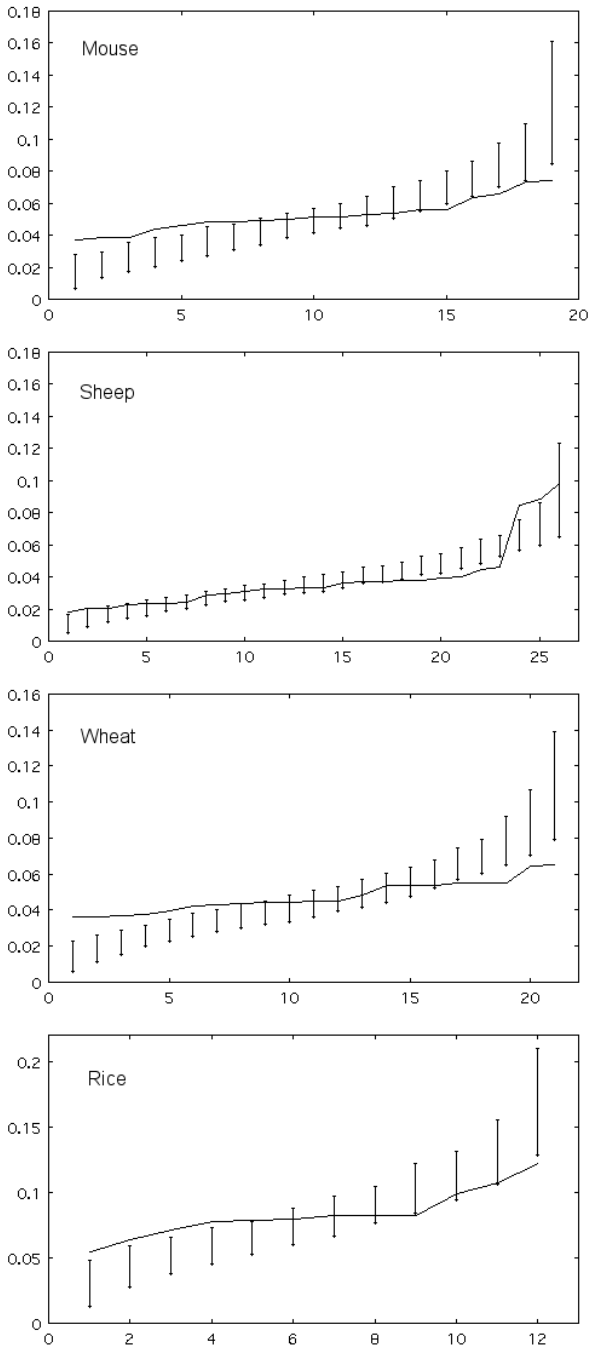

FIGURE 3: Fit of the model with recombination fitness to data from mice, sheep, wheat and rice.

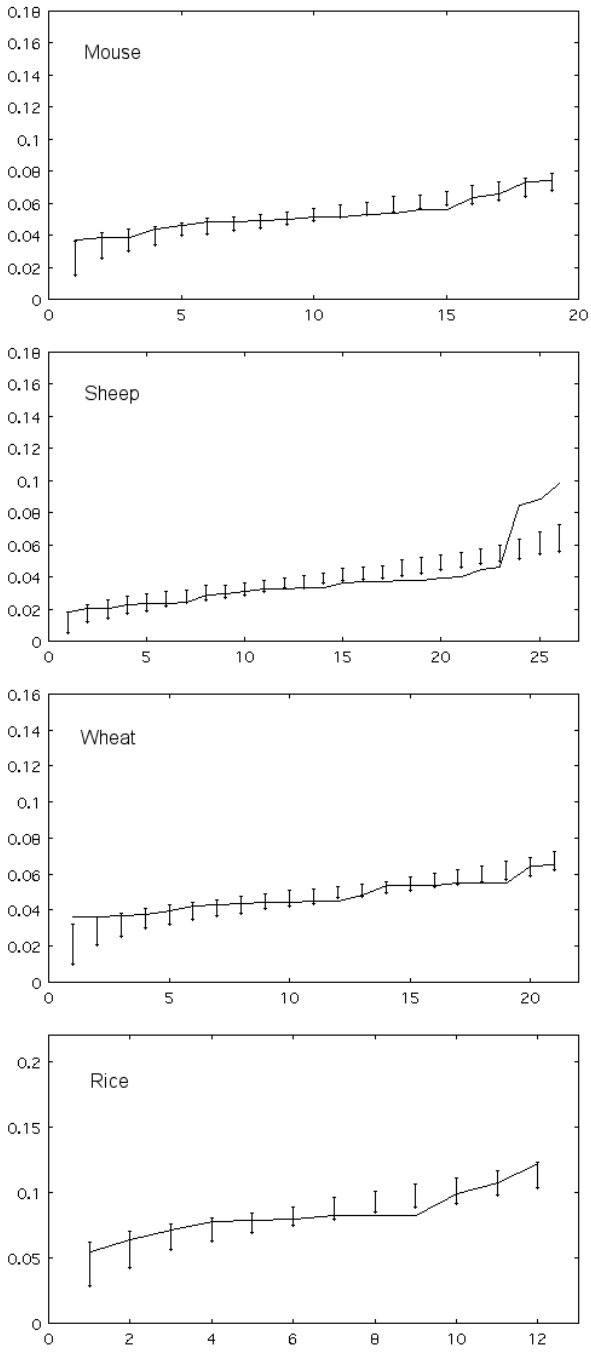

FIGURE 4: Fit of the truncated model to data from mice, sheep, wheat and rice.

As in the case of our first model with fitness (see Theorem 2), the new model has stationary distribution $C'_T\phi(\boldsymbol{u})$, where C'_T is a new constant chosen to make the sum of the probabilities one. Figure 4 shows the new truncated model fit for our four bad cases. The truncation makes the fit for sheep worse for the three largest chromosomes and now has chromosomes 16 to 23 too small. The fits for mouse, wheat, and rice improved since the data were more uniform than the previous prediction, and the new selection forces chromosome sizes to be more uniform.

Of course, the mere fact that the new model fits better does not mean that selection against chromosomes that are too long is acting to make chromosome size distributions more uniform. In wheat, for example, the hexaploidization mentioned above is another explanation for the departure from the model's equilibrium distribution. However, comparison of observed chromosome lengths with that of a model of random translocations suggests that some form of selection is causing the distribution of lengths to be more uniform.

Acknowledgements

The authors would like to express their appreciation to Todd Vision for many helpful comments before and after the paper was written. The research of De, Ferguson, and Sindi was supported by an REU grant from the National Science Foundation. Durrett's research is partially supported by a grant from the probability program at NSF. This work was done in a Research Experience for Undergraduates held at Cornell University, June 19–August 11, 2000.

References

DURRETT, R. (1999). *Essentials of Stochastic Processes.* Springer, New York.

GALE, M. D. AND DEVOS, K. M. (1998). Comparative genetics in the grasses. *Proc. Nat. Acad. Sci. USA* **95**, 1971–1974.

MOORE, G. *et al.* (1995). Was there a single ancestral cereal chromosome? *Trends Genetics* **11**, 81–82.

NADEAU, J. H. AND SANKOFF, D. (1998). The lengths of undiscovered conserved segments in comparative maps. *Mammalian Genome* **9**, 491–495.

NADEAU, J. H. AND TAYLOR, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Nat. Acad. Sci. USA* **81**, 814–818.

O'BRIEN, S. J. *et al.* (1999). The promise of comparative genomics in mammals. *Science* **286**, 458–481.

OHNO, S. (1967). *Sex Chromosomes and Sex-linked Genes.* Springer, New York.

SANKOFF, D. AND FERRETTI, V. (1996). Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Res.* **6**, 1–9.

SCHUBERT, I. AND OUD, J. L. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell* **88**, 515–520.

WILSON, W. A. *et al.* (1999). Inferences on the genome structure through comparative analysis of rice, maize, and the domesticated panicoids. *Genetics* **153**, 453–473.

URLs for data sets

Anubis Map Selector (pig), Roslin Institute. Available at http://www.ri.bbsrc.ac.uk/cgi-bin/mapviewer/.

Combined Physical and Genetic Maps of *S.cerevisiae* (yeast), Stanford Genomic Resources. Available at http://genome-www4.stanford.edu/cgi-bin/SGD/PGMAP/pgMap/.

Consensus Map of Hexaploid Wheat, Graingreens Database, National Agriculture Library. Available at gopher://greengenes.cit.cornell.edu:70/11/.maps/Triticum/.

Mouse Genome Informatics, The Jackson Laboratory. Available at http://www.informatics.jax.org/mgihome/other/mouse_facts1.shtml.

RATMAP: the Rat Genome Database, Göteborg University. Available at http://ratmap.gen.gu.se/.

Rice Genetic Map in Nature Genetics, Rice Genome Research Program. Available at http://rgp.dna.affrc.go.jp/dnabank/.

Science: the Human Gene Map, National Center for Biotechnology Information. Available at http://www.ncbi.nlm.nih.gov/SCIENCE96/.

Sheep: sizes of sheep maps. Available at http://rubens.unimelb.edu.au/∼jillm/pages/smapsize.htm.