Discrete Mathematics and Theoretical Computer Science **AC**, 2003, 17–26

A phase transition in the random transposition random walk

Nathanaël Berestycki[1,2] and Rick Durrett[1 †]

[1] *Cornell University, Department of Mathematics. Mallott Hall, Ithaca, NY 14853, U.S.A.*
[2] *Ecole Normale Supérieure, Département de Mathématiques et Applications, 45, rue d'Ulm F-75005 Paris, France*
`berestycki@dma.ens.fr, rtd1@cornell.edu`

Our work is motivated by Bourque and Pevzner (2002)'s simulation study of the effectiveness of the parsimony method in studying genome rearrangement, and leads to a surprising result about the random transposition walk in continuous time on the group of permutations on n elements starting from the identity. Let D_t be the minimum number of transpositions needed to go back to the identity element from the location at time t. D_t undergoes a phase transition: for $0 < c \leq 1$, the distance $D_{cn/2} \sim cn/2$, i.e., the distance increases linearly with time; for $c > 1$, $D_{cn/2} \sim u(c)n$ where u is an explicit function satisfying $u(x) < x/2$. Moreover we describe the fluctuations of $D_{cn/2}$ about its mean at each of the three stages (subcritical, critical and supercritical). The techniques used involve viewing the cycles in the random permutation as a coagulation-fragmentation process and relating the behavior to the Erdös–Rényi random graph model.

Keywords: random transposition, random graphs, phase transition, coagulation-fragmentation

1 General motivation

The relationship between the orders of genes in two species can be described by a signed permutation. For example the relationship between the human and mouse X chromosomes may be encoded as (see Pevzner and Tesler (2003))

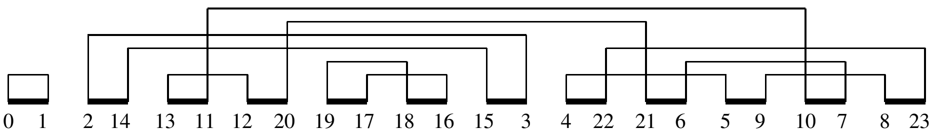
$$1 \quad -7 \quad 6 \quad -10 \quad 9 \quad -8 \quad 2 \quad -11 \quad -3 \quad 5 \quad 4$$

In words the two X chromosomes can be partitioned into 11 segments. The first segment of the mouse X chromosome is the same as that of humans, the second segment of mouse is the 7th human segment with its orientation reversed, etc. The parsimony approach to estimation of evolutionary changes of the X chromosome between human and mouse is to ask: what is the minimum number of reversals (i.e., moves that reverse the order of a segment and therefore change its *sign*) needed to transform the arrangement above back into $1,\ldots,11$?

Hannehalli and Pevzner (1995) developed a polynomial algorithm for answering this question. The first step in preparing to use the HP algorithm is to double the markers. When segment i is doubled we

†Both authors are partially supported by a joint NSF-NIGMS grant

Fig. 1: Breakpoint graph for human-mouse X chromosome comparison



replace it by two consecutive numbers $2i-1$ and $2i$, e.g., 6 becomes 11 and 12. A reversed segment $-i$ is replaced by $2i$ and $2i-1$, e.g., -7 is replaced by 14 and 13. The doubled markers use up the integers 1 to 22. To these we add a 0 at the front and a 23 at the end. Using commas to separate the ends of the markers we can write the two genomes as follows:

$$\begin{array}{rl} \text{mouse} & 0,1\,2,14\,13,11\,12,20\,19,17\,18,16\,15,3\,4,22\,21,6\,5,9\,10,7\,8,23 \\ \text{human} & 0,1\,2,3\,4,5\,6,7\,8,9\,10,11\,12,13\,14,15\,16,17\,18,19\,20,21\,22,23 \end{array}$$

The next step is to construct the breakpoint graph (see figure 1) which results when the commas are replaced by edges that connect vertices with the corresponding numbers. In the picture we write the vertices in their order in the mouse genome. Commas in the mouse order become thick lines (black edges), while those in the human genome are thin lines (gray edges).

Each vertex has one black and one gray edge so its connected components are easy to find: start with a vertex and follow the connections in either direction until you come back to where you start. In this example there are five cycles:

$$\begin{array}{c} 0-1-0 \qquad 2-14-15-3-2 \qquad 4-22-23-8-9-5-4 \\ 19-17-16-18-19 \qquad 13-11-10-7-6-21-20-12-13 \end{array}$$

To compute a lower bound for the distance now we take the number of commas seen when we write out one genome. In this example that is 1 plus the number of segments (12), then we subtract the number of connected components in the breakpoint graph. In this example that is 5, so the result is 7. This is a lower bound on the distance since any reversal can at most reduce this quantity by 1, and it is 0 when the two genomes are the same. We can verify that 7 is the minimum distance by constructing a sequence of 7 moves that transforms the mouse X chromosome into the human order. We leave this as an exercise for the reader. Here are some hints: (i) To do this it suffices to at each step choose a reversal that increases the number of cycles by 1. (ii) This never occurs if the two chosen black edges are in different cycles. (iii) If the two black edges are in the same cycle and are (a,b) and (c,d) as we read from left to right, this will occur unless in the cycle minus these two edges a is connected to d and b to c, in which case the number of cycles will not change. For example in the graph above a reversal that breaks black edges 19-17 and 18-16 will increase the number of cycles but the one that breaks 2-14 and 15-3 will not.

In general the distance between genomes can be larger than the lower bound from the breakpoint graph. There can be obstructions called *hurdles* that can prevent us from decreasing the distance and hurdles can be intertwined in a *fortress of hurdles* that takes an extra move to break. (See Hannehalli and Pevzner (1995).) In symbols if π is the signed permutation that represents the relative order and orientation of

36	37	17	40	16	15	14	63	10	9
55	28	13	51	22	79	39	70	66	5
6	7	35	64	33	32	60	61	18	65
62	12	1	11	23	20	4	52	68	29
48	3	21	53	8	43	72	58	57	56
19	49	34	59	30	77	31	67	44	2
27	38	50	26	25	76	69	41	24	75
71	78	73	47	54	45	74	42	46	

Tab. 1: Order of the genes in *D. repleta* compared to their order in *D. melanogaster*

segments in the two genomes

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

where $d(\pi)$ is the distance from the identity, n is the number of markers, $c(\pi)$ is the number of components in the breakpoint graph, $h(\pi)$ is the number of hurdles, and $f(\pi)$ is the indicator of the event π is a fortress of hurdles. See Section 5.2 of Durrett (2002) or Chapter 10 of Pevzner (2000) for more details.

To motivate our main question, we will introduce a second data set. Ranz et al. (2001) located 79 genes on chromosome 2 of *D. repleta* and on chromosome arm 3R of *D. melanogaster*. If we number the genes according to their order in *D. repleta* then their order in *D. melanogaster* is given by table 1. This time we do not know the orientation of the segments but that is not a serious problem. Using simulated annealing one can easily find an assignment of signs that minimizes the distance, which in this case is 54. Given the large number for rearrangements relative to the number of markers we should ask: when is the parsimony estimate reliable?

Bourque and Pevzner (2002) have approached this question by taking 100 markers in order and performing k randomly chosen reversals and computing Δ_k, the minimum number of reversals ($d(\pi)$) needed to return to the identity and then plotting the average value of $\Delta_k - k \leq 0$ (the circles in fig. 2). They concluded based on this and other simulations (see figure 2) that the parsimony distance on n markers was a good one as long as the number of reversals performed was at most $0.4n$. The smooth curve, which we will describe below, gives the limiting behavior of $(D_{cn} - cn)/n$ (as a function of c).

For simplicity we will consider the analogous problem for random transpositions. In that case the distance from the identity can be easily computed: it is the number of markers n minus the number of cycles in the permutation. For an example, consider the following permutation of 14 objects written in its cyclic decomposition:

$$(1\,7\,4)\,(2)\,(3\,12)\,(5\,13\,9\,11\,6)\,(8\,10\,14)$$

which indicates that $1 \to 7$, $7 \to 4$, $4 \to 1$, $2 \to 2$, $3 \to 12$, $12 \to 3$, etc. There are 5 cycles so the distance from the identity is 9. If we perform a transposition that includes markers from two different cycles (e.g., 7 and 9) the two cycles merge into 1, while if we pick two in the same cycle (e.g., 13 and 11) it splits into two.

The situation is similar but slightly more complicated for reversals. There if we ignore the complexity of hurdles, the distance is $n+1$ minus the number of components in the breakpoint graph. A reversal that involves edges in two different components merges them into 1 but a reversal that involves two edges of the same cycle may or may not increase the number of cycles. To have a cleaner mathematical problem, we will consider the biologically less relevant case of random transpositions, and ask a question that in

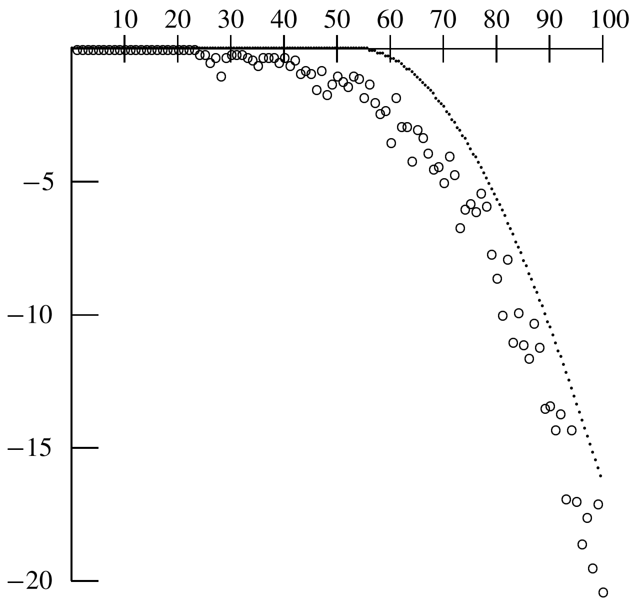


Fig. 2: Average values of Δ_k computed by simulation (circles) and approximated by Theorem 3 (smooth line)

terms of the rate 1 continuous time random walk on the symmetric group is: how far from the identity are we at time $cn/2$?

2 The coagulation-fragmentation process and the random graph process

Let $(S_t, t \geq 0)$ be the continuous-time random walk on the symmetric group $\mathfrak{S}_n$ starting at the identity, in which at times of a rate one Poisson process, we perform a transposition of two elements chosen uniformly at random.

Definition 1. *The distance to the identity D_t is the minimum number of transpositions you need to perform on S_t to go back to the identity element. In particular, if N_t is the number of transpositions performed up to time t (thus a Poisson r.v. with mean t), then $D_t \leq N_t$.*

As mentioned earlier D_t is given by the obvious but crucial formula $D_t = n - |S_t|$ where $|S_t|$ is the number of cycles in the cycle decomposition of S_t. The cycles evolve according to the dynamics of a coagulation-fragmentation process. When a transposition (i, j) occurs, if i and j belong to two different cycles then the cycles merge. On the contrary, if they belong to the same cycle, this cycle is split into two cycles.

We will be working exclusively in the continuous-time setting but it will be convenient to introduce the discrete-time analogue of $(S_t, t \geq 0)$, which we denote by $(\sigma_k, k \in \mathbb{N})$. Then $(\Delta_k, k \in \mathbb{N})$ will stand for its distance to the identity and we have $\Delta_k \leq k$.

In order to avoid confusions between the cycles of the permutation and the cycles in the graph that we define in the next paragraph, we will denote the first by σ-cycles and the second by G-cycles. From the definition it can be seen that the ranked sizes of the σ-cycles form a (non-homogeneous) coagulation-fragmentation process (see Pitman (2002a) and Pitman (2002b), Aldous (1997b)) in which components of size x and y merge at rate $K_n(x,y) = xy/n^2$ and components of size x split at rate $F_n(x) = x(x-1)/n^2$ and are broken at a uniformly chosen random point.

To study the evolution of the cycles in the random permutation, we construct a random (undirected) graph process. Start with the initial graph on vertices $\{1,\dots,n\}$ with no edge between the vertices. When a transposition of i and j occurs in the random walk, draw an edge between the vertices i and j. It is elementary to see that at time t this graph is a realization of the Erdös–Rényi random graph $G(n,p)$, see Bollobàs (1985) or Janson et al. (2000), in which edges are independently present with probability $p = 1-\exp(-t/\binom{n}{2})$. It is also easy to see that in order for two integers to be in the same σ-cycle it is necessary that they are in the same component of the random graph.

To get a result in the other direction, let F_t denote the event that a fragmentation occurs at time t. It is clear that

$$D_t = N_t - 2\sum_{s\leq t} \mathbf{1}_{\{F_s\}} \tag{1}$$

A fragmentation occurs in the random permutation when a transposition occurs between two integers in the same σ-cycle, so tree components in the random graph correspond to cycles in the random walk. Unicyclic components (with an equal number of vertices and edges) correspond to σ-cycles that have experienced exactly one fragmentation, but we need to know the order in which the edges were added to determine the resulting σ-cycles. For more complex components, the relationship between the random graph and the permutation is less clear. Fortunately, these can be ignored in the subcritical and the critical regimes.

3 Limit Theorems

We will now describe our results and sketch their proofs. A version of this article with complete proofs will be published later.

3.1 The subcritical regime

The observations above lead easily to:

Theorem 1. *Let $0 < c < 1$. Then $\mathbb{E}[D_{cn/2}] \underset{n\to\infty}{\sim} cn/2$ and the number of fragmentations*

$$Z_c := \sum_{s\leq \frac{1}{2}cn} \mathbf{1}_{\{F_s\}} \xrightarrow[n\to\infty]{\mathcal{D}} \text{Poisson}(\kappa(c)) \tag{2}$$

where $\kappa(c) = (-\log(1-c)-c)/2$. In fact the convergence holds for the process $\{Z_c : 0 \leq c < 1\}$ with the limit being a Poisson process with compensator $\kappa(c)$.

Remark. Theorem 1 has the following analogue in discrete time : Let $k = \lfloor cn/2 \rfloor$, then

$$\Delta_k - k \xrightarrow[n\to\infty]{\mathcal{D}} \text{Poisson}(\kappa(c)) \tag{3}$$

Thus the fluctuations are of order 1 in the subcritical discrete regime.

Sketch of the proof. The process $\{Z_c, 0 \le c < 1\}$ is a càdlàg counting process. Therefore by arguments from Jacod and Shiryaev (1987) it is enough to show that its compensator converges to a deterministic limit being the intensity $\kappa(c)$. If $f_k(t)$ is the number of vertices that belong to components of size k divided by n, the rate at which fragmentations occur is just $\sum_k f_k(t)(k-1)/n$. Hence the compensator is just the integral w.r.t time of this rate. We first show that the variance converges to 0 and then calculate its expectation. By exchangeability $\mathbb{E}[f_k(t)] = \mathbb{P}[|\mathcal{C}_1| = k]$ where $|\mathcal{C}_1|$ is the size of the component that contains 1 at time t. It is not hard to see that this quantity at time $bn/2$ converges in distribution to the total progeny τ of a branching process with offspring distribution Poisson(b), which happens to be subcritical since $b < 1$ (see lemma 52 in Pitman (2002a)). Using the exploration random walk associated with such a branching process, we find that the expected value of τ is $1/(1-b)$ (see Pitman (1997)). Hence by integrating w.r.t b we get the desired expected value, $\kappa(c)$. $\square$

To prepare for later developments, it is useful to take a second combinatorial approach to this result. We begin with Cayley's result that there are k^{k-2} trees with k labelled vertices. At time $cn/2$ each edge is present with probability $\approx cn/2\binom{n}{2} \approx c/n$ so the expected number of trees of size k present is

$$\binom{n}{k} k^{k-2} \left(\frac{c}{n}\right)^{k-1} \left(1 - \frac{c}{n}\right)^{k(n-k)+\binom{k}{2}-k+1}$$

since each of the $k-1$ edges need to be present and there can be no edges connecting the k point set to its complement or any other edges connecting the k points. For fixed k the above is

$$\approx n\frac{k^{k-2}}{k!} c^{k-1} \left(1 - \frac{c}{n}\right)^{kn}$$

The quantity in parentheses at the end converges to e^{-ck} so we have an asymptotic formula for the the number of tree components at time $cn/2$. As a side result we get :

Corollary 1. *The probability distribution of the total progeny T of a* Poisson(c) *branching process with $c < 1$ is given by* $\mathbb{P}[T = k] = \frac{1}{c}\frac{k^{k-1}}{k!}(ce^{-c})^k$

See section 4.1 of Pitman (1998) for another proof of this result. It was first discovered by Borel (1942) and the distribution of T is called the Borel distribution. It is a particular case of the so-called Borel-Tanner distribution, see Devroye (1992) and Pitman (1997) for further references. In this context it appeared in the problem of the total number of units served in the first busy period of a queue with Poisson arrivals and constant service times. Of course, this becomes a branching process if we think of the customers that arrive during a person's service time as their children.

3.2 The critical regime

Now let us consider the more delicate and also more interesting critical regime. It is well known in the theory of random graphs that the critical regime takes place at times $t_n^{\text{crit}}(\lambda) = \frac{1}{2}n(1+\lambda n^{-1/3}), \lambda \in \mathbb{R}$ (see

Aldous (1997a) for an extremely interesting account of some properties of the critical random graph such as cluster growth, and its relation to the multiplicative coalescent). Until this range of times we are still in the subcritical regime so that the arguments in the proof of theorem 1 are still valid. More precisely, we can show that if $c_n(r) = 1 - n^{-r/3}$ for $0 \le r \le 1$, then the expected number of fragmentations up to time $c_n(r)n/2$ is again given by $\kappa(c_n(r)) \underset{n\to\infty}{\sim} (r/6)\log n$. Hence define :

$$W_n(r) := \left(\frac{6}{\log n}\right)^{1/2} (Z_{c_n(r)} - \frac{r}{6}\log n) = \left(\frac{6}{\log n}\right)^{1/2} \left(\sum_{s \le c_n(r)\frac{n}{2}} \mathbf{1}_{\{F_s\}} - \frac{r}{6}\log n \right) \tag{4}$$

Theorem 2. *The following convergence holds with respect to the Skorokhod topology on the space of càdlàg functions on* $[0,1]$ *:*

$$W_n(\cdot) \xrightarrow[n\to\infty]{\mathcal{D}} W(\cdot) \tag{5}$$

where $\{W(r), 0 \le r \le 1\}$ *is a standard Brownian Motion on* $[0,1]$.

Sketch of the proof. $W_n(r)$ is by definition a martingale whose jumps are asymptotically zero and whose quadratic variation process is r thanks to our time-change $c_n(r) = 1 - n^{-r/3}$. Therefore it converges to Brownian Motion. □

After time $(1 - n^{-1/3})n/2 = t_n^{\text{crit}}(-1)$ we are in the critical range of the random graph. Then expected value estimates (see Łuczak et al. (1994) and Janson et al. (1993)) imply that the number of fragmentations between times $t_n^{\text{crit}}(-1)$ and $t_n^{\text{crit}}(+1)$ is bounded in expectation and hence can be ignored. Therefore we have the

Corollary 2. *Let* $\lambda \in \mathbb{R}$ *be any fixed real number. Then the number of fragmentations up to time* $t_n^{crit}(\lambda)$ *satisfies :*

$$\left(\frac{6}{\log n}\right)^{1/2} \left(\sum_{s \le t_n^{\text{crit}}(\lambda)} \mathbf{1}_{\{F_s\}} - \frac{1}{6}\log n \right) \xrightarrow[n\to\infty]{\mathcal{D}} W(1), \tag{6}$$

a standard Gaussian random variable. In particular for $\lambda = 0$ *the central limit theorem holds at time* $n/2$.

3.3 The supercritical regime

This is the most interesting case, and also the hardest one. We start by establishing a law of large numbers. For all $c > 0$ define $\beta_k(c) = \frac{1}{c}\frac{k^{k-1}}{k!}(ce^{-c})^k$ so that for $c < 1$ it coincides with the Borel distribution of Corollary 1. When $c > 1$,

$$\lim_{n\to\infty} \mathbb{P}[|\mathcal{C}_1| = k] = \beta_k(c)$$

still holds but the $\beta_k(c)$'s no longer sum up to 1 because there is a probability $\beta_\infty(c) > 0$ that $\mathcal{C}_1$ is the giant component (this is also the probability that a supercritical branching process with offspring distribution Poisson(c) doesn't die out). Then $\beta_\infty(c) = 1 - \sum_{k\ge1}\beta_k(c)$.

Let us denote by $\Upsilon(c)$ a random variable that take value $1/k$ with probability $\beta_k(c)$ when $1 \le k < \infty$ and 0 with probability $\beta_\infty(c)$.

Theorem 3. *Let $c > 0$ be a fixed positive number. Then the expected number of cycles in the random walk at time $cn/2$ is $g(c)n + O(\sqrt{n})$, where :*

$$g(c) := \mathbb{E}[\Upsilon(c)] = \sum_{k=1}^{\infty} \frac{1}{c} \frac{k^{k-2}}{k!} (ce^{-c})^k, \ c > 0 \tag{7}$$

and for $c < 1$, $g(c) = 1 - c/2$. In particular the distance is given by $D_{cn/2} = u(c)n + O(\sqrt{n})$ where $u(c) = 1 - g(c) < c/2, c > 0$

Note that the theorem is valid for all regimes. The behavior of the function g before and after $c = 1$ is very different : thus there is phase transition in the behavior of the distance of the random walk to the identity at time $n/2$ from linear to sublinear.

Sketch of the proof. In the critical regime the dynamics of the large components is quite complicated but (i) there can never be more than $\sqrt{n}$ components of size $\sqrt{n}$ or larger and (ii) an easy argument shows that the number of fragmentations occurring to clusters of size $\leq \sqrt{n}$ is at most $O(\sqrt{n})$. These two observations plus results from the theory of random graphs (Bollobàs (1985), Theorem 5.12) imply the result stated above. □

Remark. Theorem 3 extends easily to the cycle distance for random reversals and thus explains the result found by simulation in Bourque and Pevzner (2002). However in studying fluctuations it is important to know whether a reversal that acts on a single cycle increases the distance or leaves it the same, so our results on fluctuations are restricted to the case of random transpositions.

Theorem 4. *Let $\sigma(c)^2 = \mathrm{var}(\Upsilon(c))$. Then*

$$\frac{D_{cn/2} - u(c)n}{\sigma(c)n^{1/2}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1) \tag{8}$$

Sketch of the proof. This is much harder than Theorem 3 since if we restrict our attention to cycles of size $\leq n^a$ with $a > 1/2$ to make the number of large cycles $o(\sqrt{n})$ then the number of fragmentations that involve large components will be $O(n^a)$ and hence too big to ignore. To prove Theorem 4 one must use the fact that the fragmentation of the large components is a (time-changed) $M/M/\infty$ queue. Pieces of a fixed size k break off the giant component at a rate that depends on the total mass of large components, but each fragment of that size is re-absorbed at k times that rate, so at any time there are only $O(\log n)$ extra pieces. Once this estimation is done, the next step is to realize that the number $\Gamma(c)$ of components at time $cn/2$, is given by $\Gamma(c) = \sum_i \frac{1}{C_i}$ where C_i is the size of the component containing i. $-1/C_i$ are increasing functions of the i.i.d. random variables that define the random graph, so they are associated. They are also asymptotically uncorrelated, so the desired conclusion follows from an argument of Newman and Wright Newman and Wright (1981). □

Remark. Let $\{\Upsilon_i(c), i \geq 1\}$ and $\{\zeta_i(c), i \geq 1\}$ denote i.i.d random variables such that $\Upsilon_i(c) \overset{\text{(law)}}{=} \Upsilon(c)$ and $\zeta_i(c) \overset{\text{(law)}}{=} 1 - \Upsilon(c)$. Then the above argument gives that the number of components of the supercritical random graph behaves asymptotically as $\sum_{1\leq i\leq n} \Upsilon_i(c)$ so that the distance to the identity, which is just n minus the number of components, behaves asymptotically as $\sum_{1\leq i\leq n} \zeta_i(c)$. This provides an explanation for both Theorem 3 and 4.

3.4 Equilibrium behavior

Our final topic is to consider the behavior of the cycle structure of the random permutation as $c \to \infty$. As is well known, (Pitman (2002a), Arratia et al. (2001)), the cycle sizes of a random permutation when divided by n tend to a Poisson-Dirichlet distribution. In Section 2, we observed that the cycles themselves are a coagulation-fragmentation process. Ranking and letting $n \to \infty$ and gives the following result :

Theorem 5. *The Poisson-Dirichlet distribution with parameters 0 and 1 is an invariant measure for the ranked coagulation-fragmentation process on the space of real partitions of 1, with corresponding kernels given by $K(x,y) = xy$ and $F(x) = K(x,x) = x^2$.*

This result has already been recently derived in various ways (including this weak convergence argument) in Pitman (2002b) and Mayer-Wolf et al. (2002). The uniqueness of the invariant measure for this process was conjectured by Vershik and proved in 2002 in Diaconis et al. (2003).

References

D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Prob.*, 25(2):812–854, 1997a.

D. Aldous. Deterministic and stochastic models for coalescence (aggregation, coagulation) : a review of the mean-field theory for probabilists. Technical report, Dept. of Statistics, U.C. of Berkeley, 1997b.

R. Arratia, A. Barbour, and S. Tavaré. *Logarithmic combinatorial structures : a Probabilistic approach*. book, preprint, 2001.

B. Bollobàs. *Random Graphs*. Cambridge University Press, 1985.

E. Borel. Sur l'emploi du théorème de Bernouilli pour faciliter le calcul d'une infinité de coefficients. Application au problème de l'attente à un guichet. *C.R. Acad. Sci. Paris.*, 214:452–456, 1942.

G. Bourque and P. A. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12:26–36, 2002.

L. Devroye. The branching process method in the Lagrange random variate generation. Technical report, McGill University, `cgm.cs.mcgill.ca/~luc/branchingpaper.ps`, 1992.

P. Diaconis, E. Mayer-Wolf, O. Zeitouni, and M. Zerner. Uniqueness of invariant distributions for split-merge transformations and the Poisson-Dirichlet law. *Ann. Prob.*, 2003. to appear.

R. Durrett. *Probability : Theory and Examples*. Duxbury Press, second edition, 1996.

R. Durrett. *Probability models for DNA Sequence evolution*. Probability and Its Applications. Springer-Verlag, New York, 2002.

R. Durrett. Shuffling chromosomes. *Ann. App. Prob., to appear*, 2003.

S. Hannehalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 27^{th} Annual Symposium on the Theory of Computing*, pages 178–189, 1995. full version in the *Journal of the ACM*, 46:1-27.

J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer New-York, 1987.

S. Janson, D. E. Knuth, T. Łuczak, and B. Pittel. The birth of the giant component. *Rand. Struct. Algor.*, 4(3), 1993.

S. Janson, T. Łuczak, and A. Ruczinski. *Random Graphs*. Wiley-Interscience, New York, 2000.

T. Łuczak, B. Pittel, and J. C. Wierman. The structure of a random graph near the point of the phase transition. *Trans. Amer. Math. Soc.*, 341(2):721–748, February 1994.

E. Mayer-Wolf, O. Zeitouni, and M. P. W. Zerner. Asymptotics of certain coagulation-fragmentation processes and invariant poisson-dirichlet measures. *Electr. Journ. Prob.*, 7:1–25, 2002.

C. Newman and A. Wright. An invariance principle for certain dependent sequences. *Ann. Prob.*, 9: 671–675, 1981.

P. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, 2000.

P. Pevzner and G. Tesler. Genome rearrangement in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, 13:37–45, 2003.

J. Pitman. Enumerations of trees and forests related to branching processes and random walks. Technical Report 482, Department of Statistics, U.C. Berkeley, `www.stat.berkeley.edu/~pitman`, 1997.

J. Pitman. Coalescent random forests. Technical Report 457, Dept. Statistics, U.C. Berkeley, 1998.

J. Pitman. Combinatorial stochastic processes. In *Lecture Notes for St. Flour Course*, Technical Report 621, Dept. Statistics, U.C. Berkeley, July 2002a. `www.stat.berkeley.edu/~pitman`.

J. Pitman. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combin. Prob. Comput.*, 11:501–514, 2002b.

J. Ranz, F. Casals, and A. Ruiz. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Research*, 11:230–239, 2001.