The genealogical tree of a chromosome

Bernard Derrida, Bernard Jung-Muller

Journal of Statistical Physics, Vol. 94, Nos. 3/4, 1999

The Genealogical Tree of a Chromosome

B. Derrida[1] **and B. Jung-Muller**[2]

Received July 20, 1998

We consider a simple neutral model to describe the genealogy of chromosomes by taking into account the effects of both recombination and coalescence. Seen as a statistical physics problem, the model looks like an inverse problem: A number of properties such as pair or three-point correlations can be computed easily, but the prediction of global properties, in particular the average number of ancestors, remains difficult. In the absence of exact solutions, these global properties can nevertheless be estimated by the usual approximations: series expansions, Monte Carlo simulations, mean-field theory. Simulations exhibit also non-self-averaging properties similar to those of mean-field spin glasses.

KEY WORDS: Steady state; spin chain; random genealogies.

1. INTRODUCTION

One of the simplest questions one can ask about the genealogy of a living population is: what is the number of ancestors of a given individual?

When the reproduction is asexual (as generally in bacteria), this question has a trivial answer: each individual has a single parent, so in principle if one looks back at an arbitrary time in the past, an individual has a single ancestor.

For sexual reproduction, this question is much more complex: at first sight, when one looks backward in history, at each generation, the number of ancestors is doubled: we have 2 parents, 4 grand parents and so on. If one extrapolates 2000 years ago with a new generation every 30 years, this

This work is dedicated to Heinz Horner on the occasion of his 60th birthday.

[1] Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, F-75231 Paris 05 Cedex, France.

[2] Laboratoire Evolution et Systématique, Université Paris-Sud, F-91405 Orsay Cedex, France, and ENGREF, F-54042 Nancy Cedex, France.

0022-4715/99/0200-0277$16.00/0

would lead to 10^{20} ancestors at the time of the roman empire and about 10^{57} at Adam's time. Of course these numbers are totally unrealistic and this is because a genealogical tree is not a true tree: some branches of the tree coalesce, meaning that some ancestors at a given generation in the past share a common ancestor one generation before. (Note that if the total mass of earth was just human bodies, the total number of humans would not exceed 10^{23}).

The problem of understanding the properties of such a "tree" is easier to consider in the simpler context of the genealogical tree of a single chromosome (human beings have 23 pairs of chromosomes) or of a sequence (a part of a chromosome). Mitochondrial DNA or the Y chromosome are always inherited from a single parent and their genealogy is strictly of the asexual type.[1] For the other chromosomes of mammals, however, the genealogy is more complex: a chromosome is either inherited from a single chromosome at the previous generation or produced by the recombination between two homologous chromosomes (after a recombination event, also called crossing-over, the new chromosome consists of one part coming from one parent and the remaining part coming from the other parent, the breaking point between these two parts being more or less random along the sequence).

The simplest way to model this recombination event is to consider that a chromosome consists of a sequence of L nucleotides (or sites). When the chromosome is inherited from a single parental chromosome, it remains unchanged (if mutations are neglected). On the other hand, during a time dt, (dt is typically the time of one generation) there is a probability rdt that the chromosome transmitted results from a recombination event with a breaking between the ith site and the $i+1$th site, meaning that all the nucleotides $1 \leqslant j \leqslant i$ come from the first parental chromosome and all the nucleotides $i+1 \leqslant j \leqslant L$ come from the second parental chromosome.

Under this process, if the population was infinite and if mating between pairs of individuals in the population was done at random (this random mating is called panmixia), the number $Q(t)$ of ancestors, at time t in the past, would be simply L in the limit $t \to \infty$ (by looking far enough in the past one would always find a recombination event which would make two consecutive nucleotides belong to different ancestors). However the population is always finite. This has the effect that two individuals in a genealogical tree have sometimes the same parent and so can inherit the same chromosome. For simplicity, we will assume that the population size remains constant in time, so that the chance that two individuals have a common parent does not vary with time. To model this effect, we consider that when we go backward in time, there is a probability dt that two ancestors inherit the same chromosome from the same parent.

We see that recombination and coalescence have opposite effects: coalescence tends to decrease and recombination tends to increase the number of ancestors of a chromosome. In the long time limit, the genealogy of a given chromosome reaches a dynamical steady state where these two tendencies equilibrate.

Coalescence theory, the aim of which is to study the statistical laws of genealogies of genes or sequences has been greatly developped[2, 3] since the pioneering work of Kingman.[4] The genetic diversity of a population is closely related to the rate of coalescence: the longer it takes the lineages of two sequences to coalesce, the larger is the present genetic diversity between the two sequences. Hudson[5] was the first to consider sequences where recombination can occur randomly along the sequence. Hudson's retrospective process considers a sample of n sequences, the genealogy of which is studied backward in time until all the homologous sites of the sample have a single ancestral site. The problem we study here, namely the genealogy of a single chromosome or sequence, concerns in some sense what happens in the past beyond the end of Hudson's process. This aspect has received little attention until recently[6] because it is not directly related to the observable genetic diversity of the population at present time.

The model we study here (which is very similar to the model considered recently by Wiuf and Hein[6]) is a simple formulation of the problem of the genealogy of a chromosome (or of a sequence) as the dynamics of a spin chain of L spins. At any given time, each spin can take an arbitrary color. For example, a spin configuration of $L = 10$ spins may be

$$1\,1\,2\,2\,3\,1\,2\,3\,1\,4 \tag{1}$$

The only aspect which is relevant is the way the chain is partitionned into different colors, i.e., which spins belong to the same color. So (1) means only that sites 1, 2, 6, 9 carry the same color, sites 3, 4, 7 carry another color, sites 5, 8 a third color and site 10 a fourth color.

The dynamics is the following:

1. **coalescence**: during every infinitesimal time interval dt, there is a probability dt for any pair of colors α and β present in the system to coalesce, so that all the spins having color α or β adopt a common color. For example (1) becomes

$$1\,1\,1\,1\,2\,1\,1\,2\,1\,3 \quad \text{with probability} \quad dt \tag{2}$$

$$1\,1\,2\,2\,1\,1\,2\,1\,1\,3 \quad \text{with probability} \quad dt \tag{3}$$

$$1\,1\,2\,2\,3\,1\,2\,3\,1\,1 \quad \text{with probability} \quad dt \tag{4}$$

$$1\,1\,2\,2\,2\,1\,2\,2\,1\,3 \quad \text{with probability} \quad dt \tag{5}$$

$$1\,1\,2\,2\,3\,1\,2\,3\,1\,2 \quad \text{with probability} \quad dt \tag{6}$$

$$1\,1\,2\,2\,3\,1\,2\,3\,1\,3 \quad \text{with probability} \quad dt \tag{7}$$

2. **recombination**: during every infinitesimal time interval dt, there is a probability rdt that a recombination event occurs between site i and site $i+1$ for any color present in the system. This means that if this event occurs for color α, all the sites having color α at the left of i (including i) keep their color α and all the sites at the right of $i+1$ (including $i+1$) which had color α adopt a different color β which was not yet present in the system. So (1) becomes

$$1\,2\,3\,3\,4\,2\,3\,4\,2\,5 \quad \text{with probability} \quad rdt \tag{8}$$

$$1\,1\,2\,2\,3\,4\,2\,3\,4\,5 \quad \text{with probability} \quad 4rdt \tag{9}$$

$$1\,1\,2\,2\,3\,1\,2\,3\,4\,5 \quad \text{with probability} \quad 3rdt \tag{10}$$

$$1\,1\,2\,3\,4\,1\,3\,4\,1\,5 \quad \text{with probability} \quad rdt \tag{11}$$

$$1\,1\,2\,2\,3\,1\,4\,3\,1\,5 \quad \text{with probability} \quad 3rdt \tag{12}$$

$$1\,1\,2\,2\,3\,1\,2\,4\,1\,5 \quad \text{with probability} \quad 3rdt \tag{13}$$

We have now to explain why the dynamics of this spin chain can be used to model the genealogical tree of a chromosome. The values of the spins along the chain tell us how the nucleotides of a chromosome that we observe now were distributed among the ancestors a time t ago. For example (1) tells us that nucleotides at location 1, 2, 6 and 9 were carried by one individual a time t ago (this individual whose name has no importance is called 1), nucleotides at location 3, 4, 7 were carried by another ancestor (called 2), nucleotides at location 5, 8 by a third ancestor and the nucleotide at location 10 by a fourth ancestor.

The coalescence event between two colors α and β simply means that between time t and time $t+dt$ in the past, the two individuals α and β have inherited the same chromosome from their common parent.

The recombination event in color α between site i and $i+1$ on the other hand means that ancestor α has inherited the part at the left of site i from his first parent and the part at the right of site $i+1$ from his second parent.

In the present work, we study the steady state (i.e., long time limit) of this model. In Section 2, we define various quantities of interest, in particular the average number of ancestors and quantities which measure how the nucleotides are distributed among these ancestors. In Section 3, we show that these quantities can be computed exactly for small system sizes.

This gives also (see Section 4) exact expressions of the pair or three point correlations of systems of arbitrary size. In Section 5 we present the predictions of a simple mean field approximation and we discuss possible ways of improving it. In Section 6, we develop an expansion method valid in the scaling limit where L is large and r is of order $1/L$ so that most quantities of interest become functions of the product rL. Lastly in Section 7, we present the results of numerical simulations, which allow a comparison with the exact results obtained for small system sizes and with the mean field predictions. We also show that some global properties remain non-self-averaging even in the infinite size limit.

2. QUANTITIES OF INTEREST

The dynamics defined above for a chain of finite length L is a Markov process with a finite number of states. In the long time limit, the system reaches a steady state which does not evolve in time. Our goal is to calculate various properties in this steady state.

An important quantity characteristic of a configuration is the number n_i of sites of the chain which have the same color as site i (as site i contributes to n_i, one always has $n_i \geqslant 1$). For configuration (1), one has $n_1 = n_2 = n_6 = n_9 = 4$, $n_3 = n_4 = n_7 = 3$, $n_5 = n_8 = 2$ and $n_{10} = 1$.

If Q is the total number of different colors (that is the total number of ancestors of the chromosome), its expression in terms of the n_i is

$$Q = \sum_{i=1}^{L} \frac{1}{n_i} \tag{14}$$

and one has $Q = 4$ for configuration (1).

At least for finite L, this number of ancestor fluctuates and so one can try to describe its probability distribution; in particular one can try to calculate its average and its variance

$$\langle Q \rangle = \sum_{i=1}^{L} \left\langle \frac{1}{n_i} \right\rangle \tag{15}$$

$$\langle Q^2 \rangle - \langle Q \rangle^2 = \sum_{i=1}^{L} \sum_{j=1}^{L} \left[\left\langle \frac{1}{n_i n_j} \right\rangle - \left\langle \frac{1}{n_i} \right\rangle \left\langle \frac{1}{n_j} \right\rangle \right] \tag{16}$$

By analogy with spin glass problems,[7–10] one can define another quantity Y which measures the relative weights of the different ancestors

$$Y = \sum_{i=1}^{L} \frac{n_i}{L^2} \tag{17}$$

If W_α is the weight of ancestor α defined as the fraction of sites coming from this ancestor, it is easy to check that

$$Y = \sum_\alpha [W_\alpha]^2 \tag{18}$$

Under this form, it is clear that Y is (for each configuration) the probability that two sites (chosen at random among the L sites) belong to the same ancestor. This number is itself random as it depends on the configuration: for configuration (1), it takes the value $Y = 3/10$.

One can try to calculate the successive moments of Y, in particular to see whether as in other spin glass types of systems,[7–10] Y remains a non-self-averaging quantity even in the large L limit (a non-self-averaging quantity is by definition a quantity which fluctuates even in the thermodynamic limit). From (17), it is clear that

$$\langle Y \rangle = \frac{1}{L^2} \sum_{i=1}^{L} \langle n_i \rangle \tag{19}$$

$$\langle Y^2 \rangle - \langle Y \rangle^2 = \frac{1}{L^4} \sum_{i=1}^{L} \sum_{j=1}^{L} [\langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle] \tag{20}$$

Other global properties can be expressed in terms of correlations between sites. If τ_i is the color at site i, the number S of segments (a segment is a set of contiguous sites belonging to the same ancestor) is clearly

$$S = L - \sum_{i=1}^{L-1} \delta_{\tau_i, \tau_{i+1}} \tag{21}$$

(δ is the Kronecker symbol). If we call l_i the length of the segment to which belongs site i, one can also see that S can be rewritten as

$$S = \sum_{i=1}^{L} \frac{1}{l_i} \tag{22}$$

For configuration (1) $l_1 = l_2 = l_3 = l_4 = 2$, $l_5 = l_6 = l_7 = l_8 = l_9 = l_{10} = 1$ and $S = 8$.

With the lengths l_i, one can build a quantity Z similar to Y defined in (17)

$$Z = \sum_{i=1}^{L} \frac{l_i}{L^2} \tag{23}$$

As in (18), if W'_α is the weight of a segment α defined as the fraction of sites belonging to this segment, one can rewrite Z as

$$Z = \sum_\alpha [\, W'_\alpha]^2 \tag{24}$$

For each configuration, Z is the probability that two sites chosen at random belong to the same segment. In the example of configuration (1), one has $Z = 7/50$.

It is easy, using Jensen's inequality, to verify from (14, 17, 21, 23) that for any configuration

$$Q \geqslant \frac{1}{Y} \qquad \text{and} \qquad S \geqslant \frac{1}{Z} \tag{25}$$

Moreover, it is also rather obvious to see that

$$Q \leqslant S \qquad \text{and} \qquad Y \geqslant Z \tag{26}$$

Apart from global quantities, one can also study correlations between sites:

- the probability that sites i and j have the same color: $P_{i,j}$
- the probability that sites i, j and k have the same color: $P_{i,j,k}$
- the probability that sites i, j, k and l have the same color: $P_{i,j,k,l}$
- etc...
- the probability that sites i and j have the same color and k and l have the same color, but these two colors are different: $P_{i,j;k,l}$
- and so on.

Some moments of global properties can be calculated from the knowledge of these correlations. For example

$$\langle n_i \rangle = \sum_{j=1}^{L} P_{i,j} \tag{27}$$

$$\langle n_i n_k \rangle = \sum_{j=1}^{L} \sum_{l=1}^{L} P_{i,j,k,l} + P_{i,j;k,l} \tag{28}$$

and this gives

$$\langle Y\rangle = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} P_{i,j} \tag{29}$$

$$\langle Y^2\rangle - \langle Y\rangle^2 = \frac{1}{L^4}\sum_{i=1}^{L}\sum_{j=1}^{L}\sum_{k=1}^{L}\sum_{l=1}^{L} P_{i,j,k,l} + P_{i,j;k,l} - P_{i,j}P_{k,l} \tag{30}$$

Similarly one has

$$\langle \delta_{\tau_i,\tau_{i+1}}\rangle = P_{i,i+1} \tag{31}$$

and this implies that

$$\langle S\rangle = L - \sum_{i=1}^{L-1} P_{i,i+1} \tag{32}$$

Most global properties (like $\langle Q\rangle$), however cannot be written in terms of 2 or 4 point correlation functions but require the knowledge of high order correlation functions and this makes the calculation of such properties much more difficult.

3. SMALL SYSTEM SIZES

For small enough L, the number of possible configurations is sufficiently small to allow an exact solution of the steady state.

It is easy to check that the total number Ω_L of configurations is $\Omega_1 = 1$, $\Omega_2 = 2$, $\Omega_3 = 5$, $\Omega_4 = 15$ and to establish the following recursions

$$\Omega_L = \sum_{k=1}^{L} \omega_L(k)$$

where $\omega_L(k)$ is the number of configurations of a system of L sites having k different colors. Clearly, $\omega_L(k)$ satisfies the following recursion

$$\omega_L(k) = k\omega_{L-1}(k) + \omega_{L-1}(k-1)$$

(this recursion is in fact the recursion of Stirling numbers of second kind.[11]) This of course allows one to calculate all the Ω_L. This gives $\Omega_5 = 52$, $\Omega_6 = 203$, $\Omega_7 = 877$, $\Omega_8 = 4140$, $\Omega_9 = 21147$, $\Omega_{10} = 115975$. The dynamics of the spin chain defined in the introduction leads to a system of Ω_L linear equations for the weights of the configurations in the steady state.

For example when $L=2$, one finds that the weights w_{11} and w_{12} of the two possible configurations 11 and 12 evolve according to

$$\begin{aligned}\frac{dw_{11}}{dt} &= -rw_{11}+w_{12}\\ \frac{dw_{12}}{dt} &= rw_{11}-w_{12}\end{aligned} \tag{33}$$

and in the steady state, where the left hand side of (33) vanishes, this leads to

$$w_{11}=1-w_{12}=\frac{1}{1+r} \tag{34}$$

For $L=3$, the system of five linear equations can still be solved analytically[6] and this leads in the steady state for the weights of the five possible configurations 111, 112, 121, 122, 123 to

$$\begin{aligned}w_{111} &= \frac{3+5r}{(1+r)(1+2r)(3+2r)}\\ w_{112} &= w_{122}=\frac{3r+4r^2}{(1+r)(1+2r)(3+2r)}\\ w_{121} &= \frac{2r^2}{(1+r)(1+2r)(3+2r)}\\ w_{123} &= \frac{2r^2+4r^3}{(1+r)(1+2r)(3+2r)}\end{aligned} \tag{35}$$

A full analytic solution of the system of Ω_L equations becomes very quickly too difficult as L increases. It is however possible to determine numerically, and with an arbitrary accuracy, the weights of the Ω_L configurations for L up to 8 by solving numerically the system of Ω_L equations. The variations of $\langle Q\rangle$, $\langle Q^2\rangle-\langle Q\rangle^2$, $\langle Y\rangle$ and $\langle Y^2\rangle-\langle Y\rangle^2$ as functions of rL are shown in Figs. 1 to 4 for $2\leqslant L\leqslant 8$.

By the same procedure, it is also possible numerically to generate, for small enough L, the first terms of the expansion of all the desired quantities in powers of r. One observes that the coefficients have usually a polynomial dependence on the size L (the coefficient of r^n is essentially a polynomial of degree L^n) and so the computation of these coefficients for the very first

sizes determines them for all L. The fact that the coefficients are rationnal in L could probably be established by generalizing the calculation of Section 6 to the case of a discrete lattice.)

This leads to the following expressions, valid for all system sizes L, up to terms of order r^4 or higher.

$$\langle Q\rangle = 1 + (L-1)\,r - \frac{L^2-1}{3}\,r^2 + \frac{13L^3-12L^2-L}{54}\,r^3 + \cdots$$

$$\langle Q^2\rangle - \langle Q\rangle^2 = (L-1)\,r - \frac{2(L-1)(L+1)}{3}\,r^2 + \frac{L(L-1)(35L+11)}{54}\,r^3 + \cdots$$

$$\langle Y\rangle = 1 - \frac{L^2-1}{3L}\,r + \frac{L^2-1}{6}\,r^2 - \frac{27L^4-45L^2+18}{270L}\,r^3 + \cdots$$

$$\langle Y^2\rangle - \langle Y\rangle^2 = \frac{2L^4-2}{15L^3}\,r - \frac{7L^4-5L^2-2}{45L^2}\,r^2 + \frac{779L^6-952L^4-259L^2+432}{5670L^3}\,r^3 + \cdots$$

$$\langle Z\rangle = 1 - \frac{L^2-1}{3L}\,r + \frac{L^3+L^2-L-1}{9L}\,r^2 - \frac{23L^4+45L^3-25L^2-45L+2}{540L}\,r^3 + \cdots$$

$$\langle Z^2\rangle - \langle Z\rangle^2 = \frac{2L^4-2}{15L^3}\,r - \frac{5L^5+2L^4-5L-2}{45L^3}\,r^2 + \frac{\begin{pmatrix}346L^6+420L^5+91L^4+105L^3\\ -791L^2-525L+354\end{pmatrix}}{5670L^3}\,r^3 + \cdots$$

$$\langle S\rangle = 1 + (L-1)\,r - (L-1)\,r^2 + (L-1)\,r^3 + \cdots$$

$$\langle S^2\rangle - \langle S\rangle^2 = (L-1)\,r + \frac{L^2-9L+8}{3}\,r^2 - \frac{5L^3+12L^2-152L+135}{27}\,r^3 + \cdots$$

We see that for large L and for r of order $1/L$, a scaling regime is reached and all properties become a function of the reduced variable

$$R = Lr \tag{36}$$

$$\begin{aligned}
\langle Q \rangle &= 1 + R - \tfrac{1}{3}R^2 + \tfrac{13}{54}R^3 + O(R^4) \\
\langle Q^2 \rangle - \langle Q \rangle^2 &= R - \tfrac{2}{3}R^2 + \tfrac{35}{54}R^3 + O(R^4) \\
\langle Y \rangle &= 1 - \tfrac{1}{3}R + \tfrac{1}{6}R^2 - \tfrac{1}{10}R^3 + O(R^4) \\
\langle Y^2 \rangle - \langle Y \rangle^2 &= \tfrac{2}{15}R - \tfrac{7}{45}R^2 + \tfrac{779}{5670}R^3 + O(R^4) \\
\langle Z \rangle &= 1 - \tfrac{1}{3}R + \tfrac{1}{9}R^2 - \tfrac{23}{540}R^3 + O(R^4) \\
\langle Z^2 \rangle - \langle Z \rangle^2 &= \tfrac{2}{15}R - \tfrac{1}{9}R^2 + \tfrac{346}{5670}R^3 + O(R^4) \\
\langle S \rangle &= 1 + R + O(R^4) \\
\langle S^2 \rangle - \langle S \rangle^2 &= R + \tfrac{1}{3}R^2 - \tfrac{5}{27}R^3 + O(R^4)
\end{aligned} \tag{37}$$

These results show that even in the large L limit (keeping the product rL fixed at some arbitrary value R), all the global properties fluctuate and therefore are non-self-averaging. We will see in Section 5 that the expansions (37) can be recovered directly by considering a continuous version of the model which becomes valid in the large L limit.

4. EXACT CORRELATION FUNCTIONS AND THEIR CONSEQUENCES

An interesting aspect of the model is that pair or 3 point correlation functions[6] are easy to calculate. The reason is that one can forget all the remaining sites of the sequence.

The calculation of the pair correlation function is very similar to the calculation of the steady state of a sequence of $L = 2$ sites. The probabilities $P_{i,j}$ that sites i and j have the same color and $P_{i;j} = 1 - P_{i,j}$ that sites i and j have different colors evolve according to

$$\frac{d}{dt} P_{i,j} = -(r\,|j-i|)\, P_{i,j} + 1 - P_{i,j}$$

and in the steady state, this leads to

$$P_{i,j} = \frac{1}{1 + r\,|j-i|}$$

Similarly the calculation of three point functions reduces to solving a system of 5 linear equations (similar to those of the steady state of a sequence of $L=3$ sites) and one finds that for $i \leqslant j \leqslant k$

$$P_{i,j,k} = \frac{3+4(r_1+r_2)+r_1^2+3r_1r_2+r_2^2}{(1+r_1)(1+r_2)(1+r_1+r_2)(3+r_1+r_2)}$$

$$P_{i,j;k} = \frac{r_2(3+3r_1+4r_2+(r_1+r_2)^2)}{(1+r_1)(1+r_2)(1+r_1+r_2)(3+r_1+r_2)}$$

$$P_{i,k;j} = \frac{r_1r_2(2+r_1+r_2)}{(1+r_1)(1+r_2)(1+r_1+r_2)(3+r_1+r_2)}$$

$$P_{i;j,k} = \frac{r_1(3+4r_1+3r_2+(r_1+r_2)^2)}{(1+r_1)(1+r_2)(1+r_1+r_2)(3+r_1+r_2)}$$

$$P_{i;j;k} = \frac{r_1r_2(1+r_1+r_2)(2+r_1+r_2)}{(1+r_1)(1+r_2)(1+r_1+r_2)(3+r_1+r_2)}$$

where

$$r_1 = (j-i)\, r$$

$$r_2 = (k-j)\, r$$

In principle the calculation of four point functions or higher correlations can be done in a similar way but the size of the linear system increases very quickly (for the four point function, one has a system of $\Omega_4 = 15$ equations).

Using (29) and (32), one can obtain exact expressions of moments of some global properties:

$$\langle Y \rangle = \frac{1}{L^2} \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{1}{1+r\,|i-j|} \tag{38}$$

$$\langle S \rangle = \frac{1+Lr}{1+r} \tag{39}$$

and in the scaling limit $L \to \infty$ keeping $Lr = R$, this leads to the following exact expressions

$$\langle Y \rangle_{\text{exact}} = 2\,\frac{(1+R)\log(1+R)-R}{R^2} \tag{40}$$

and

$$\langle S \rangle_{\text{exact}} = 1 + R \tag{41}$$

in agreement with the expansion results (37). The exact result (41) was derived in ref. 6 where an expression for the variance of S was given ($\langle S^2 \rangle - \langle S \rangle^2 \simeq 2R$) which disagrees with (37), the disagreement being due to the fact that (37) is valid for small R whereas the expression of ref. 6 seems to be valid for large R.

5. MEAN-FIELD APPROXIMATION

In absence of an exact solution, most properties can only be estimated through approximate methods. We have seen in (14, 15) that the exact expression of $\langle Q \rangle$ requires the knowledge of $\langle 1/n_i \rangle$.

Unfortunately, it is very hard to calculate $\langle 1/n_i \rangle$. What is much easier to obtain (see Section 4), however, is $\langle n_i \rangle$

$$\langle n_i \rangle = \sum_{j=1}^{N} \frac{1}{1 + r\,|j-i|} \tag{42}$$

The simplest mean field approximation consists in replacing $\langle 1/n_i \rangle$ by $1/\langle n_i \rangle$ leading to

$$\langle Q \rangle_{\text{meanfield}} = \sum_{i=1}^{L} \frac{1}{\langle n_i \rangle}$$

Because of the fact that

$$\left\langle \frac{1}{n_i} \right\rangle \geqslant \frac{1}{\langle n_i \rangle}$$

$\langle Q \rangle_{\text{meanfield}}$ gives always a lower bound

$$\langle Q \rangle_{\text{meanfield}} \leqslant \langle Q \rangle_{\text{exact}}$$

For large L and small r with $Lr = R$ of order 1, one has

$$\langle n_i \rangle \simeq \frac{\log(1 + ri) + \log(1 + r(L-i))}{r}$$

and this leads to

$$\langle Q \rangle_{\text{meanfield}} \simeq \int_0^R \frac{dy}{\log(1+y)+\log(1+R-y)} \tag{43}$$

which leads to the expansion

$$\langle Q \rangle_{\text{meanfield}} = 1 + \frac{R}{3} - \frac{3R^2}{10} + O(R^3)$$

which disagrees with the exact expansion (37) showing the limitations of the mean field approach to predict some global properties. However, from (17, 19), as $\langle n_i \rangle$ is known exactly, the mean field expression $\langle Y \rangle_{\text{mean field}}$ coincides with the exact one given in (40)

$$\langle Y \rangle_{\text{exact}} = \frac{1}{L^2} \sum_{i=1}^{L} \langle n_i \rangle$$

A priori, with more and more efforts, one could try to calculate in addition to pair and 3 point correlations, higher correlations. For example, the knowledge of four point correlations would lead to exact expressions of $\langle n_i n_j \rangle$ (and as $\langle n_i^2 \rangle$ can be computed from 3 point functions) and this would give the exact expression of $\langle Y^2 \rangle$.

The knowledge of higher correlations and therefore of higher moments or correlations of the n_i could also be used to develop improved mean field approximations. When only $\langle n_i \rangle$ is known, one can just approximate $\langle 1/n_i \rangle$ by $1/\langle n_i \rangle$ (and as mentioned above one knows that this approximation is a lower bound). Knowing a few higher moments of the n_i could be used to obtain better approximations as well as improved (upper and lower) bounds (by looking for the distributions of integer n_i compatible with the known moments and which minimize or maximize $\langle 1/n_i \rangle$).

6. CONTINUOUS THEORY

When r is very small at fixed L, the only configuration which has a non negligible weight is the configuration where all sites have the same color. If one takes the limit $r \to 0$ and $L \to \infty$ keeping the product

$$R = Lr$$

fixed all configurations with a finite number of colors contribute. If in addition the product $R = Lr$ is small, the weight of configurations with k

segments is of order R^{k-1}, and therefore to calculate any quantity in powers of R up to order R^k, one can ignore the contribution of all configurations with $k+2$ segments or more. In this section, we show how the calculations can be done up to order R^3. Let us denote the probability of all configurations with at most 4 segments by

$$A = \text{Prob}(1)$$

$$B(x) = \text{Prob}(1\,|\,2)$$

$$C(x,\, y) = \text{Prob}(1\,|\,2\,|\,3)$$

$$D(x,\, y) = \text{Prob}(1\,|\,2\,|\,1)$$

$$E(x,\, y,\, z) = \text{Prob}(1\,|\,2\,|\,3\,|\,4)$$

$$F(x,\, y,\, z) = \text{Prob}(1\,|\,2\,|\,1\,|\,3)$$

$$G(x,\, y,\, z) = \text{Prob}(1\,|\,2\,|\,3\,|\,1)$$

$$H(x,\, y,\, z) = \text{Prob}(1\,|\,2\,|\,3\,|\,2)$$

$$I(x,\, y,\, z) = \text{Prob}(1\,|\,2\,|\,1\,|\,2)$$

These notations simply mean that A is the probability of the configuration with all spins being 1, $B(x)\,dx$ is the probability of all configurations with all sites being 1 on the first segment and all sites being 2 on the second segment, the breaking point being located between Lx and $L(x+dx)$, $C(x,\, y)\,dx\,dy$ is the probability of all configurations with 3 segments of colors 1, 2 and 3, the first breaking point being located between poisitions Lx and $L(x+dx)$ and the second breaking point between Ly and $L(y+dy)$ and so on.

Then if one ignores all configurations of 5 or more segments, one finds that these probabilities evolve according to

$$\frac{dA}{dt} = -RA + \int_0^1 dx B(x) + \int_0^1 dy \int_0^y dx D(x,\, y)$$

$$+ \int_0^1 dz \int_0^z dy \int_0^y dx I(x,\, y,\, z)$$

$$\frac{dB(x)}{dt} = -(R+1)\,B(x) + RA + \int_0^x dz C(z,\, x) + + \int_x^1 dz C(x,\, z)$$

$$+ \int_0^x dy \int_0^y dz F(z,\, y,\, x) + \int_x^1 dy \int_y^1 dz H(x,\, y,\, z)$$

$$
\begin{aligned}
\frac{dC(x,\,y)}{dt} &= -(R+3)\,C(x,\,y) + RB(x) + RB(y) + R(y-x)\,D(x,\,y) \\
&\quad + \int_0^x dz E(z,\,x,\,y) + \int_x^y dz E(x,\,z,\,y) + \int_z^1 dz E(x,\,y,\,z) \\
\frac{dD(x,\,y)}{dt} &= -(R + R(y-x) + 1)\,D(x,\,y) + C(x,\,y) + \int_y^1 dz F(x,\,y,\,z) \\
&\quad + \int_0^x dz G(z,\,x,\,y) + \int_x^y dz G(x,\,z,\,y) \\
&\quad + \int_y^1 dz G(x,\,y,\,z) + \int_0^x dz H(z,\,x,\,y) \\
\frac{dE(x,\,y,\,z)}{dt} &= -6E(x,\,y,\,z) + RC(x,\,y) + RC(x,\,z) + RC(y,\,z) \\
\frac{dF(x,\,y,\,z)}{dt} &= -3F(x,\,y,\,z) + E(x,\,y,\,z) + RD(x,\,y) \\
\frac{dG(x,\,y,\,z)}{dt} &= -3G(x,\,y,\,z) + E(x,\,y,\,z) + RD(x,\,z) \\
\frac{dH(x,\,y,\,z)}{dt} &= -3H(x,\,y,\,z) + E(x,\,y,\,z) + RD(y,\,z) \\
\frac{dI(x,\,y,\,z)}{dt} &= -I(x,\,y,\,z) + F(x,\,y,\,z) + H(x,\,y,\,z)
\end{aligned}
\tag{44}
$$

The steady state solution is then

$$
\begin{aligned}
A &= 1 - R + \tfrac{2}{3}R^2 - \tfrac{23}{54}R^3 + \cdots \\
B(x) &= R - \tfrac{4}{3}R^2 + \tfrac{5}{18}R^3(x^2 + (1-x)^2) + R^3 + \cdots \\
C(x,\,y) &= \tfrac{2}{3}R^2 + \tfrac{2}{9}R^3(y-x) - R^3 + \cdots \\
D(x,\,y) &= \tfrac{2}{3}R^2 - \tfrac{7}{9}R^3(y-x) - R^3 + \cdots \\
E(x,\,y,\,z) &= \tfrac{1}{3}R^3 + \cdots \\
F(x,\,y,\,z) &= \tfrac{1}{3}R^3 + \cdots \\
G(x,\,y,\,z) &= \tfrac{1}{3}R^3 + \cdots \\
H(x,\,y,\,z) &= \tfrac{1}{3}R^3 + \cdots \\
I(x,\,y,\,z) &= \tfrac{2}{3}R^3 + \cdots
\end{aligned}
$$

From these weights, one can calculate all properties in power series of R (of course up to the order R^3). For example

$$\langle Q\rangle = A + 2\left[\int_0^1 B(x)\,dx + \int_0^1 dx \int_x^1 dy\, D(x, y) + \int_0^1 dx \int_x^1 dy \int_y^1 dz\, I(x, y, z)\right] + 3\left[\int_0^1 dx \int_x^1 dy\, C(x, y) + \int_0^1 dx \int_x^1 dy \int_y^1 dz (F(x, y, z) + G(x, y, z) + H(x, y, z))\right] + 4\int_0^1 dx \int_x^1 dy \int_y^1 dz\, E(x, y, z) = 1 + R - \tfrac{1}{3}R^2 + \tfrac{13}{54}R^3 + O(R^4)$$

and one recovers this way the expressions (37) guessed in Section 3.

7. SIMULATIONS

The dynamics of the spin chain defined in Section 1 is easy to simulate by a Monte Carlo method for rather large L. The results obtained for $L = 5, 20, 80, 320, 1280$ after 10^7 updates are shown in Figs. 1–4. We see that as L increases, all these quantities become a function of the product

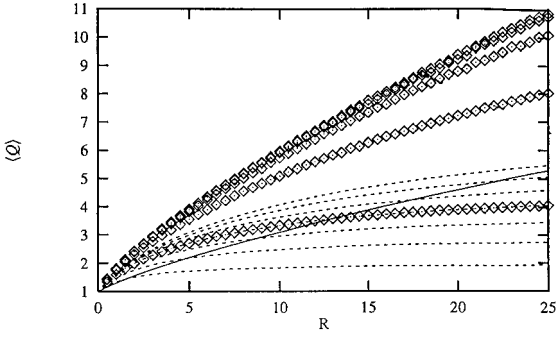

Fig. 1. Figure 1 shows the average number of ancestors (in the steady state) $\langle Q\rangle$ as a function of $R = rL$. The dashed lines represent exact calculations for small system sizes $2 \leqslant L \leqslant 8$ (the L dependence is monotonous and $L = 2$ is the lower curve). The diamonds represent the results of the Monte Carlo calculation for $L = 5, 20, 80, 320, 1280$. Lastly the plain line represents the mean field prediction (43).

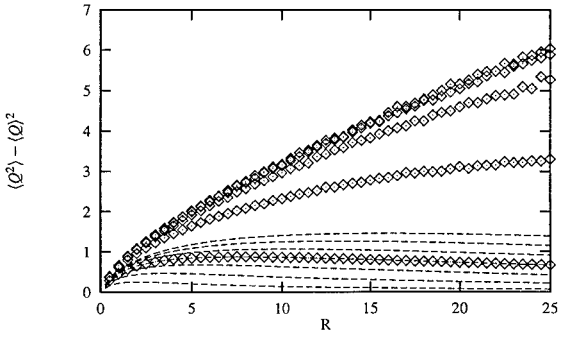

Fig. 2. Figure 2 shows the variance $\langle Q \rangle^2 - \langle Q \rangle^2$ as a function of $R = rL$. The dashed lines represent exact calculations for small system sizes $2 \leqslant L \leqslant 8$ (the L dependence is monotonous and $L = 2$ is the lower curve). The diamonds represent the results of the Monte Carlo calculation for $L = 5, 20, 80, 320, 1280$.

$rL = R$ as expected when R is of order 1 (see Sections 3 and 5). For $L = 5$, the agreement with the exact result of Section 3 is perfect and this gives confidence in the Monte Carlo procedure.

When compared to the mean field predictions in Figs. 1 and 3, the Monte Carlo results agree very well with the mean field prediction of $\langle Y \rangle$, as expected since for $\langle Y \rangle$ the mean field theory is exact. On the other hand, for $\langle Q \rangle$, the mean field prediction is clearly unaccurate and therefore better approximation schemes are needed.

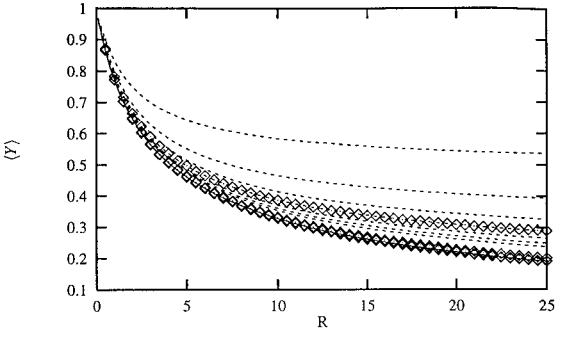

Fig. 3. Figure 3 shows $\langle Y \rangle$ as a function of $R = rL$. The dashed lines represent exact calculations for small system sizes $2 \leqslant L \leqslant 8$ (the L dependence is monotonous and $L = 2$ is the lower curve). The diamonds represent the results of the Monte Carlo calculation for $L = 5, 20, 80, 320, 1280$. Lastly the plain line, hardly visible because it coincides with the Monte Carlo data for the largest sizes, represents the mean field prediction (40) which is exact for $L \to \infty$.

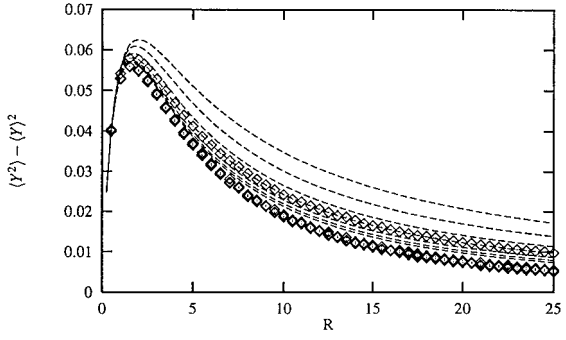

Fig. 4. Figure 4 shows the variance $\langle Y \rangle^2 - \langle Y \rangle^2$ as a function of $R = rL$. The dashed lines represent exact calculations for small system sizes $2 \leqslant L \leqslant 8$ (the L dependence is monotonous and $L = 2$ is the lower curve). The diamonds represent the results of the Monte Carlo calculation for $L = 5, 20, 80, 320, 1280$.

In Monte Carlo simulations, it is of course easy to measure with good statistics a large number of properties. As we know from (37) that Y defined by (17) is non-self-averaging, one can try to measure its distribution $\Pi(Y)$. When L becomes large, (here we choose $L = 1000$), we have measured the distribution $\Pi(Y)$ for several values of the product $rL = R$. Figures 5–8 show the results of our simulations for $R = 1, 4, 7$ and 10. Clearly, Y is a non-self-averaging quantity and its distribution $\Pi(Y)$ exhibits, as in many other systems (spin glasses, random maps, random trees...)[8, 9, 12, 13] clear singularities at $Y = 1/2$, $Y = 1/3$ on top of a very large peak at $Y = 1$ not shown on the figures to keep the rest of the distribution visible. Therefore, one expects as in many of these other systems,[9, 14] the distribution of Y to be singular at all values $Y = 1/n$.

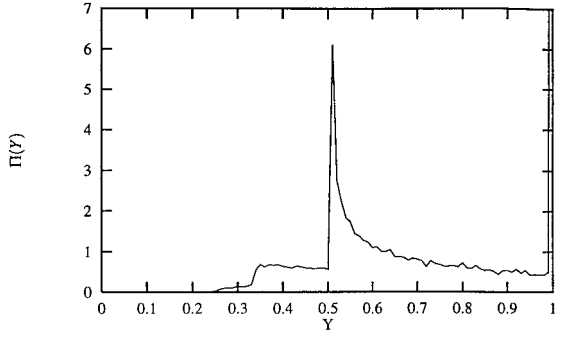

Fig. 5. Figure 5 shows the distribution $\Pi(Y)$ for $R = 1$ for a sample of 10^5 values of Y.

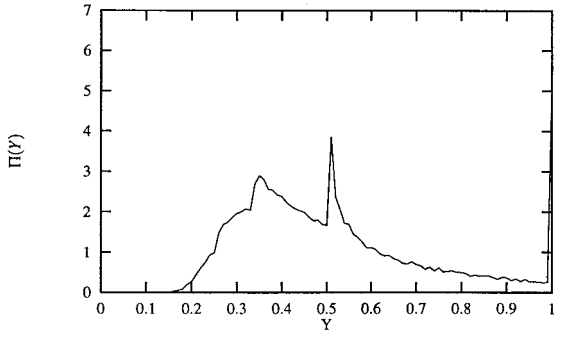

Fig. 6. Figure 6 shows the distribution $\Pi(Y)$ for $R=4$ for a sample of 10^5 values of Y.

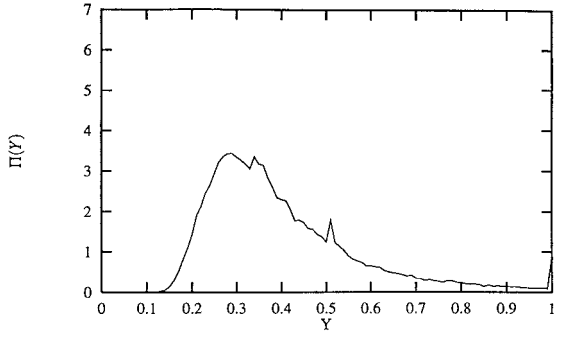

Fig. 7. Figure 7 shows the distribution $\Pi(Y)$ for $R=7$ for a sample of 10^5 values of Y.

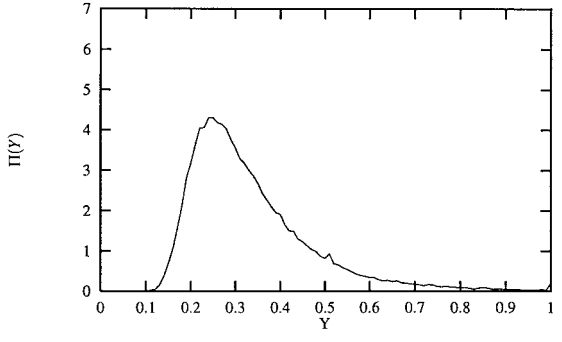

Fig. 8. Figure 8 shows the distribution $\Pi(Y)$ for $R=10$ for a sample of 10^5 values of Y.

8. CONCLUSION

In this paper we have seen how a simple model combining the effects of coalescence and recombination (without mutations) can be formulated as a dynamical spin chain.

From the point of view of statistical physics, this problem is interesting because some quantities like correlations involving a small number of sites can be calculated exactly, although most global properties like the distribution of the number of ancestors are difficult to obtain. It poses the important question of finding the optimal approximate schemes consistent with the correlations which are calculated exactly. We have seen in Section 5 that the average $\langle n_i \rangle$ is easy to calculate from the known pair correlation. The calculation of higher moments of n_i is not impossible as it reduces to the calculation of higher correlations (finding the correlations between k points is equivalent to solving a system of Ω_k linear equations with $\Omega_2 = 2$, $\Omega_3 = 5$, $\Omega_4 = 15$, $\Omega_5 = 52$...). If we knew a few higher moments of n_i, one could improve the mean field predictions of $\langle Q \rangle$. One could for example estimate $\langle Q \rangle$ using a maximum entropy principle consistent with the known moments. With more moments of the n_i, one could also, as mentioned earlier, obtain improved bounds for $\langle Q \rangle$.

An interesting aspect of the model is that in the whole range where $R = rL$ is of order 1, most global properties (Q, Y,...) defined in Section 2 exhibit non-self-averaging effects reminiscent of spin glasses and random genealogical trees.[8, 10, 12]

The model discussed here is of course an extreme simplification of the biological reality[6] where there is no selection and the population has a constant size and no structure. Moreover, the mechanism of recombination is usually more complex than totally random along the sequence (with interferences),[15, 16] with the possibility of more than one recombination at each generation.[17]

The parameter r is the ratio between the rate of recombination between two adjacent sites and the rate of coalescence between two lineages. One can estimate $r \sim 10^{-8} N$ for a population of N individuals[5] and the number L of nucleotides of a human chromosome is of order 10^8.[6] The product $R = rL$ is therefore usually very large. This means that our results, valid for R of order 1, could be only relevant for rather small populations or rather short sequences.

The number of ancestors Q is clearly an important parameter from a biological point of view. However it does not describe the full reality, in particular it does not tell us how equally the nucleotides are distributed among the ancestors. In this work, we have proposed to consider other quantities, in particular Y coming from spin glass theory, which give a

measure of the repartition of the nucleotides among the ancestors. Interesting enough, $\langle Y \rangle$ can be calculated exactly.

ACKNOWLEDGMENTS

We thank F. Austerlitz, A. Franc, P. H. Gouyon, and E. Klein for many useful discussions.

REFERENCES

1. R. L. Cann, M. Stoneking, and A. C. Wilson, Mitochondrial DNA and human evolution, *Nature* **325**:31 (1987).
2. R. Hudson, Gene genealogies and the coalescent process, *Oxford Surveys in Evolutionnary Biology*, Vol. 7, p. 1, D. Futuyma and J. Antonovics, eds. (Oxford University Press, 1991).
3. P. Donnelly and S. Tavaré, Coalescents and genealogical structure under neutrality, *Annu. Rev. Genet.* **29**:401 (1995).
4. J. F. Kingman, The coalescent, *Stoch. Proc. Appl.* **13**:235 (1982).
5. R. Hudson, Properties of a neutral allel model with intragenic recombination, *Theor. Pop. Biol.* **23**:183 (1983).
6. C. Wiuf and J. Hein, On the number of ancestors to a DNA sequence, *Genetics* **147**:1459 (1997).
7. M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, and M. Virasoro, Replica symmetry breaking and the nature of the spin glass phase, *J. Physique* **45**:843 (1984).
8. M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, 1987).
9. B. Derrida and H. Flyvbjerg, Statistical properties of randomly broken objects and of multivalley structures in disordered systems, *J. Phys. A* **20**:5273 (1987).
10. B. Derrida, From random walks to spin glasses, *Physica D* **107**:186 (1997).
11. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).
12. B. Derrida and L. Peliti, Evolution in a flat fitness landscape, *Bull. Math. Biol.* **53**:355 (1990).
13. P. G. Higgs, Frequency distributions in population genetics parallel those in statistical physics, *Phys. Rev. E* **51**:95 (1995).
14. L. Frachebourg, I. Ispolatov, and P. L. Krapivsky, Extremal properties of random systems, *Phys. Rev. E* **52**:R5727 (1995).
15. M. S. McPeek and T. P Speed, Modeling interference in genetic recombination, *Genetics* **139**:1031 (1995).
16. G. S Roeder, Sex and the single cell, *Proc. Natl. Acad. Sci. USA* **92**:10450 (1995).
17. S. Ohno, The Maltusian parameter of ascents: what prevents the exponential increase of one's ancestors? *Proc. Natl. Acad. Sci. USA* **93**:15276 (1996).