Open access • Journal Article • DOI:10.1209/EPL/I2006-10224-4

Noisy traveling waves: Effect of selection on genealogies — Source link

Éric Brunet, Bernard Derrida, Alfred H. Mueller, Stéphane Munier

Institutions: École Normale Supérieure, Pierre-and-Marie-Curie University, Columbia University, École Polytechnique

Published on: 01 Oct 2006 - EPL (IOP Publishing)

Topics: Population

Noisy traveling waves: Effect of selection on genealogies

E Brunet, Bernard Derrida, A. Mueller, S Munier

EUROPHYSICS LETTERS 1 October 2006

Europhys. Lett., **76** (1), pp. 1–7 (2006)
DOI: 10.1209/epl/i2006-10224-4

Noisy traveling waves: Effect of selection on genealogies

E. Brunet[1], B. Derrida[1], A. H. Mueller[2] and S. Munier[3]

[1] *Laboratoire de Physique Statistique, École Normale Supérieure 24 rue Lhomond, 75231 Paris cedex 05, France*
[2] *Department of Physics, Columbia University - New York, NY 10027, USA*
[3] *Centre de Physique Théorique, Unité mixte de recherche du CNRS (UMR 7644) École Polytechnique - 91128 Palaiseau, France*

received 5 July 2006; accepted in final form 2 August 2006
published online 1 September 2006

PACS. 02.50.-r – Probability theory, stochastic processes, and statistics.
PACS. 05.40.-a – Fluctuation phenomena, random processes, noise, and Brownian motion.
PACS. 89.75.Hc – Networks and genealogical trees.

Abstract. – For a family of models of evolving population under selection, which can be described by noisy traveling-wave equations, the coalescence times along the genealogical tree scale like $\ln^\alpha N$, where N is the size of the population, in contrast with neutral models for which they scale like N. An argument relating this time scale to the diffusion constant of the noisy traveling wave leads to a prediction for α which agrees with our simulations. An exactly soluble case gives trees with statistics identical to those predicted for mean-field spin glasses by Parisi's theory.

Traveling-wave equations such as the F-KPP equation $\partial_t h = \partial_x^2 h + h - h^2$ [1–3] describe how a stable medium $h = 1$ invades an unstable medium $h = 0$. They were first introduced to study how an advantageous gene propagates through a population, h being the fraction of the population with the advantageous gene. They also appear in other contexts such as disordered systems [4–6], QCD [7], reaction-diffusion [8,9], fragmentation [10] or chemistry [11].

Traveling-wave equations often represent a mean-field picture where the fluctuations at the microscopic scale are ignored. The effect of these fluctuations can be represented [9,11–14] by a noise term ($\partial_t h = \partial_x^2 h + h - h^2 + \epsilon\eta(x,t)\sqrt{h-h^2}$, where $\eta(x,t)$ is a white noise). Determining quantitatively the effect of such a weak noise ($\epsilon \ll 1$) on the front position is a subject of active research. There is increasing evidence that the dynamics of the position of the front is dominated by the fluctuations near its tip [15–17] and that there is a shift in the velocity of the front [8,18–22], logarithmic in the amplitude ϵ of the noise, as predicted by a simple cut-off theory [23].

In the present letter, we consider models of an evolving population under selection, which can be described by noisy traveling-wave equations. Instead of focusing on the time dependence of the position of the front, we look at the problem from a different perspective: we determine how coalescence times in the genealogy depend on the size of the population. Our simulations as well as a simple argument indicate that these coalescence times scale as the inverse of the diffusion constant of the front.

We consider a population of fixed size N with asexual reproduction. Each individual i is characterized by a real number x_i measuring its adequacy to the environment, and the population is fully specified by these N positions x_i's on the adequacy axis. At each new generation, all the individuals disappear after giving birth to some offsprings. We consider the following variants:

In *Model A*, each individual gives birth to k offsprings, and the j-th offspring of individual i is at position $x_i + \epsilon_{i,j}$, where the $\epsilon_{i,j}$ are uncorrelated random numbers chosen according to some distribution $\rho(\epsilon)$. Thus, each individual inherits its parent's adequacy, and $\epsilon_{i,j}$ accounts for the effects of mutation. Then comes the selection step: out of the kN new individuals, we only keep the N best ones, the ones with the highest x_i's. Typically, we will take $k = 2$ and $\rho(\epsilon)$ uniform between 0 and 1. A similar model was proposed recently [16, 24–26] to study the evolution under competitive selection of a population of DNA molecules *in vitro*. The population undergoes several cycles where, in the first part of a cycle, each molecule is amplified by a fixed number k with possible mutations and in the second part of the cycle selection acts by keeping only the best $1/k$ fraction of these offsprings, defined as the molecules with the highest binding energies to a given target. In this picture, x_i represents this binding energy.

In Model A, selection is strict: we keep at each generation the best N individuals among the kN offsprings, without any randomness. To check whether our results remain valid under a less stringent selection, we also investigate *Model A′*: instead of keeping the best N offsprings, we choose the N surviving individuals uniformly among the best $3N/2$ offsprings, thus making the influence of the x_i's weaker in the selection process. We will see below that Models A and A′ have very similar behaviours.

In *Model B*, each individual i has infinitely many offsprings, with positions distributed according to a Poisson point process of density $\psi(x - x_i)$ (*i.e.*, with probability $\psi(x - x_i)\,\mathrm{d}x$, there is an offspring of individual i at position x). As in Model A, we only keep the N best offsprings. Here, $\psi(\epsilon)$ is a positive function such that $\int \psi(\epsilon)\mathrm{d}\epsilon = \infty$ (for the population to never disappear) and which decays fast enough as $\epsilon \to \infty$ to ensure that these best offsprings have finite positions. Having infinitely many offsprings at the first step is of course unrealistic, but after selection, each individual has only a finite number of offsprings. The two main reasons for considering Model B are to check the robustness of our results and to exhibit an exactly soluble case for one particular $\psi(\epsilon)$.

The genealogical tree of an evolving population can be characterized in many ways [27, 28]. Here we measure average coalescence times $\langle T_p \rangle$ defined as follows: T_p is the age of the most recent common ancestor of p individuals chosen at random at generation g, and $\langle T_p \rangle$ is the average of T_p over all choices of these p individuals and over all generations g.

In the absence of selection (for example, when each individual has k offsprings as in model A, but with N survivors chosen uniformly among these kN offsprings), these $\langle T_p \rangle$ grow linearly with N and their ratios take, for large N, the simple values (independent of k) of the Kingman coalescent [29, 30]:

$$\langle T_p \rangle \sim N, \qquad \frac{\langle T_3 \rangle}{\langle T_2 \rangle} \to \frac{4}{3}, \qquad \frac{\langle T_4 \rangle}{\langle T_2 \rangle} \to \frac{3}{2}, \qquad \frac{\langle T_p \rangle}{\langle T_2 \rangle} \to 2 - \frac{2}{p}. \tag{1}$$

One goal of the present work is to show that the effect of selection changes completely eq. (1): the time scale of these coalescence times $\langle T_p \rangle$ becomes

$$\langle T_p \rangle \sim \left[\ln N \right]^{\alpha} \tag{2}$$

and the ratios are compatible with the values characterizing the Bolthausen-Sznitman coales-

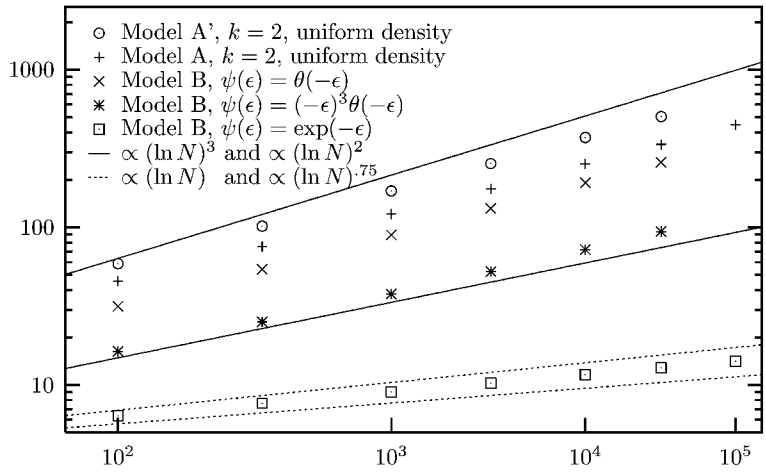

Fig. 1 – Times $\langle T_2 \rangle$ *vs.* N, for different models. The scale of the N axis is $\ln \ln N$. The data points are compared to several power laws of $\ln N$ shown by the straight lines.

cent [28, 31, 32]:

$$\frac{\langle T_3 \rangle}{\langle T_2 \rangle} \simeq \frac{5}{4}, \qquad \frac{\langle T_4 \rangle}{\langle T_2 \rangle} \simeq \frac{25}{18}. \tag{3}$$

In figs. 1 and 2, we show the results of simulations for different cases: Models A and A′ with $k = 2$ and a uniform density $\rho(\epsilon) = 1$ for $0 < \epsilon < 1$, and Model B for three different choices of the density $\psi(\epsilon)$: $\psi_1(\epsilon) = \theta(-\epsilon)$, $\psi_2(\epsilon) = (-\epsilon)^3\theta(-\epsilon)$ and $\psi_3(\epsilon) = e^{-\epsilon}$.

Typically we simulated populations of sizes ranging from $N = 10^2$ to 10^5 over 10^7 generations. We measured the times $\langle T_p \rangle$ by recording at each generation g the age $T_2(i, j)$ of the most recent common ancestor of individuals i and j. One then gets $\langle T_2 \rangle$ by averaging $T_2(i, j)$

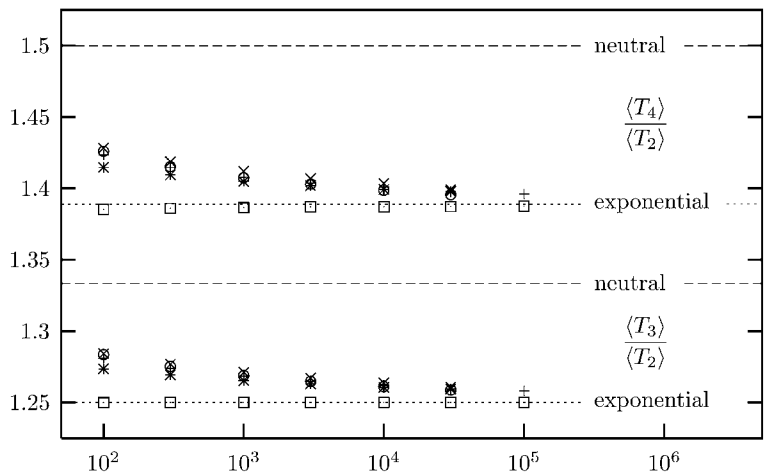

Fig. 2 – Ratios $\langle T_3 \rangle / \langle T_2 \rangle$ and $\langle T_4 \rangle / \langle T_2 \rangle$ *vs.* N, for the same models (same symbols) as in fig. 1. The dashed lines represent the neutral case eq. (1), and the dotted lines correspond to eq. (3), *i.e.* Model B with $\psi(\epsilon) = e^{-\epsilon}$.

over i, j and g. As the matrix $T_2(i,j)$ is ultrametric, no additional information is needed to compute the $\langle T_p \rangle$: for instance, $T_3(i,j,k) = \max[T_2(i,j), T_2(i,k)]$. For large sizes N, we actually took advantage of ultrametricity by representing the matrix $T_2(i,j)$ as a tree: at each step, we only kept track of the current N individuals and of all the most recent common ancestors of any pair of them. There are at most $N-1$ such ancestors, so both memory and execution time grow linearly with N, instead of N^2 if we were manipulating the full matrix.

In all cases except for Model B with the exponential distribution, which is special as we shall see below, fig. 1 indicates that the exponent α defined in eq. (2) is in the range

$$2 \leq \alpha_{\text{measured}} \leq 3. \tag{4}$$

For Model B with the exponential distribution, however, our data suggest a significantly smaller value $0.75 \leq \alpha \leq 1$.

Figure 2 shows the ratios $\langle T_3 \rangle / \langle T_2 \rangle$ and $\langle T_4 \rangle / \langle T_2 \rangle$ for the same models as in fig. 1. In all cases, including Model B with $\psi(\epsilon) = e^{-\epsilon}$, these ratios take for large N values similar to eq. (3) which differ noticeably from their values (1) in the absence of selection. At present, we do not have a general theory to explain these numerical results.

Only for Model B with $\psi(\epsilon) = e^{-\epsilon}$ (the special case), can one calculate these ratios exactly using a method similar to [5]. If at generation g, the population consists of N individuals $x_1(g), \ldots, x_N(g)$, the probability of having one of their offsprings in the interval $y, y + \mathrm{d}y$ is

$$\sum_{i=1}^{N} \psi\big(y - x_i(g)\big)\ \mathrm{d}y = \sum_{i=1}^{N} e^{x_i(g)-y}\ \mathrm{d}y = e^{X_g - y}\ \mathrm{d}y,$$

where

$$X_g = \ln\left[e^{x_1(g)} + e^{x_2(g)} + \cdots + e^{x_N(g)}\right].$$

So, for Model B with $\psi(\epsilon) = e^{-\epsilon}$, the offsprings of the whole population are the same as those of a single effective individual at position X_g. This means that one can write $x_i(g+1) = X_g + y_i$, where $y_1, y_2, ..., y_N$ are the N largest values of a Poisson point process on the line with exponential density. The y_i's are therefore distributed according to

$$\mathrm{P}(y_N < y_{N-1} < \cdots < y_1) = e^{-(y_1+y_2+\cdots+y_N)-e^{-y_N}}. \tag{5}$$

With this simplification, one can see that the differences $\Delta X_g = X_{g+1} - X_g$ are independent identically distributed random variables. At generation g, the N numbers $x_i(g)$ form a cloud of points which does not spread in time, very much like a quantum N-particle bound state. This cloud has a well-defined velocity v_N and diffusion constant D_N. As the differences ΔX_g are independent, v_N and D_N are given by

$$v_N = \langle \Delta X_g \rangle, \qquad D_N = \langle [\Delta X_g]^2 \rangle - \langle \Delta X_g \rangle^2,$$

where the expectations are over the distribution (5) of the y_i's and

$$\Delta X_g = X_{g+1} - X_g = \ln\left[e^{y_1} + ... + e^{y_n}\right].$$

Calculations similar to those of [5] lead for large N to

$$v_N = \ln\ln N + \frac{\ln\ln N + 1}{\ln N} + \cdots, \qquad D_N = \frac{\pi^2}{3\ln N} + \cdots. \tag{6}$$

We now turn to computing $\langle T_p \rangle$ in the exponential case. If one chooses randomly $p \geq 2$ individuals at generation $g+1$, one can calculate the probability q_p that they have the same ancestor at generation g:

$$q_p = \left\langle \frac{\sum_i e^{py_i}}{[\sum_i e^{y_i}]^p} \right\rangle \simeq \frac{1}{p-1}\frac{1}{\ln N} + \cdots . \tag{7}$$

One can also show that for large N and fixed p, events with more than one coalescence within the p individuals between two successive generations have a probability of order $(\ln N)^{-2}$ at most. Thus, for large N, the genealogical tree of a sample of p individuals consists of single coalescence events separated by times of order $\ln N$.

From the knowledge of the q_p's, one can obtain for large N the probability $r_p(k)$ that p individuals at generation $g+1$ have exactly $k < p$ ancestors at generation g:

$$\begin{aligned} r_p(k) &= \sum_{j=0}^{k-1} \frac{(-)^{j-k+1} p!}{j!(k-1-j)!(p-k+1)!} q_{p-j} \\ &= \frac{p}{(p-k)(p-k+1)}\frac{1}{\ln N} + \cdots \end{aligned} \tag{8}$$

and one has

$$\langle T_p \rangle = 1 + \langle T_p \rangle + \sum_{k<p} r_p(k)\ [\langle T_k \rangle - \langle T_p \rangle]. \tag{9}$$

Using the fact that $\langle T_1 \rangle = 0$ this immediately gives

$$\langle T_2 \rangle \simeq \ln N \tag{10}$$

(in reasonable agreement with our simulations of fig. 1) and all the ratios $\langle T_n \rangle / \langle T_2 \rangle$. For $n = 3$ and 4, this gives eq. (3) which is therefore asymptotically exact in the exponential case.

We now return to the general case. Our models are branching processes which are known to be related to fronts of the F-KPP type [33]. Let us now see how one can associate to Model B a noisy traveling-wave equation (a similar calculation can be done for Model A). At generation g the whole population can be characterized by a function $h_g(x)$ which counts the fraction of individuals i such that $x_i(g) > x$. Obviously, $h_g(x)$ has the shape of a front ($h_g(-\infty) = 1$ and $h_g(\infty) = 0$). From the definition of Model B, one can show that $h_g(x)$ satisfies

$$h_{g+1}(x) = \min\left[1, \int h_g(x-\epsilon)\psi(\epsilon)\mathrm{d}\epsilon + \frac{\eta_g(x)}{\sqrt{N}}\right], \tag{11}$$

where $\eta_g(x)$ is some (correlated in x) noise of zero-mean with a variance equal to the integral appearing in eq. (11). One can show as in [5] that this noise is Gaussian far from the tip of the front.

In the large-N limit, eq. (11) becomes deterministic. One can look for traveling-wave solutions moving at a velocity v. In the region where $h \ll 1$, the front has an exponential shape $h_g(x) \simeq A\exp[-\gamma(x - vg)]$, where γ is related to v by

$$v = v(\gamma) = \frac{1}{\gamma}\ln\left[\int e^{\gamma\epsilon}\psi(\epsilon)\mathrm{d}\epsilon\right]. \tag{12}$$

If the front is of the F-KPP type [3], its velocity for steep enough initial conditions is the smallest one for which γ is real. Furthermore, for finite but large N, one expects [15, 17, 34]

a correction to this velocity of order $[\ln N]^{-2}$ and a diffusion constant scaling like $[\ln N]^{-3}$. We checked that these predictions are compatible with our numerical simulations for all the models that we tested when $v(\gamma)$ has a minimum value (all the models of figs. 1, 2 except the exponential case have this property). For other choices such as $\psi(\epsilon) = \exp[-\epsilon]$, however, $v(\gamma) = \infty$ for all γ, and the front is not of the F-KPP type. It is therefore not surprising that the genealogy of the exponential case is special.

In all cases, the times $\langle T_2 \rangle$ or $\langle T_p \rangle$ give the order of magnitude of the age of the most recent common ancestor of the whole population and, therefore, the time scale on which the population looses memory about its genealogy. Now, the position of this most recent common ancestor has fluctuations of order 1, and this contributes a random shift of order 1 to the displacement of the front. Therefore, the fluctuating part of the position at generation g is the sum of roughly $g/\langle T_2 \rangle$-independent random variables of order 1. Within this picture, the diffusion constant but also all the cumulants of the position would scale like

$$D_N \sim 1/\langle T_2 \rangle \sim 1/\langle T_p \rangle \ . \tag{13}$$

That all cumulants have the same large-N–dependence was indeed one of the main results of our previous work [17]. Note that eq. (13) does hold (compare eqs. (6) and (10)) in the exponential case. Assuming that it remains valid in general, we would then predict

$$\alpha_{\text{prediction}} = 3, \tag{14}$$

since there is now good numerical evidence [15,34] as well as analytic arguments [17] in favor of $D_N \simeq [\ln N]^{-3}$ in the generic case, *i.e.* if $v(\gamma)$ has a finite minimum. We think that this prediction agrees with our measurements (4) up to finite-size corrections: in fact, for the diffusion constant itself, it already turned out [15,17,34] that the large-N asymptotic regime was only observed for much larger systems than the ones studied here.

Beyond the fact that the time scale is logarithmic in N, which is not surprising for models of evolution in the presence of strong selection [35–37], the times $\langle T_p \rangle$ are characteristic of the statistical properties of the genalogical trees of samples of a few individuals. For the exponential model, these statistics are totally specified by the fact that there is at most one single coalescence event at each generation with probability given by eq. (7). Surprisingly, these coalescence probabilities (up to the factor $\ln N$ which fixes the time scale) are the same as those which emerged from the theory of spin glasses [28, 31, 32], so that trees, in the exponential Model B here, have exactly the same statistics as the ultrametric trees of Parisi's mean-field theory of spin glasses [38, 39]. So far we have not been able to develop a replica approach for noisy traveling waves to justify this connection. There is however some hope to do so since noisy traveling waves appear in the study of directed polymers in a random medium [5], a system for which Parisi's theory is known to be valid at the mean-field level [4].

The numerical data presented in this paper show that, for the class of models we considered, selection has drastic effects on the genealogies: the coalescence times become logarithmic in the population size (2) instead of linear and the statistics of the coalesence times are modified. The accuracy of our simulations is not sufficient to be sure that the exponent α and the ratios of coalescence times are universal (for all models for which $v(\gamma)$ has a finite minimum). We however gave an argument (13) which supports the conjecture (14). Of course, developing an analytical approach susceptible of proving or disproving this universality is a challenging open problem. Another open issue is whether thinking in terms of genealogies is limited to the family of selection models discussed here or could be extended to more general noisy traveling-wave equations.

Lastly, it would be interesting to know what our ratios (3) would become in other models of evolution with selection such as [40, 41] and if there is a chance of estimating them from experimental data on genetic diversity.

* * *

This work was partially supported by the US Department of Energy.

REFERENCES

[1] Fisher R. A., *Ann. Eugen.*, **7** (1937) 355.
[2] Kolmogorov A., Petrovsky I. and Piscounov N., *Bull. Univ. État Moscou, A*, **1** (1937) 1.
[3] van Saarloos W., *Phys. Rep.*, **386** (2003) 29.
[4] Derrida B. and Spohn H., *J. Stat. Phys.*, **51** (1988) 817.
[5] Brunet É. and Derrida B., *Phys. Rev. E*, **70** (2004) 016106.
[6] Carpentier D. and Le Doussal P., *Nucl. Phys. B*, **588** (2000) 531.
[7] Iancu E., Mueller A. H. and Munier S., *Phys. Lett. B*, **606** (2005) 342.
[8] Breuer H.-P., Huber W. and Petruccione F., *Europhys. Lett.*, **30** (1995) 69.
[9] Doering C. R., Mueller C. and Smereka P., *Physica A*, **325** (2003) 243.
[10] Krapivsky P. L. and Majumdar S. N., *Phys. Rev. Lett.*, **85** (2000) 5492.
[11] Lemarchand A., Lesne A. and Mareschal M., *Phys. Rev. E*, **51** (1995) 4457.
[12] Mueller C. and Sowers R. B., *J. Funct. Anal.*, **128** (1995) 439.
[13] Mai J., Sokolov I. M. and Blumen A., *Phys. Rev. Lett.*, **77** (1996) 4462.
[14] Panja D., *Phys. Rep.*, **393** (2004) 87.
[15] Brunet É. and Derrida B., *J. Stat. Phys.*, **103** (2001) 269.
[16] Kloster M., *Phys. Rev. Lett.*, **95** (2005) 168701.
[17] Brunet É., Derrida B., Mueller A. H. and Munier S., *Phys. Rev. E*, **73** (2006) 056126.
[18] Moro E., *Phys. Rev. E*, **69** (2004) 060101(R).
[19] Kessler D. A., Ner Z. and Sander L. M., *Phys. Rev. E*, **58** (1998) 107.
[20] Colon J. G. and Doering C. R., *J. Stat. Phys.*, **120** (2005) 421.
[21] Pechenik L. and Levine H., *Phys. Rev. E*, **59** (1999) 3893.
[22] Escudero C., *Phys. Rev. E*, **70** (2004) 041102.
[23] Brunet É. and Derrida B., *Phys. Rev. E*, **56** (1997) 2597.
[24] Peng W., Gerland U., Hwa T. and Levine H., *Phys. Rev. Lett.*, **90** (2003) 088103.
[25] Snyder R. E., *Ecology*, **84** (2003) 1333.
[26] Kloster M. and Tang C., *Phys. Rev. Lett.*, **92** (2004) 038101.
[27] Schweinsberg J., *Electron. J. Probab.*, **5** (2000) 1.
[28] Pitman J., *Ann. Probab.*, **27** (1999) 1870.
[29] Kingman J. F. C., *J. Appl. Probab. A*, **19** (1982) 27.
[30] Tavaré S., Balding D. J., Griffiths R. C. and Donelly P., *Genetics*, **145** (1997) 505.
[31] Bolthausen E. and Sznitman A.-S., *Commun. Math. Phys.*, **197** (1998) 247.
[32] Ruelle D., *Commun. Math. Phys.*, **108** (1987) 225.
[33] McKean H. P., *Commun. Pure Appl. Math.*, **28** (1975) 323.
[34] Moro E., *Phys. Rev. E*, **70** (2004) 045102(R).
[35] Kaplan N. L., Hudson R. R. and Langley C. H., *Genetics*, **123** (1989) 887.
[36] Durrett R. and Schweinsberg J., *Theor. Popul. Biol.*, **66** (2004) 129.
[37] Durrett R. and Schweinsberg J., *Stoch. Processes Appl.*, **115** (2005) 1628.
[38] Parisi G., *J. Phys. A*, **13** (1980) 1101.
[39] Mézard M., Parisi G., Sourlas N., Toulouse G. and Virasoro M. A., *J. Phys. (Paris)*, **45** (1984) 843.
[40] Tsimring L., Levine H. and Kessler D. A., *Phys. Rev. Lett.*, **76** (1996) 4440.
[41] Kessler D. A., Levine H., Ridgway D. and Tsimring L., *J. Stat. Phys.*, **87** (1997) 519.