On the Mean Convergence Time of Evolutionary Algorithms without Selection and Mutation*

Hideki Asoh
Electrotechnical Laboratory
Tsukuba, Ibaraki 305, Japan
asoh@etl.go.jp

Heinz Mühlenbein
GMD, Schloß Birlinghoven
D-53754 Sankt Augustin, Germany
muehlen@gmd.de

Abstract

In this paper we study random genetic drift in a finite genetic population. Exact formulae for calculating the mean convergence time of the population are analytically derived and some results of numerical calculations are given. The calculations are compared to the results obtained in population genetics. A new proposition is derived for binary alleles and uniform crossover. Here the mean convergence time τ is almost proportional to the size of the population and to the logarithm of the number of the loci. The results of Monte Carlo type numerical simulations are in agreement with the results from the calculation.

1 Introduction

Two opposite tendencies operate on natural populations: *natural selection*, or the propensity to adapt to a given environment; and *polymorphism*, or the propensity to produce variation to cope with changing environments. With the explosion of data reporting polymorphism on the biochemical level, the long-standing problem of the relative importance of nonrandom and random processes in the genetic structuure of populations has revived in the form of a selectionist-neutralist controversy. The most prominent neutralist is Kimura[8].

The controversy has stimulated the study of many stochastic models including the infinite alleles model and the sampling formula. Later Kimura and many others extensively applied diffusion analysis to the study of stochastic genetic models[7]. The problems considered include the analysis of random sampling effects due to small populations, the balance in small populations of recurrent mutation and random genetic drift, the expected time of fixation of a mutant gene.

Random genetic drift is also important for *evolutionary algorithms*. It is a source of reducing the variation of the population. But if the variation is reduced then the response

*Technical report GMD-AS-TR-94-12

to selection becomes less in the next generation [10]. In this paper we will compute the expected time until convergence for different genetic models. Convergence means that all genotypes in the population become equal. One model deals with recombination by unifrom crossover. This model, which is the most important for evolutionary algorithms, has not been investigated before. We derive exact formulae for calculating the mean convergence time and compare them with the results from Monte Carlo type simulations. These results show that uniform crossover recombination increases the convergence time only slightly.

The outline of the paper is as follows. In the next section we analyse the classical simple sampling case as a preparation for treating the case with uniform crossover. In section 3 we present the main result. Section 4 is for the comparison with simulations, and section 5 is for discussion and conclusion.

2 Random drift with simple sampling

2.1 Two alleles

Consider a population of N individuals. Assume that each individual has only one gene (one locus) in which there are two different alleles termed "A" and "a". There is no mutation, crossover, and selection. The generations are discrete and the size of the population is fixed, that is, in each new generation we sample N offspring from the gene pool of N ancestors with replacements. This model is approximately equivalent with the classical diploid model with $N/2$ individuals which is usually treated in the literatures of quantitative genetics.

We can describe the status of the population by the number of individuals which have genotype "A". Let the set of possible states be $\Theta = \{0, 1, ..., \mathrm{N}\}$. The development of the state of the population can be described by a simple Markov chain[2][6]. We denote the probability of the population to be in state $i \in \Theta$ at time t as $P_i(t)$. Then the transition probability of the Markov chain from the state i to the state j is denoted as $q(j|i)$ $i, j \in \Theta$.

The product law leads us to the following formula for $q(j|i)$

$$q(j|i) = \left(\begin{array}{c} N \\ j \end{array} \right) \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j}, \tag{1}$$

and the relation

$$P_j(t) = \sum_{i=0}^{N} q(j|i) P_i(t-1) \tag{2}$$

holds.

When we use vector-matrix notation, we denote the transition probability matrix as $Q = (q_{ji})$, where $q_{ji} = q(j-1|i-1)$ and probability vector as $\mathbf{P}(t)$. The i-th element of $\mathbf{P}(t)$ is $P_{i-1}(t)$. Then we can write the equation (2) as

$$\mathbf{P}(t) = Q\mathbf{P}(t-1) \tag{3}$$

and naturally the equation

$$\mathbf{P}(t) = Q^t\mathbf{P}(0) \tag{4}$$

holds. Here Q^t is the power t of the matrix Q.

If the population is in the state 0 or N, the population is homogeneous. Hence, the probability of the population converging by time $t = k$ can be expressed as

$$s(k) = P_0(k) + P_N(k). \tag{5}$$

The probability of convergence just at time $t = k > 0$ is

$$c(k) = s(k) - s(k-1). \tag{6}$$

We can calculate the mean convergence time τ using $c(k)$ as

$$\tau = \sum_{k=1}^{\infty} k\ c(k). \tag{7}$$

When k is large enough, $kc(k)$ decreases as k increases and converges to 0 quickly. By taking a large enough K we can approximately calculate τ as

$$\tau \approx \sum_{k=1}^{K} k\ c(k) = Ks(K) - (\sum_{k=0}^{K-1} s(k)). \tag{8}$$

Table 1 and figure 1 show some results of numerical calculations using the equation (8). We tested values of K and found that $K = 1000$ is large enough for obtaining the given results.

p_A	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$	
1/2	2.00	4.55	9.89	20.76	42.71	$\tau \approx 1.4N$
3/4	—	3.69	7.94	16.71	32.48	$\tau \approx 1.0N$
7/8	—	—	5.26	11.02	22.85	$\tau \approx 0.7N$

Table 1. Mean convergence time τ for simple sampling for 2 alleles (1)

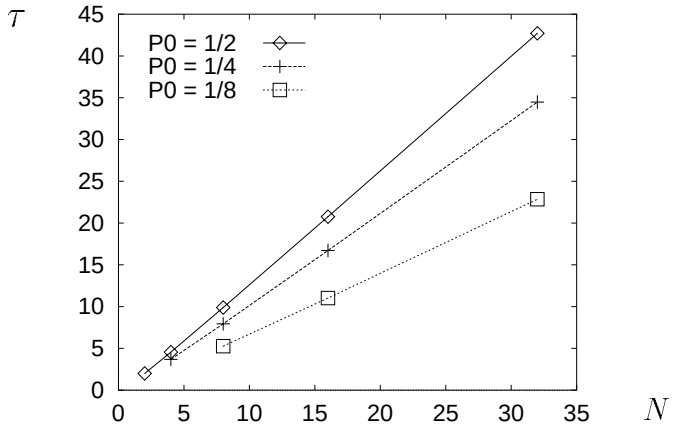

Figure 1. Population size versus mean convergence time (1)

Here the value of p_A means that in the initial population $p_A * N$ individuals have genotype "A" and the rest $((1 - p_A) * N)$ have genotype "a". We can summarize the results in the following proposition.

Proposition 1 *Let each individual have one gene with two alleles "A" and "a". Then in a population of size N with random sampling, the mean convergence time τ increases almost proportionally with the population size N, and in case of $p_A = 1/2$ (half of the initial population have allele "A", the rest have allele "a") $\tau \approx 1.4N$ holds.*

Kimura et al. approximately analysed the equivalent genetic model using the diffusion equations. The first order approximation of $s(k)$ is[7]

$$s(k) \approx 1 - 6p_A(1 - p_A)e^{-t/2N}. \tag{9}$$

From this formula one can calculate τ as $\tau = 12p_A(1 - p_A)N$. This gives $\tau = 3N$ for $p_a = 1/2$. If more terms are used $\tau = 2.8N$ is obtained for $p_A = 1/2$[5]. Considering that in their model each individual has two chromosomes (diploid), their results and the above exact calculation are consistent.

If $p_A > 1/2$, the initial population is biased and has a greater tendency to converge to genotype "A" and lesser tendency to converge to genotype "a". We call these cases as "all A" and "all a" respectively. For these cases, we can calculate the conditional mean of the convergence time under the condition of "all A" or "all a". These results are shown in the following.

p_A	Final State	$N = 4$	$N = 8$	$N = 16$	$N = 32$	
3/4	all A	2.99	6.43	13.60	28.16	$\tau \approx 0.9N$
3/4	all a	5.78	12.47	26.06	53.45	$\tau \approx 1.7N$
7/8	all A	—	4.08	8.55	17.83	$\tau \approx 0.6N$
7/8	all a	—	13.57	28.30	57.98	$\tau \approx 1.9N$
15/16	all A	—	—	5.26	10.89	$\tau \approx 0.35N$
15/16	all a	—	—	29.34	60.10	$\tau \approx 1.92N$

Table 2. Mean convergence time for simple sampling for 2 alleles (2)

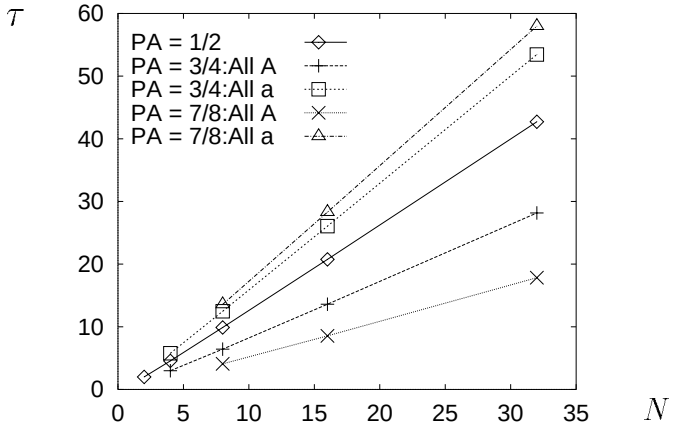

Figure 2. Population size versus mean convergence time (2)

Let $P_A(\infty)$ and $P_a(\infty)$ denote the probability of the population converging to "all A" and "all a" respectively. As for these probabilities, we can prove the following theorem:

Theorem 1 *Consider a genetic population of size N. Let each individual have only one gene with two alleles "A" and "a", and in the initial state, $p_A N$ individuals have allele "A" and the rest have allele "a". Then in a randomly mating population,*

$$P_A(\infty) = p_A, \quad P_a(\infty) = 1 - p_A. \tag{10}$$

Proof Using the formula

$$\sum_{i=0}^{N} q(i|k)\frac{N-i}{N} = \sum_{i=0}^{N} \left(\begin{array}{c} N \\ i \end{array} \right) \left(\frac{k}{N}\right)^i \left(\frac{N-k}{N}\right)^{N-i} \frac{N-i}{N} = \frac{N-k}{N}, \tag{11}$$

we can caliculate $Q^\infty = \lim_{t\to\infty} Q^t$ as

$$Q^\infty = \left(\begin{array}{cccccc} 1 & (N-1)/N & (N-2)/N & \cdots & 1/N & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1/N & 2/N & \cdots & (N-1)/N & 1 \end{array} \right). \tag{12}$$

With this Q^∞ we can readily get $\mathbf{P}(\infty)$ ∎

That is, if the initial population has n times more individuals with genotype "A" at the initial stage, then the probability of the population converging to "all A" is also n times larger.

2.2 Many alleles

When we have more than two alleles, the Markov chain describing the development of the population becomes rather complicated. If we have $m \geq 2$ alleles, we must consider a Markov chain with $(\sum_{i_m=0}^{N} \sum_{i_{m-1}=0}^{i_m} \cdots \sum_{i_2=0}^{i_3} 1)$ states. Each state is characterized by the number of each genotype included in the population. Although we can calculate the mean convergence time in principle, it is almost impossible to do the calculation, and we will not describe the formulae here. However, if the number of alleles is very large and we can assume that all individuals have a different genotype at the initial stage, there is a simple trick to calculate the mean convergence time.

Let each different genotype be a_i $(i = 1, ..., N)$. The probability of the convergence to a genotype a_i is equal for all i and is $1/N$. We denote this probability as $P_{a_i}(\infty)$. The conditional mean convergence time τ_{a_i} under the condition that the population converges to a_i is also equal for all i. Let this conditional mean be τ_c. The "unconditional" mean convergence time τ can be calculated as

$$\tau = \sum_{a_i} \tau_{a_i} P_{a_i}(\infty) = \sum_{a_i} \frac{1}{N} \tau_c, \tag{13}$$

and is equal to τ_c in this case.

Now a fact worth noticing is that τ_c is equal to the conditional mean convergence time in the two alleles case under the condition of converging "all A" from the initial state $p_A = 1/N$. According to this consideration, we can calculate τ for the case with very large number of alleles by the same formula as for the case with two alleles. The results are shown in the following table 3. We also show the results from Monte Carlo type simulations done by Mühlenbein et al.[11].

	$N=2$	$N=4$	$N=8$	$N=16$	$N=32$	
Exact	2.00	5.78	13.57	29.34	61.12	$\tau \approx 2.0N$
Simulation	—	—	13.6	29.4	60.3	

Table 3. Mean convergence time for simple sampling with many alleles

The remarkable fact is that τ is still proportional to N and only slightly larger than the case with two alleles. Now we can state the following proposition.

Proposition 2 *Let the number of alleles be sufficiently large. Then in a population of size N with random sampling the mean convergence time τ increases almost proportionally with the population size N, and $\tau \approx 2.0N$ holds approximately.*

Note that in this case τ is mathematically equivalent to the the mean fixation time of a mutant gene introduced in a population. In Crow and Kimura[1] $\tau \approx 4.0N$ is derived for the diploid case using the diffusion equation model. Our results are consistent with theirs.

3 Genetic drift with uniform crossover

In this section we investigate how much recombination by uniform crossover can reduce the influence of genetic drift. Uniform crossover is an adaptation of Mendel's chance model to haploid organisms. It is used in many genetic algorithms.

We assume that each individual has one chromosome and each chromosome has n loci. We denote the set of alleles for the i-th locus as Θ_i.

Let the chromosome of parents be $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$. Here $x_i, y_i \in \Theta_i$. Then the offspring $\mathbf{z} = (z_1, ..., z_n)$ is computed by the uniform crossover operation according to the following probability;

$$Prob[z_i = x_i] = 0.5, \quad Prob[z_i = y_i] = 0.5.$$

In the following we assume that all Θ_i are the same. Then the probability of fixing (converging) each locus till time $t = k$ is same for all i and we denote it as $r(k)$. Because each locus behaves statistically independent, the probability of fixing all n loci till the time $t = k$ is easily calculated as $r(k)^n$. Now we got the following theorem,

Theorem 2 *Let the number of loci be n. Then the mean convergence time τ of the population with uniform crossover operation is*

$$\tau = \sum_{k=1}^{\infty} k \, (r(k)^n - r(k-1)^n). \tag{14}$$

We will now assume that each locus has two alleles. In this case, we can put $r(k) = s(k)$, where $s(k)$ was introduced in the previous section. If k is large enough $kc_n(k) = k(s(k)^n - s(k-1)^n)$ is decreasing for k and converges to 0 very rapidly. By taking a large enough K we can approximately calculate τ as

$$\tau \approx \sum_{k=1}^{K} k\ c_n(k) = Ks(K)^n - (\sum_{k=0}^{K-1} s(k)^n). \tag{15}$$

Table 4 and figures 3 show a result of numerical calculations using the above equation (10). We tested some value of K and found that $K = 5000$ is large enough. The horizontal axis of the figures are scaled by $\log_2$.

n	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 32$
1	2.00	4.55	9.89	20.76	42.71
2	2.67	6.32	13.72	28.71	58.90
4	3.50	8.35	18.10	37.80	77.38
8	4.42	10.55	22.86	47.63	97.38
16	5.37	12.86	27.82	57.91	118.24
32	6.36	15.21	32.90	68.41	139.56
64	7.34	17.60	38.03	79.03	161.07
128	8.34	19.99	43.19	89.71	182.62
256	9.34	22.39	48.37	100.42	204.03
512	10.33	24.80	53.55	111.14	225.05
1024	11.33	27.21	58.74	121.87	245.17

Table 4. Mean convergence time with uniform crossover ($p_A = 1/2$)

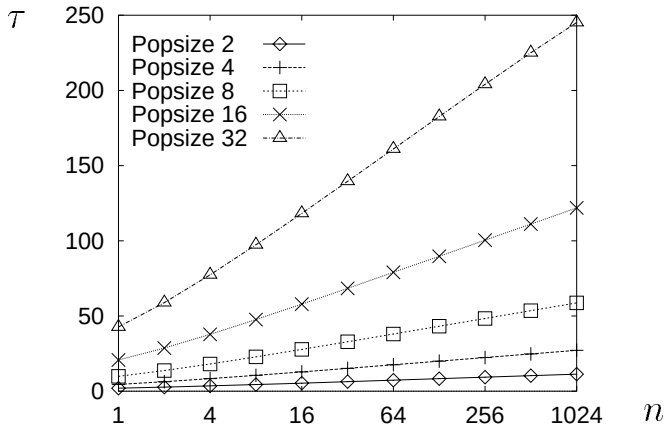

Figure 3. Number of loci versus mean convergence time τ ($p_A = 1/2$)

As you can see from the figure, the mean convergence time τ increases almost proportionally to $\log n$.

Next figure 4 is about the relation with the population size N.

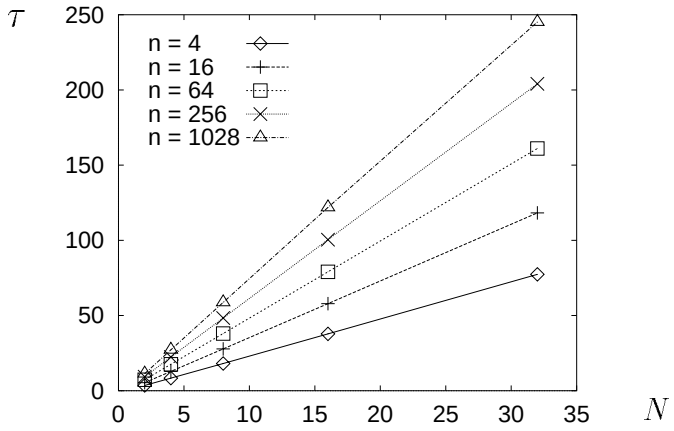

Figure 4. Population size versus mean convergence time ($p_A = 1/2$)

This figure shows that τ increases proportionally to N. To maintain consistency with the results in the previous section, we approximate the above numerical results by a simple formula

$$\tau \approx C_0 N \ (a \log_e n + 1.0)^b. \tag{16}$$

Here C_0 is a constant which depends on p_A and from table 1. The optimal value of a and b which minimize the squared error have been computed and we got the following approximative formulae for some values of p_A.

Proposition 3 *Let the number of loci be n. Let each gene have two alleles. Then the mean convergence time τ of the population with uniform crossover is approximately*

$$\tau \approx 1.4N \ (0.5 \log_e n + 1.0)^{1.1} \quad \text{for} \quad p_A = 1/2, \tag{17}$$

$$\tau \approx 1.0N \ (0.7 \log_e n + 1.0)^{1.1} \quad \text{for} \quad p_A = 3/4, \tag{18}$$

$$\tau \approx 0.7N \ (0.8 \log_e n + 1.0)^{1.2} \quad \text{for} \quad p_A = 7/8. \tag{19}$$

4 Comparison with simulations

We have also done numerical (Monte Carlo type) experiments with our Parallel Genetic Algorithm Simulator "PeGAsuS". The initial population is generated randomly, that is, each locus has a probability 1/2 to have the value 0 and 1/2 to have the value 1. In the simulations self-fertilization is prohibited. This is different from the theoretical analysis. However these differences are not essential here.

In figure 5 the results of simulations are shown, and in figure 6 we make comparison between our exact calculation and simulation.

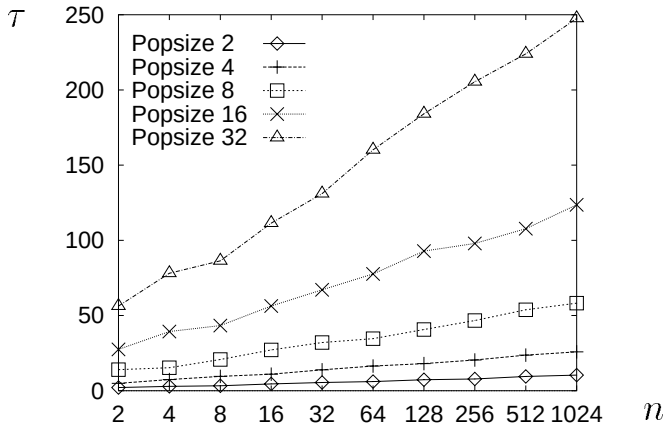

Figure 5. Number of loci versus mean convergence time (Monte Carlo type Simulation)

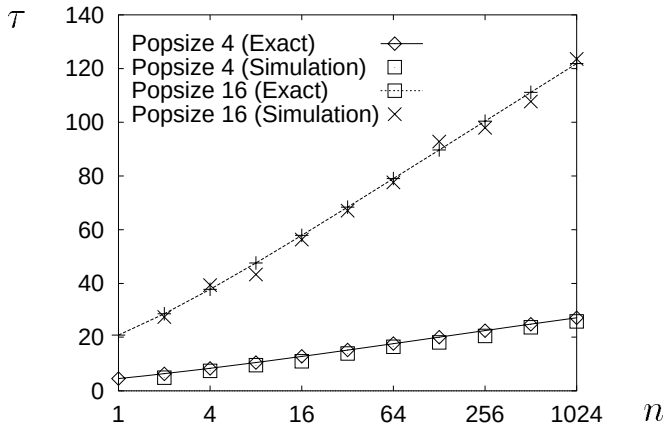

Figure 6. Comparison between exact calculation and simulation

The agreement between the analytical fit and the simulations are very good.

5 Discussion and conclusion

We have derived exact formulae for calculating the mean convergence time of random genetic drift in a random mating population without selection and mutation. Exact numerical calculations using the formulae show that in all cases treated here, the mean convergence time τ is approximately proportional to the size of the population N and to the logarithm of the number of loci n. This means that genetic drift is an important factor for reducing the variance of the population. But the reduction of the variance will reduce the increase of of the average fitness of the population [11]

The above results have been compared with the results from Monte Carlo type experiments. The fit between them is very good. The simulation results also suggest that the

standard deviation of the convergence time increases rather rapidly with population size N. Although in this paper we evaluate only the mean of convergence time, the extension for calculating the variance is straightforward.

An analytical derivation of the proposition 3 and its extension to the case with n genes and large number of alleles are left for future work. Evolutionary Algorithms provide the field of theoretical quantitative genetics with many interesting experimental phenomena in artificial situations. Many topics remain to be investigated in the future.

Acknowledgements

The Monte Carlo type simulation results in section 2 and section 4 are from Andreas Reinholtz. Dirk Schlierkamp-Voosen helped to use PeGAsuS. Byoung-Tak Zhang and Bill Buckles carefully read the manuscript and gave us useful comments. This work was done while one of the authors (Hideki Asoh) was at GMD as a guest researcher. He thanks GMD for that opportunity, and also to the Science and Technology Agency in Japan for supporting his stay. This work is a part of the SIFOGA project supported by Real World Computing Partnership.

References

[1] Crow,J.F. and Kimura,M. *An Introduction to Population Genetics Theory*, Harper and Row, New York, 1970.

[2] Feller,W. *An Introduction to Probability Theory and its Applications. vol.1* (3rd ed.), John Wiley & Sons, New York, 1957.

[3] Fisher,R.A. On the dominance ratio. Proc. Roy. Soc. Edinburgh **42**, 321-341, 1922.

[4] Goldberg,D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Readin, 1989.

[5] Hartl,D.L. *A Primer of Population Genetics*, Sinauer Associates, Sunderland, 1981.

[6] Karlin,S. and Taylor H.M. *A First Course in Stochastic Processes.* (2nd ed.), Academic Press, New York, 1975.

[7] Kimura,M. Diffusion Models in Population Genetics, *J. Appl. Prob.* **1**, 177-232, 1964.

[8] Kimura,M. *The Neutral Theory of Molecular Evolution*, Cambridge Univ. Press, 1983.

[9] Mühlenbein,H. Evolutionary algorithms: Theory and applications, in E.Aarts and J.K.Lenstra (eds.) *Local Search in Combinatorial Optimization*, Wiley, 1993.

[10] Mühlenbein,H.and Schlierkamp-Voosen,D. Predictive models for the Breeder Genetic Algorithm, *Evolutionary Comptation* **1**, 25-49, 1993.

[11] Mühlenbein,H. and Schlierkamp-Voosen,D. The science of breeding and its application to the breeder genetic algorithm BGA, preprint, 1994.

[12] Wright,S. Evolution in Mendelian populations, *Genetics* **16**, 97-159, 1931.