Convergence Analysis of Evolution Strategies with Random Numbers of Offspring

Olivier François

Institut National Polytechnique de Grenoble,
TIMC-IMAG, Faculté de Médecine,
38706 La Tronche, France

Abstract. Hitting times of the global optimum for evolutionary algorithms are usually available for simple unimodal problems or for simplified algorithms. In discrete problems, the number of results that relate the convergence rate of evolution strategies to the geometry of the optimisation landscape is restricted to a few theoretical studies. This article introduces a variant of the canonical $(\mu + \lambda)$-ES, called the Poisson-ES, for which the number of offspring is not deterministic, but is instead sampled from a Poisson distribution with mean λ. After a slight change on the rank-based selection for the μ parents, and assuming that the number of offspring is small, we show that the convergence rate of the new algorithm is dependent on a geometric quantity that measures the maximal width of adaptive valleys. The argument of the proof is based on the analogy of the Poisson-ES with a basic Mutation-or-Selection evolutionary strategy introduced in a previous work.

Keywords: Evolution Strategies, Discrete Optimisation, Convergence Theory, Markov Chains, Large Deviations, Mutation or Selection.

1 Introduction

Evolution Strategies (ES) have generated considerable interest during the last decades, both in the practical and in the theoretical issues [3,5]. Until recently, however, the number of mathematical results about the behaviour of ES has remained rather limited, especially in the field of discrete optimisation. The early theoretical analyses indeed concentrated on continuous optimisation problems, and they were mainly based on the so-called rate-of-progress theory, examining the average gain of the algorithm after a single step of the algorithm [4]. In the continuous setting, global convergence results and results on hitting times of the global optimum are now at least available for simple unimodal problems like the sphere or quadratic functions [1,6], or for simplified algorithms like the $(1 + 1)$-ES [15].

Regarding discrete or combinatorial optimisation, the convergence analysis of evolutionary algorithms has also focused on simple cases, the most representative of which may be the *one-max* problem [9]. Numerous studies have obtained deep insights on such simple problems [18,19], like bounds for the runtime of simple

G. Rudolph et al. (Eds.): PPSN X, LNCS 5199, pp. 21–30, 2008.

EA on pseudo-boolean functions [20,21]. Although some remarkable progress has been achieved for more complex problems [17], the transfer of results and techniques to new problems remain an open question. At the exception of the simulated annealing algorithm [14,19] and of a few variants of evolution strategies that were based on mutation or selection instead of mutation plus selection [10,12], few explicit results have linked hitting times of the global optimum to the geometrical features of the discrete optimisation landscape for an arbitrary optimisation problem. Nevertheless, these scarce results have revealed to be of fundamental interest as they have emphasised the importance of the depths of the adaptive valleys in the simulated annealing algorithm [14], and the importance of their widths in rank-based selection evolutionary algorithms [10].

One difficulty with building a convergence theory for discrete optimisation evolution strategies is the determinism of the selection schemes based on fitness rankings. In this article, we introduce a stochastic variant of ES that converts the usual deterministic offspring assumption made in these algorithms, into an assumption of a stochastic number of offspring. We show that this modification is crucial for characterising the convergence of evolution strategies by means of geometrical quantities.

The article is organised as follows. In section 2, we introduce the Poisson-Evolution Strategy (Poisson-ES) in which the number of offspring is randomly sampled according to the Poisson distribution with mean λ. This modification of the canonical ES will be accompanied by a slight change on the deterministic rank-based selection for the parents, which objective is to prevent premature convergence. Section 3 states our main results about hitting times of the global optimum that are valid for small λ. These results underline the role of the width of adaptive valleys for determining the rate of evolution toward the global optimum. Section 4 presents a short simulation study illustrating the fact that the theory can also predict the behaviour of the modified algorithm for values of λ that are not close to zero.

2 The Poisson-ES

Consider a finite set, V, and assume that we seek the maximum of an injective objective function f defined on V

$$f: \quad V \to \mathbb{R}_+ .$$

Here injectivity is more a convenient assumption than a necessary condition. It facilitates proofs and leads to more elegant statements (see [12] for a more general setting). The canonical $(\mu + \lambda)$-ES is usually defined as follows. At each generation, the algorithm generates a deterministic number of offspring, λ, from μ parents, and simultaneously applies a mutation operator to the λ offspring. Then, μ individuals are selected among the $(\mu+\lambda)$ parents plus offspring to form the parental population in the next generation.

Here, we introduce a variant of the $(\mu + \lambda)$-ES, that generates stochastic – in place of deterministic – numbers of offspring in each generation. In the variant

under consideration, the number of offspring, Λ, is sampled from a Poisson distribution with mean λ, for some value $\lambda > 0$. The probability that the algorithm generates k offspring in a given generation is then equal to

$$\Pr(\text{generate } k \text{ offspring}) = \frac{\lambda^k e^{-\lambda}}{k!}\,, \quad k \geq 0\,.$$

One key property of the Poisson distribution regarding the further analysis of the stochastic dynamics of the algorithm is that

$$\Pr(\text{generate exactly 1 offspring}) = \lambda + o(\lambda)\,,$$

and

$$\Pr(\text{generate} \geq 2 \text{ offspring}) = o(\lambda)\,.$$

In the perspective of a convergence analysis, λ will be thought of as being slowly decreased to zero, like in the simulated annealing algorithm (see [19]). The convergence of annealing schedules will be examined in the next section.

Since our state space V is an arbitrary finite state space, properly defining a mutation operator requires a graph structure, (V, E), that represents how offspring can be generated from the parents. To ensure the irreducibility of the finite Markov chain model for the algorithm, we additionally assume that the graph (V, E) is connected. The mutation operator can then be defined as any particular random walk on the connected graph (V, E).

In the canonical ES, the selection of individuals present in the next generation is usually performed after a deterministic ranking of the parents and offspring. One possible issue with this mode of selection is that random walkers may get trapped into sub-optimal solutions. For example, this can happen if no improvement can be reached by random walking from the last-ranked graph vertex represented in the population. To avoid this issue, we use a slightly modified type of rank-based selection. Instead of selecting μ parents, we actually select $\mu - \Lambda$ parents according to their rank, and then we include the Λ offspring to form the next generation population. This selection scheme requires that the Poisson sampling distribution is conditioned on the event $\Lambda < \mu$, a condition which does not change the above stated key property of the sampling distribution for $\lambda \ll \mu$.

To explain how the modified selection scheme approximates the traditional $(\mu + \lambda)$-ES, we can look at the intermediate generation. After the mutation is applied but before selection is performed, the population consists of the parents plus the offspring,

$$(a_{(1)}, \ldots a_{(\mu)}) + \Lambda \text{ mutant individuals},$$

where the $a_{(i)}$ denote individuals ranked by decreased order of fitness values. Since we are more specifically interested in the behaviour the algorithm for small values of λ, we can approximate the parental population as μ copies of the current best fit individual, $a_{(1)}$,

$$(a_{(1)}, \ldots, a_{(1)}) + \Lambda \text{ mutant individuals}.$$

Accordingly the loss in diversity when replacing μ parents by $\mu - \Lambda$ parents is generally negligible, if not null. Strongly unfavorable mutations produce offspring viable for one generation, but their carriers usually do not transmit their phenotypes in the subsequent generations. The algorithm behaves similarly to the canonical ES except during peak shifts, which are likely to occur more rapidly in the modified version. Finally, the Poisson-ES for discrete optimisation can be summarised as follows.

The $(\lambda + \mu)$-Poisson-ES. The algorithm iteratively applies the following steps until some stopping criterion is met.

1. Conditional on $\Lambda < \mu$, draw $\Lambda \sim \text{Poisson}(\lambda)$.
2. Generate Λ offspring from the μ parents and apply mutation to the offspring.
3. Select $\mu - \Lambda$ parents according to ranked-based selection.
4. Add the Λ mutant offspring to form the next generation population.

3 Convergence Results for the Poisson-ES

In this section, we describe our main results regarding the hitting time of the optimum for an arbitrary injective objective function f defined on a general search space V when the parameter λ is small. When λ is close to zero, the dynamics of the algorithm are strongly dominated by the selection process, which tends to aggregate individuals into homogeneous (homozygous) populations. Mutations, that usually occur at small rates, can essentially be viewed as perturbations of the selection process [13]. In this context, the behaviour of the algorithm is strongly related to S. Wright's concept of a *fitness landscape* and to the presence of *adaptive valleys* [22]. It has long been acknowledged that the widths and the depths of the adaptive valleys may influence the convergence time of evolutionary algorithms [16]. However, there is a lack of theoretical results that can quantify the convergence rate of an algorithm by means of such geometrical quantities.

Let us represent the fitness landscape by the values of the objective function for each vertex of the graph (V, E). In this section, we define a geometrical quantity that intuitively measures the width of the largest adaptive valley in the fitness landscape. For two vertices a and b, which are viewed as two evolutionary distant individuals by the algorithm, let the distance $d(a, b)$ be defined as the length of the shortest path from a to b in (E, V). In other words, the distance is the minimal number of mutations required to transform a into b. The geometrical quantity of interest is [10,12]

$$\ell_* = \max_{a \neq a_{\text{opt}}} \min_{b: f(b) > f(a)} d(a, b)\,,$$

where a_{opt} is the (assumed unique) global optimum of f, and a can be restricted to the set of locally sub-optimal phenotypes [10,12]. This quantity measures the greatest distance between a locally optimal individual and a descendent with a higher selective value.

Our main result can be stated as follows.

Theorem 1. *Let (E, V) be a finite connected graph, and let f be an injective function defined on V. Consider the Poisson-ES with parameters $\mu > \lambda$. Let T_{opt} be the hitting time of the optimal solution a_{opt}, and t_{opt} be its expected value*

$$t_{\text{opt}} = \mathrm{E}_v[T_{\text{opt}}]\,,$$

where v is identified to an arbitrary locally optimal population such that $a_{\text{opt}} \notin v$. Then, we have

$$\lim_{\lambda \to 0} \frac{\log(t_{\text{opt}})}{\log(1/\lambda)} = \ell_*\,.$$

In addition, the standard deviation of T_{opt} is equivalent to the expected value

$$\mathrm{sd}[T_{\text{opt}}] \sim t_{\text{opt}}\,, \quad \lambda \to 0\,.$$

This theorem states that a rough approximation of the mean hitting time can be formulated as

$$t_{\text{opt}} \approx C\lambda^{-\ell_*}\,,$$

for some unknown constant C that depends on μ. Implicitly, the theorem tells us that the spectral gap – that is, one minus the second eigenvalue – of the Markov chain modeling the Poisson-ES is logarithmically equivalent to λ^{ℓ_*} for small λ. Remark that the constant C is not explicit, and may be very large depending on the complexity of the problem under consideration. This happens for example when $\ell_* = 1$, a situation that corresponds to an enumerative sampling strategy. This sort of limitation is also present in the simulated annealing algorithm, where the enumerative strategy leads to a minimum critical depth [14]. In fact, according to [12], and similarly to what has been obtained for the simulated annealing algorithm, the theorem suggests that the convergence towards the optimum can be controlled by a logarithmically decreasing number of offspring

$$\lambda_t = (1+t)^{-\gamma}, \quad \gamma > \ell_*\,, \quad t \geq 0\,. \tag{1}$$

To see this, we can introduce an artificial decreasing temperature schedule (T_t) and perform the following change of parameter

$$\lambda_t = e^{-1/T_t}\,.$$

According to [14,12], a necessary and sufficient condition for convergence of the annealed algorithm to the global optimum is then

$$\sum_{t=1}^{\infty} \lambda_t^{\ell_*} = \infty\,,$$

that justifies the form of equation (1) for λ_t.

The proof of the theorem is based on the theory of large deviations applied to Markov chains with rare transitions [13]. It makes use of the Laplace method for computing sums of exponentials. The arguments for the mean hitting time strictly parallel those given in [12] for the Mutation-or-Selection ES (see also [8]). The result for the standard deviation, as well as other results that confirm the geometric-like behaviour of the hitting times, can be derived from [7]. The complete proof is too long to be reproduced here, and we can only give an outline below.

The Mutation-or-Selection ES is based on the following steps. Let p be a mutation probability. At generation t, let $a^t = a$ denote the current population which we also assume to be of fixed size, μ. To update the current state of the population, the algorithm iterates the following operations

1. Select the best individual from the current population, $a_{(1)}$.
2. For each a_i, $i = 1, \ldots, \mu$, either mutate the individual a_i with probability p, or replace it by $a_{(1)}$.

The connection between the MoS-ES and the Poisson-ES arises as the number of offspring in the MoS-ES is also random, and it is distributed according to the Binomial distribution, $\mathrm{Bin}(\mu, p)$. Most of the large deviation analysis is based on the Laplace method as p goes to zero, and makes use of the following key property of the $\mathrm{Bin}(\mu, p)$ distribution

$$\Pr(\text{generate exactly 1 offspring}) = p + o(p)\,,$$

and

$$\Pr(\text{generate} \geq 2 \text{ offspring}) = o(p)\,,$$

which is very similar to the property stated for the Poisson distribution in the previous section. For the MoS-ES [12], we have previously shown that

$$\lim_{p \to 0} \frac{\log(t_{\mathrm{opt}})}{\log(1/p)} = \ell_*\,.$$

In fact, a rough justification of Theorem 1 would consist in setting $\lambda = p/\mu$ and argueing that the $\mathrm{Bin}(\mu, p)$ distribution can be replaced by a Poisson distribution of mean λ according to the classic Binomial-Poisson approximation in probability theory. Although the guess is correct, the argument can easily be seen to be flawed. However the result transfers to the Poisson-ES after a step by step replication of the proof given in [12].

4 Numerical Illustration

To assess the value of the large-deviation approximation for intermediate values of λ, that is,

$$t_{\mathrm{opt}} \approx C\lambda^{-\ell_*}\,,$$

we performed a comparative evaluation of the performances of the Poisson-ES and the MoS-ES on a very simple test problem. The design of experiments and the fit of the simulated data to the large-deviation approximation can be done using an experimental method based on regression, also described in [2,11].

In our example, the objective function was defined on the set of integers $V = [1, \ldots, 60]$ as

$$f(x) = 1 + (103 - x)x + 120 \sin(x)\,, \quad x \in [1, \ldots, 60]\,,$$

and the mutations were implemented as the (reflected) random walk on V. We added ± 1 with equal probability to each a_i in $\{2, \ldots, 59\}$. The states 1 and 60 could be moved into 2 and 59, respectively. The corresponding fitness landscape is represented in figure 1. This simple optimisation problem is illustrative of the behaviour of the algorithm for a large class of toy problems. It is easy to predict the behaviour of the algorithm, and to compute some geometrical quantities, and can be generalised to many dimensions without difficulties. We used $\mu = 10$ individuals, and the algorithms were started from the homogeneous population $a^1 = (1, \ldots, 1)$. In this example, the adaptive valleys were narrow and easy to cross as we started from a^1, but their width increased as the algorithms moved toward the optimum. The critical parameter ℓ_* was computed as

$$\ell_* = 5\,.$$

In this test problem, the ES were then expected to improve quickly from the starting population, but they were also expected to make slower progress as they approached the global optimum.

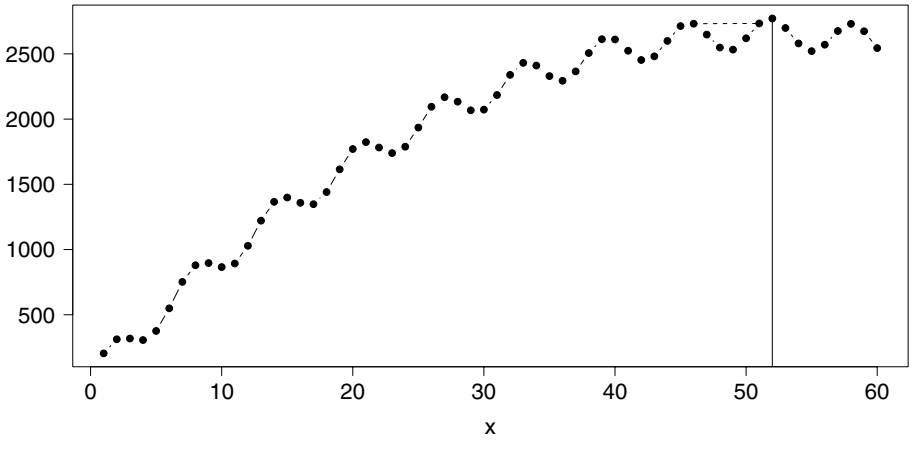

Fig. 1. The toy objective function: $f(x) = 1 + (103 - x)x + 120 \sin(x)$, $x = 1, \ldots, 60$. The optimum is reached at $x = 52$ and, the width of the largest additive valley is represented by a dashed line. We have $\ell_* = 5$.

To evaluate the performances of the algorithms, we regressed the logarithm of the hitting times on the logarithm of λ (or the logarithm of p) in order to estimate ℓ_* as the slope of the regression

$$\log(T_{\text{opt}}) = \log(C) + \ell_* \log(1/\lambda) + \epsilon\,.$$

To obtain comparable results for the Poisson-ES and for the MoS-ES, we set $p = \lambda/\mu$, and we ran simulations for values of p in the interval $(0.15, 0.4)$, where p is the mutation probability. We obtained 250 replicates of the hitting times for regularly spaced values of p. The corresponding values of λ fall in the interval $(1.5, 4)$. The fact that experimental values of λ were not close to zero makes departures from the theoretical predictions rather likely. These values were nevertheless more conform to standard user-defined ones than would have been the very small values suggested by Theorem 1. Figure 2 shows that the data fit the log-log regression rather well ($R^2 = 0.76$, $P \approx 0$ for the Poisson-ES, and $R^2 = 0.81$, $P \approx 0$ for the MoS-ES), providing evidence that the log-hitting times were actually explained by the logarithm of λ. The coarse approximation $t_{\text{opt}} \approx C\lambda^{-\ell}$ could then considered valid for values of λ not close to zero. The coefficients of the regression model were computed as 3.3 (intercept) and 4.2 (slope) for the Poisson-ES, and they were computed as 2.7 (intercept) and 5.9 (slope) in the MoS-ES. The slope values 4.2 and 5.9 were close to the value $\ell_* = 5$ predicted by the theory of large deviations. In this example, we noticed that the Poisson-ES ran slightly faster than the MoS-ES to the population with the highest selective values.

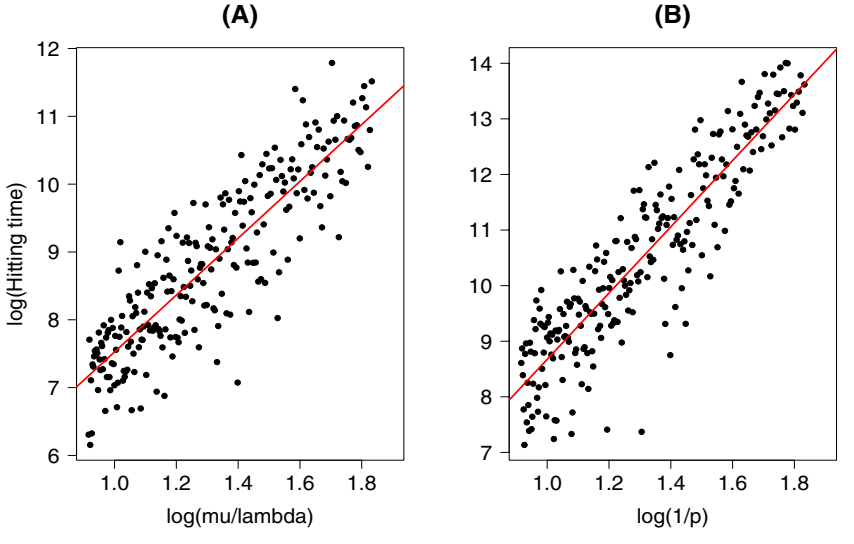

Fig. 2. Regression of the log hitting time on $\log(1/p)$, where p is the mutation probability. The slope of the regression corresponds to the critical quantity ℓ_*. (A) Poisson-ES. (B) MoS-ES, $p = \lambda/\mu$, $\mu = 10$.

5 Discussion

This article has introduced a variant of the canonical ES in which the number of offspring is not deterministic, but is instead sampled from a Poisson distribution with mean λ. After a slight change on the rank-based selection for the μ parents, we showed that the new ES resembles the basic Mutation or Selection-ES introduced in [10]. The Poisson-ES and the MoS-ES provide interesting models for obtaining in-depth insights on the convergence of evolutionary algorithms. Based on the similarity between the two algorithms, we stated a convergence theorem for arbitrary discrete optimization problems, that emphasises the role of the width of the adaptive valleys. Yet, analogs of MoS-ES or of discrete Poisson-ES have not been studied in continuous optimization problems, but it is natural to expect that geometric quantities similar to those influencing the behaviour of the discrete algorithms are likely to determine the convergence rate of the continuous algorithms as well.

Stochastic parameters are the basis for designing adaptive or self-adaptive algorithms. Since the cost of an algorithm is a function of the mutation load through the number of fitness evaluations, an ES should end with $\lambda \approx 0$ when getting close to the optimum. In contrast, being far from the optimum would probably require that the number of offspring is large $\lambda \gg 1$. This idea can be implemented in the Poisson-ES using an explicit convergent annealing scheme. We also believe that this study opens new directions for self-adaptation in discrete ES, because it indicates that increasing λ or performing faster walks in the bottom of the valleys is likely to improve the convergence rate of the algorithm.

Acknowledgments

I wish to thanks Anne Auger for her comments on a previous draft of the manuscript. This work is supported by a grant from the Agence Nationale de la Recherche BLAN06-3146280.

References

1. Auger, A.: Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains. Theor. Comput. Sci. 334, 35–69 (2005)
2. Bartz-Beielstein, T.: Experimental Research in Evolutionary Computation – The New Experimentalism. Natural Computing Series. Springer, Berlin
3. Beyer, H.-G.: The Theory of Evolution Strategies. Natural Computing Series. Springer, Heidelberg (2001)
4. Beyer, H.-G., Schwefel, H.-P., Wegener, I.: How to analyse evolutionary algorithms. Theor. Comput. Sci. 287, 101–130 (2002)
5. Beyer, H.-G., Schwefel, H.-P.: Evolution strategies – A comprehensive introduction. Natural Computing 1, 3–52 (2002)
6. Bienvenüe, A., François, O.: Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. Theor. Comput. Sci. 306, 269–289 (2003)

7. Cercueil, A., François, O.: Sharp asymptotics for fixation times in stochastic population genetics models at low mutation probabilities. Journal of Statistical Physics 110, 311–332 (2003)
8. Cerf, R.: Asymptotic convergence of genetic algorithms. Adv. Appl. Probab. 30, 521–550 (1998)
9. Droste, S., Jansen, T., Wegener, I.: On the analysis of the (1 + 1) EA. Theor. Comput. Sci. (276), 51–81 (2002)
10. François, O.: An evolutionary algorithm for global minimization and its Markov chain analysis. IEEE Trans. Evol. Comput. 2, 77–90 (1998)
11. François, O., Lavergne, C.: Design of evolutionary algorithms: A statistical perspective. IEEE Trans. Evol. Comput. 5, 129–148 (2001)
12. François, O.: Global optimization with exploration/selection algorithms and simulated annealing. Ann. Appl. Probab. 12, 248–271 (2002)
13. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems. Springer, New York (1984)
14. Hajek, B.: Cooling schedules for optimal annealing. Math. Oper. Research 13, 311–329 (1988)
15. Jägersküpper, J.: Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. Theor. Comput. Sci. 379, 329–347 (2007)
16. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1996)
17. Neumann, F., Wegener, I., Randomized, I.: local search, evolutionary algorithms, and the minimum spanning tree problem. In: Deb, K., et al. (eds.) GECCO 2004. LNCS, vol. 3102, pp. 713–724. Springer, Heidelberg (2004)
18. Rudolph, G.: Finite Markov chain results in evolutionary computation: A tour d'horizon. Fundam. Inform. 35, 67–89 (1998)
19. Schmitt, L.M.: Theory of genetic algorithms. Theor. Comput. Sci. 259, 1–61 (2001)
20. Wegener, I., Witt, C.: On the optimization of monotone polynomials by simple randomized search heuristics. Combin. Probab. Comput. 14, 225–247 (2005)
21. Witt, C.: Runtime Analysis of the (μ+1) EA on Simple Pseudo-Boolean Functions. Evol. Comput. 14, 65–86 (2006)
22. Wright, S.: The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proceedings of the VI International Congress of Genetics, pp. 356–366 (1932)