Dimensional Analysis of Allele-Wise Mixing Revisited

Dirk Thierens

Department of Computer Science
Utrecht University, The Netherlands
dirk.thierens@cs.ruu.nl

Abstract. This paper revisits an important, yet poorly understood, phenomenon of genetic optimisation, namely the mixing or juxtapositioning capacity of recombination, and its relation to selection. Mixing is a key factor in order to determine when a genetic algorithm will converge to the global optimum, or when it will prematurely converge to a suboptimal solution. It is argued that from a dynamical point of view, selection and recombination are involved in a kind of race against time: the number of instances of good building blocks is quickly increased from generation to generation by the selection phase, but in order to create optimal solutions these building blocks have to be juxtaposed by the crossover operator and this also takes some time to occur. If the selection of building blocks goes too fast - relative to the rate at which crossover can juxtapose or mix them - then the population will prematurely converge to a suboptimal solution. Previous work (Goldberg, Deb & Thierens, 1993) made a first step toward a better understanding of mixing in genetic algorithms, and also introduced the use of dimensional analysis in GA modelling. In this paper we extend this work by integrating some of the insights gained from the modelling of the dynamic behaviour of GAs on infinite populations (Mühlenbein & Schlierkamp-Voosen, 1993; Thierens & Goldberg, 1994; Bäck, 1995; Miller & Goldberg, 1995). The resulting dimensional model quantifies the allele-wise mixing process: it specifies the boundary in the GA parameter space between the region of reliable convergence at one side, and the region of premature convergence at the other. Although the model is limited to simple bit-wise mixing, the lessons learned from it are quite general and are also valid for more difficult, building-block based problems.

1 Introduction

Genetic algorithms are complex adaptive systems and therefore it should come as no surprise that different modelling approaches are needed to highlight different aspects of their behaviour. Traditional analysis of genetic algorithms - as stated by the schema theorem - focuses on the selective growth of good building blocks and on the destructive effects of recombination on this growth (Holland, 1975; Goldberg, 1989). An alternative approach studies the GA convergence properties by modelling the dynamic change of the population fitness distribution (Vose & Liepins, 1991; Whitley, 1993; Mühlenbein & Schlierkamp-Voosen, 1993; Thierens

& Goldberg, 1994; Bäck, 1995; Miller & Goldberg, 1995). These analytical tools are all quite useful but unfortunately they do not answer all the questions one would like to get answered. For instance the analyses do not give an answer to the important question whether a finite sized population will prematurely converge or not, nor do they consider the effect of the different GA parameters on this phenomenon.

In this paper we study the dynamic interaction between the selection phase and the recombination phase. In contrast with the schema theorem analysis our attention is now focused on the constructive capacities of recombination. It is argued that selection and recombination are basically involved in a race against time. Selection makes copies of the best strings and since the population size is fixed a loss of diversity occurs. At the same time recombination has to mix or juxtapose the good genetic material, but due to the loss of diversity it has only limited time to achieve this. If selection goes too fast then recombination will have not enough time to bring all the optimal alleles together and the population will prematurely converge to a suboptimal solution.

A very useful methodology to quantify this selection-recombination race is the so called dimensional analysis methodology (Ipsen, 1960). Dimensional analysis tries to identify the basic dimensions or key features of a certain process and establishes the functional relationship between them. Important key features for genetic algorithms are the selection pressure, the recombination rate, the population size, the string length and the building block length. The dimensional analysis methodology was first applied to the study of GAs in (Goldberg, Deb & Thierens, 1993) where a first step at allele-wise modelling was performed. In (Thierens & Goldberg, 1993) the methodology was used to show the limitations of simple genetic algorithms when they have no linkage information about the building blocks. A general discussion of the methodological aspects and their use for GA analysis can be found in (Goldberg, 1994). This paper extends the work done in (Goldberg, Deb & Thierens, 1993) and builds a complete dimensional model for allele-wise mixing by integrating some of the insights gained from the modelling of the dynamic behaviour of GAs on infinite populations (Mühlenbein & Schlierkamp-Voosen, 1993; Thierens & Goldberg, 1994; Bäck, 1995; Miller & Goldberg, 1995). In addition we also perform an extensive set of experiments to check the validity of the dimensional model.

2 A Dimensional Model for Allele-Wise Mixing

The purpose of mixing analysis is to study the relationship between the key genetic algorithm parameters that determine whether or not the optimal solution will be found in a reliable way. When a given parameter choice violates this relationship the GA will prematurely converge to a suboptimal solution. In this paper we look at the mixing limitations on the level of alleles, and thus the modelling considers fitness functions with independent allele fitness contributions. Two extreme cases can be considered. In the first case all allele fitness

values are equal - which gives us the well studied bit-counting function - and the proportion of the optimal alleles increases at the same rate in the entire string. In the second case the allele fitness values are exponentially scaled, which causes a domino-like convergence behaviour. The optimal alleles converge sequentially and at any time only a very limited number of genes is actually converging in the so called convergence window. Genes outside the convergence window are either already fully converged or they still have to start converging (Rudnick, 1992). From the mixing point of view, the domino convergence is an easy problem: only a few genes are actually converging at the same time and juxtaposing them is a simple matter. The mixing task becomes much harder with equal allele values because all genes are converging simultaneously and thus they all have to be juxtaposed simultaneously. It has also been shown that the time complexity of the convergence rate is of order $O(\sqrt{\ell})$ for equally scaled allele values, and of order $O(\ell)$ for the exponentially scaled allele values (Thierens, 1995). Since the convergence rate is faster for the uniform allele fitness distribution there is less time to mix all the alleles, and thus the most difficult situation for allele-wise mixing occurs when optimising the bit counting function. Therefore we will develop a dimensional mixing model for this case. The notation used throughout the text is summarised in table 1.

2.1 Recombination

n	population size	p_c	crossover probability
l	string length	p_x	allele swapping probability
s	tournament size	p_r	gene exchange probability
I	selection intensity	r	recombination rate ($= p_c p_r$)
t_x	mixing time	$p(t)$	proportion optimal alleles
t_s	selection time		

Table 1. *Notation.*

We start our dimensional modelling with the analysis of the recombinative capacity of crossover. Whether or not the GA will prematurely converge depends on the relative speed of selection and recombination. To quantify the recombination speed we express the recombinative capacity of crossover in terms of time or equivalently of number of generations.

First the probability that two arbitrary genes are exchanged is simply the probability of applying the crossover operator p_c times the probability p_r that an actual exchange between the two genes takes place. This exchange probability is dependent on the type of crossover operator used. For instance for 1-point crossover we have $p_r = \delta_{ij}/(l-1)$ (with δ_{ij} the distance between the two genes), while for parameterised uniform crossover the gene exchange probability is $p_r =$

$2p_x(1-p_x)$ (with p_x the allele swapping probability). We call the probability of gene exchange the recombination rate r:

$$r = p_c p_r$$

Under repeated use of crossover, the number of generations that are needed in expectation to exchange two arbitrary genes is thus given by $1/nr$. This gene exchange has to happen with all the l genes, so the number of generations that we have to apply recombination or the mixing time t_x is proportional to:

$$t_x \propto \frac{l}{nr} \tag{1}$$

2.2 Selection

The speed at which selection makes the population converge is of course dependent on the selective pressure and on the particular selection algorithm used. A number of studies have derived an exact dynamical model of the convergence of different selection algorithms for normally distributed fitness functions such as the bit counting function (Mühlenbein & Schlierkamp-Voosen, 1993; Thierens & Goldberg, 1994; Bäck, 1995; Miller & Goldberg, 1995). For tournament selection and (μ/λ) or truncation selection this model is given by:

$$p(t) = 0.5(1 + sin(\frac{I}{\sqrt{l}}t + c_0))$$

where I is the selection intensity and $c_0 = arcsin(2p(0)+1)$ a constant dependent on the initial proportion of optimal alleles. For a randomly initialised population the number of generations needed to converge ($p(g_{conv}) = 1$) is thus given by:

$$g_{conv} = \frac{\pi}{2}.\frac{\sqrt{l}}{I}$$

The selection time t_s is therefore proportional to:

$$t_s \propto \frac{\sqrt{l}}{I} \tag{2}$$

Note that for unscaled proportionate selection the time to convergence is given by $g_{conv} = l \log l$. Clearly this is unacceptably slow and in practice unscaled proportionate selection is never used so we will not consider it here.

2.3 Allele-wise mixing model

Now that we have a dimensional relation of both recombination and selection, we need to combine them. When selection acts more slowly than exchange - when $t_x \leq t_s$ - we expect the optimal alleles to have time to be juxtaposed and thus creating good strings which can then be given more copies by selection. This juxtapositioning of the target alleles can then subsequently proceed until all

optimal alleles are put together. However when selection is too fast - when $t_s < t_x$ - too many copies are created of only a few good strings before recombination has had the time to exchange the optimal alleles. As a result some alleles will not be part of strings that receive extra copies and their proportion in the population will decline until they finally disappear. A lack of good mixing thus results in premature convergence.

To quantify this mixing failure we interrelate the previous dimensional relations (Equations 1 and 2) for the recombination and selection process. If we want to avoid a mixing failure we need $t_x \leq t_s$ or:

$$\frac{l}{nr} \leq c \, \frac{\sqrt{l}}{I}$$

where c is a constant factor since dimensional analysis gives us only the functional relationship between the parameters. Rearranging the selection and recombination parameters at one side and the size parameters at the other side we obtain the following mixing model:

$$\frac{\sqrt{\mathbf{l}}}{\mathbf{n}} \leq \mathbf{c} \, \frac{\mathbf{r}}{\mathbf{I}} \tag{3}$$

3 Experimental verification

The use of dimensional modelling allows us to build simple analytical models that expresses the functional relationship between the key parameters. For instance the model tells us that in order to maintain proper convergence the minimal population size needs to be increased by a factor $\sqrt{2}$ whenever the string length is doubled. Dimensional analysis forces us to take a birds eye view on the whole convergence process. It is therefore extremely important to validate the model with experimental results. Since we have four independent variables (population size n, string length l, selection pressure s and the recombination rate $r = p_c p_r$) we have $\binom{4}{2} = 6$ functional relations between two variables while keeping the other independent variables fixed. Exhaustively performing the 6 sets of experiments assures us of the correctness of the dimensional model. In the following experiments the results are obtained by doing 20 independent runs for all the parameter combinations shown. We call the experiments a success whenever in at least 19 of these 20 runs the population converged to the optimal string. At one side of the curve we have reliable convergence to the optimal solution, while at the other side premature convergence occurs. Although the 19-out-of-20 success criterion might seem rather arbitrary, the particular choice has no influence on the functional relationships. The value of the constant c in the dimensional model might change a little bit, but this is of no importance. All experiments are carried out with tournament selection with tournament size s. Since the selection intensity I is approximately proportional to the tournament size s ($\ln s \approx \sqrt{2I}$), we can quantify the selection pressure by the tournament size as was done in (Goldberg, Deb & Thierens, 1993).

1. **string length vs. selection pressure**
   First we look at the functional relation between the string length l and the selection pressure s. We fix the population size $n = 200$, the crossover probability $p_c = 1.0$ and the disruption probability $p_r = 0.3$. We use parameterised uniform crossover so the crossover swapping probability is $p_x = 0.18$ ($p_r = 2p_x(1 - p_x)$). Plugging these values into the mixing model (Equation 3) gives us a prediction of the boundary between successful and failing convergence:

$$\sqrt{l} = \frac{60\ c}{\ln s}$$

   Figure 1 shows the data and the corresponding fitted curve

$$l = \frac{320}{(\ln s)^2}$$

   The predicted functional relation and the data coincide quite well and the corresponding value of the constant factor in the dimensional model is approximately: $c = 0.298$.

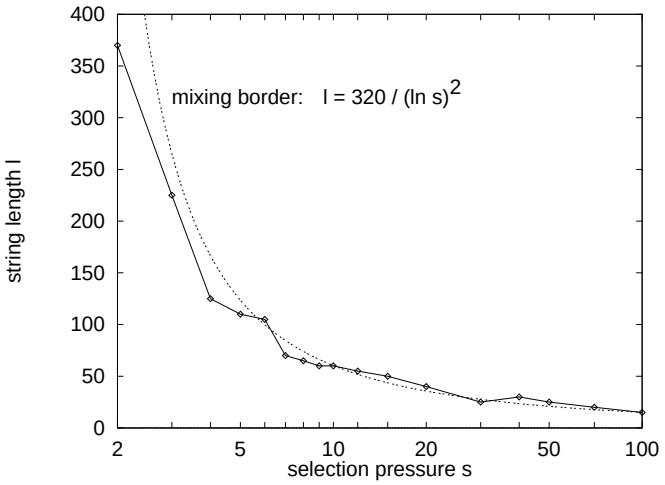


Fig. 1. *Dimensional relation between the selection pressure and the string length.*

2. **disruption probability vs. string length**
   The second experiment checks the relation between the disruption probability p_r and the string length l. The population size $n = 200$, the selection pressure $s = 10$ and the crossover probability $p_c = 1.0$. Note that the choice of the values of the fixed variables is such that the resulting mixing boundary falls within a reasonable range of the varying variables.

The predicted relation is now :

$$p_r = \frac{0.0115}{c} \sqrt{l}$$

Figure 2 shows a fitted curve equal to

$$p_r = 0.045\sqrt{l}$$

which gives us for the constant factor: $c = 0.256$.

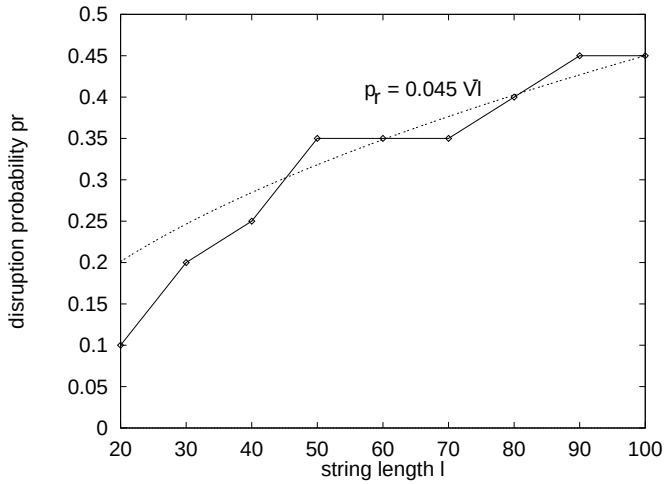


Fig. 2. *Dimensional relation between the string length and the disruption factor.*

The experimental data shows good agreement although for very low string lengths ($l < 30$) the necessary disruption probability is actually a little bit lower than predicted. This can be understood however if we realize that for such small string lengths the search problem becomes almost trivial.

3. **crossover probability vs. selection pressure**
For the relation between the selection pressure and the crossover probability we take the string length $l = 100$, the population size $n = 200$ and the disruption probability $p_r = 0.5$, the predicted functional relation now becomes:

$$p_c = \frac{\ln s}{10\ c}$$

Figure 3 shows the experimental results and the fitted curve:

$$p_c = 0.35 \ln s$$

which results in a constant factor : $c = 0.286$.
For large values of the selection pressure s the GA behaviour is no longer determined by the mixing but by the cross-competition between the genes (Goldberg, Deb & Thierens, 1993).

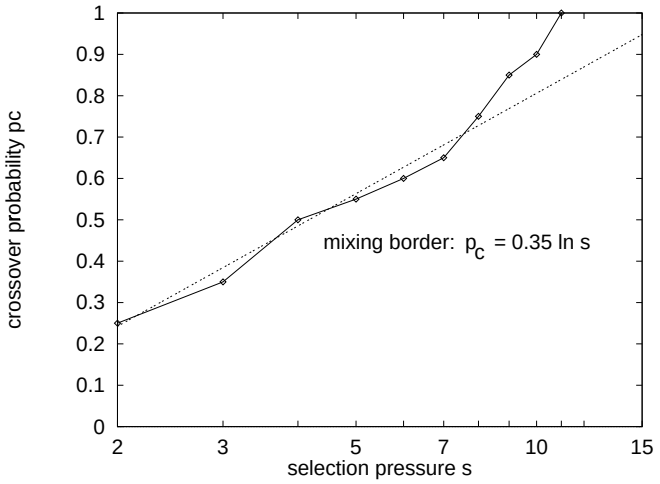


Fig. 3. *Dimensional relation between the selection pressure and the crossover probability.*

4. **population size vs. disruption probability**
   The fourth set of experiments establishes the relation between the population size n and the disruption probability p_r. Taking for the string length $l = 100$, the selection pressure $s = 4$ and the crossover probability $p_c = 1$ the predicted relation is:

$$n = \frac{13.86}{c\ p_r}$$

   Figure 4 shows the experimental results and the fitted curve:

$$n = \frac{50}{p_r}$$

   and a corresponding constant factor: $c = 0.277$.
5. **population size vs. selection pressure**
   For a string length $l = 100$, crossover probability $p_c = 0.75$ and disruption probability $p_r = 0.20$ ($\Rightarrow p_x = 0.11$) the relation between the population size n and the selection pressure s is:

$$n = \frac{66.67}{c} \ln s$$

   Experimental results (Figure 5) are

$$n = 236 \ln s$$

   or $c = 0.282$. Again we notice that for very large values of the selection pressure s cross-competition starts to influence the convergence results.

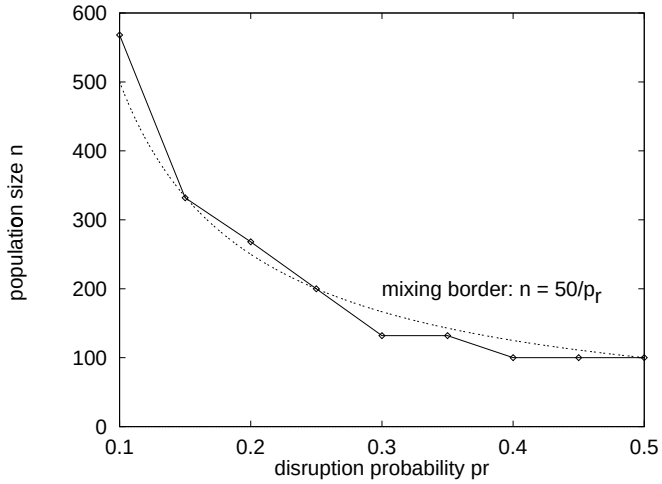
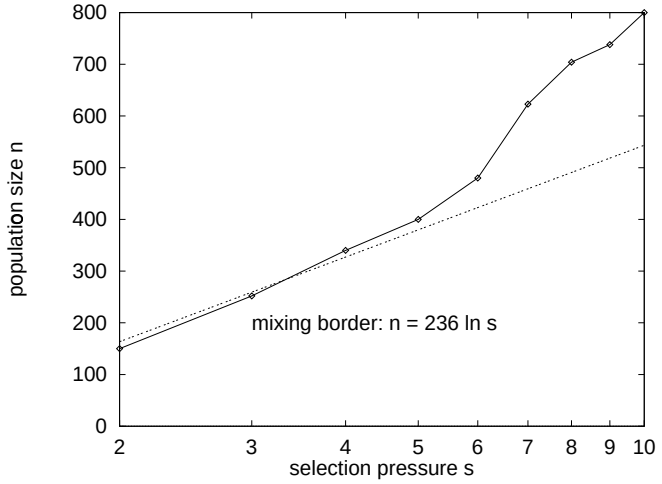


Fig. 4. *Dimensional relation between the disruption factor and the population size.*



r

Fig. 5. *Dimensional relation between the selection pressure and the population size.*

6. **population size vs. string length**
   Finally the last set of experiments looks at the relation between the population size and the string length. Fixing the selection pressure $s = 2$, the crossover probability $p_c = 1.0$ and the disruption probability $p_x = 0.1 (\Rightarrow p_r = 0.18)$ the dimensional model predicts:

$$n = \frac{7.7}{c} \sqrt{l}$$

   Experimental results are shown in Figure 6. The fitted curve is:

$$n = 27.3\sqrt{l}$$

which gives us for the constant factor: $c = 0.282$[1].

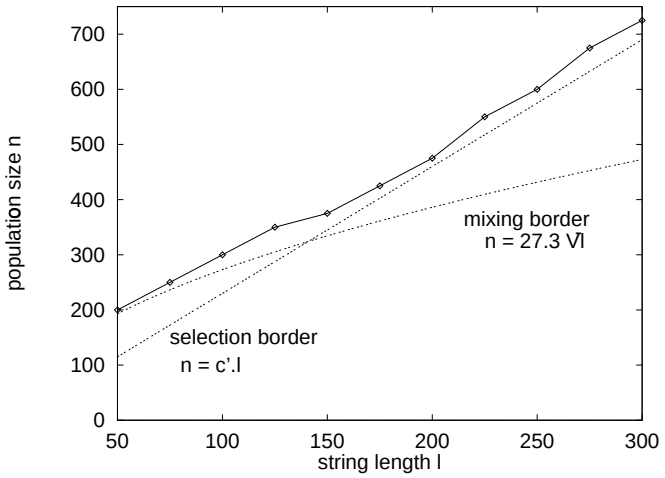


Fig. 6. *Dimensional relation between the string length and the population size.*

4 Conclusion

In this paper we have derived a dimensional model of allele-wise mixing. This model gives the functional relationship between the key genetic algorithm parameters and determines a boundary between reliable convergence and premature convergence. From a dynamical viewpoint, selection and recombination are involved in a times race, and when selection goes too fast relative to the recombination or mixing time, then the GA will fail to converge to the global optimum. The model quantifies the mixing boundary in terms of the selection pressure, the recombination rate, the population size, and the string length. Experimental results show good agreement with the predicted functional relations.

[1] The 6 fitted curves gave an approximate value to the constant factor of respectively $c = \{0.298,\ 0.256,\ 0.286,\ 0.277,\ 0.282,\ 0.282\}$. Considering the presence of the constant $\sqrt{\pi}$ in the normal distribution modelling of GA convergence it is interesting to note that: $1/2\sqrt{\pi} = 0.282$. However further analysis would be needed to test the sensitivity of the constant value on a number of factors, such as the choice of success criterion and the curve fitting method.

Acknowledgement

We would like to thank David E. Goldberg for the many useful discussions on the importance of dimensional modelling in complex systems analysis.

References

1. Bäck T. (1995). Generalised Convergence Models for Tournament- and (μ, λ)-Selection. Proceedings of the Sixth International Conference on Genetic Algorithms.
2. Goldberg D.E. (1989). *Genetic Algorithms in Search, Optimisation and Machine Learning.* Addison Wesley Publishing Company.
3. Goldberg D.E. (1994). *First Flights at Genetic-Algorithm Kitty Hawk.* IlliGAL Report No. 94008. University of Illinois at Urbana-Champaign, Illinois Genetic Algorithm Laboratory.
4. Goldberg D.E., & Deb K. (1991). A comparative analysis of selection schemes used in genetic algorithms. *Proceedings of Foundations of Genetic Algorithms FOGA-I.* ed. G. Rawlings. pp.69-93. Morgan Kaufmann.
5. Goldberg D.E., Deb K., & Thierens D. (1993). Toward a better understanding of mixing in genetic algorithms. *Journal of the Society for Instrumentation and Control Engineers, SICE Vol.32,* No.1 pp.10-16.
6. Holland J.H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.
7. Ipsen D.C.(1960). *Units, dimensions, and dimensionless numbers.* McGraw-Hill.
8. Miller B. L., & Goldberg D. E. (1995). *Genetic algorithms, selection schemes, and the varying effects of noise* IlliGAL Report No. 95009. Illinois Genetic Algorithm Laboratory. University of Illinois at Urbana-Champaign,
9. Mühlenbein H. & Schlierkamp-Voosen D. (1993). Predictive Models for the Breeder Genetic Algorithm. I. Continuous Parameter Optimisation. *Evolutionary Computation 1(1):25-49,* MIT Press.
10. Rudnick M. (1992). *Genetic Algorithms and Fitness Variance with an Application to the Automated Design of Artificial Neural Networks.* Unpublished doctoral dissertation, Oregon Graduate Institute of Science and Technology, Beaverton.
11. Thierens D., & Goldberg D.E. (1993). Mixing in Genetic Algorithms. *Proceedings of the Fifth International Conference on Genetic Algorithms ICGA-93.* ed. S. Forrest. pp.38-45. Morgan Kaufmann.
12. Thierens D., & Goldberg D.E. (1994). Convergence Models of Genetic Algorithm Selection Schemes. *Lecture Notes in Computer Science, Vol. 866: Parallel Problem Solving from Nature PPSN-III.* Jerusalem (Il). eds. Y. Davidor, H.P. Schwefel, R. Männer. pp.119-129. Springer-Verlag.
13. Thierens D. (1995). *Analysis and Design of Genetic Algorithms.* PhD thesis, Dept. Electrical Engineering, Kath. Univ. Leuven, Belgium.
14. Vose M., & Liepins G. (1991). Punctuated Equilibria in Genetic Search. *Complex Systems 5:31-44*
15. Whitley D. (1993). An Executable Model of a Simple Genetic Algorithm. *Foundations of Genetic Algorithms II.* Morgan Kaufmann.