Individuals' Genealogy and the Population Diversity

Krzysztof Adamski[1,*], Jarosław Arabas[1]
Łukasz Bartnik[1], Arkadiusz Chrustowski[2]
Krzysztof Jamróz[1], Rafał Wardziński[3]

Warsaw University of Technology
[1]Institute of Electronic Systems
[2]Institute of Telecommunications
[3]Institute of Control and Computation Engineering
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
*corresponding author: kadamski@elka.pw.edu.pl

Abstract. In this paper we study the genotype diversity of the base population. We provide evidence that the diversity is closely related to the genealogy of individuals and to the degree of mutation. We give analytical formulas for the diversity distribution in the space of binary and real vectors. The theoretical results are supported by numerical experiments.

1 Introduction

In the field of evolutionary algorithms (EAs) it is widely believed that the population diversity is responsible for the robustness of an EA and the ability to locate global maximum of the fitness function. Some authors suggest that the EA should be stopped if diversity falls below a certain level. With this paper we attempt to explain diversity as the effect of the genealogy (the history of reproduction and inheritance) and the mutation.

The idea to directly analyze the genealogy trees was introduced by Walczak [6], though the idea of observing the history of reproduction and inheritance was raised much earlier. In [2, 5] it is observed that if only selection is applied (neither mutation nor crossover), the population becomes uniform (i.e. contains copies of a single element) in a relatively short time. This effect is caused by the fact that the population size is finite, and the stronger the selective pressure, the shorter is the time until homogeneity. The number of generations needed for homogeneity is called the *takeover time*. Moreover, it has been empirically observed [1] that the takeover time is correlated with the population diversity. In this paper we introduce the concept of the *closest common ancestor* (CCA) and study the probability distribution of the number of generations when the CCA appears for any pair of individuals in the current population.

It appears that if we know the number of generations ago when a pair of chromosomes has its CCA, then we can give a probability distribution that describes the genotype difference and the distance between these chromosomes. Thus it is possible to give a probability distribution of the distance between arbitrary pair of chromosomes in the

base population, and the population diversity is the expected value of that distribution. Additionally, the takeover time is the expected value of the number of generations back to the CCA for a randomly chosen pair of chromosomes from the base population.

The paper is organized in the following way. In Section 2 we make a theoretical discussion of the relation between the genealogy of individuals and the population diversity in arbitrary Banach space. Section 3 presents the theoretical results specific for the binary space. In Section 4 we give a simulation example in the binary space aimed at illustrating the prediction possibility of the theory-based results. Section 5 concludes the paper.

2 Genealogy of individuals and the population diversity

Consider a generational evolutionary algorithm with selection and mutation only. Assume that EA operates in a measurable Banach space (e.g. R^n or in the space of n dimensional binary vectors $\{0,1\}^n$), and the distance between two vectors is denoted by $||\mathbf{x}-\mathbf{y}||$ (e.g. Euclidean or Hamming distance), where $||\mathbf{a}||$ denotes the norm of a vector $\mathbf{a}$. Consider an arbitrary pair of individuals, $\mathbf{x}$ and $\mathbf{y}$ contained in the same base population $\mathbf{P}^t$. Assuming that $\mathbf{x}$ and $\mathbf{y}$ are picked from $\mathbf{P}^t$ at random with the uniform distribution, we conclude that their distance $d = ||\mathbf{x}-\mathbf{y}||$ is a random value with an unknown distribution. The aim of this section is to give formula for the probability distribution $P_d(d|\mathbf{P}^t)$. Note that P_d gives us information about the distribution of individuals from $\mathbf{P}^t$ in the search space, and its expected value can serve as a diversity measure.

Consider the genealogy tree defined in the following way. Each node of the tree is an individual. The tree is directed, and the edge leading from the individual $\mathbf{x}$ to the individual $\mathbf{y}$ means that $\mathbf{y}$ is the result of mutating $\mathbf{x}$. An example of a genealogy tree is depicted in 1. Nodes have been organized in a "layered" manner — individuals from the same base population are group in a column, allowing for easy identification of base populations in consecutive generations.

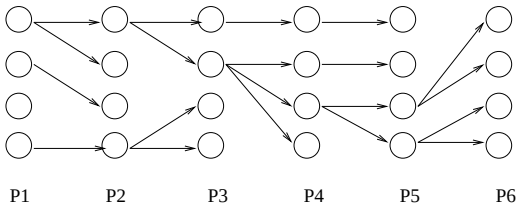

Figure 1. Example of a genealogy tree. Nodes correspond to individuals, and a link shows that an individual is the result of mutating another individual.

We can derive $P_d(d|\mathbf{P}^t)$ from the distribution describing the mutation process and the genealogy tree. Consider a pair of chromosomes $\mathbf{x}, \mathbf{y} \in \mathbf{P}^t$ and assume that k generations before (i.e. in the population $\mathbf{P}^{t-k}$) there was an individual $\mathbf{z}$ who was the common ancestor of $\mathbf{x}$ and $\mathbf{y}$, and that in generations $t-k+1, ..., t-1$ there was no common ancestor of $\mathbf{x}$ and $\mathbf{y}$. Then $\mathbf{z}$ is the CCA for $\mathbf{x}$ and $\mathbf{y}$. Note that it is possible that in generations $t-k+1, ..., t-1$ there was at least one individual $\mathbf{v}$ with the same genotype as $\mathbf{z}$, but $\mathbf{v}$ is still not a common ancestor for $\mathbf{x}$ and $\mathbf{y}$, since we are interested

in the genealogy tree rather than in the similarity in a genotype space. Assume that the mutation acts as follows

$$\mathbf{x}' = \mathbf{x} + \mathbf{m} \tag{1}$$

where $\mathbf{x}, \mathbf{x}'$ are the mutated chromosome and its result, respectively, $\mathbf{m}$ is the value of the random variable describing the mutation process, and "+" is the addition operator specific for the search space (e.g. sum of vectors in R^n or exclusive or in $\{0,1\}^n$). Then $\mathbf{x}$ and $\mathbf{y}$ differ from $\mathbf{z}$ in sum of k vectors being independently driven from the random variable describing the mutation process. Denote the mutation distribution by P_m.

According to (1) the distribution of both $\mathbf{x} - \mathbf{z}$ and $\mathbf{y} - \mathbf{z}$ is given by the k-fold autoconvolution of the mutation distribution. If we assume that P_m is symmetrical, i.e. $P_m(\mathbf{x}) = P_m(-\mathbf{x})$, then the distribution of differences $\mathbf{d} = \mathbf{x} - \mathbf{y}$ is $2k$-fold autoconvolution of P_m

$$P_1(\mathbf{d}|k) = \underbrace{P_m * ... * P_m}_{2k}(\mathbf{d}) \tag{2}$$

We have expressed P_1 as a conditional distribution to stress its dependence on the value of k.

The distribution of differences $P_1(\mathbf{d}|k)$ uniquely defines the distribution of distances $P_2(d|k)$, where $d = ||\mathbf{d}||$. If we denote $P_a(k|\mathbf{P}^t)$ the ancestry distribution — the probability distribution that a pair of vectors in $\mathbf{P}^t$ comes from the CCA originated k generations before — then we end up with the formula for probability distribution $P_d(d|\mathbf{P}^t)$ of the distance between pair of points from $\mathbf{P}^t$

$$P_d(d|\mathbf{P}^t) = \sum_{k=1}^{t} P_2(d|k) P_a(k|\mathbf{P}^t) \tag{3}$$

The actual population diversity in the generation t is the sample of the random variable given by $P_d(d|\mathbf{P}^t)$. Its expected value equals linear combination of expected difference values for each value of k

$$E[P_d(d|\mathbf{P}^t)] = \sum_{k=1}^{t} E[P_2(d|k)] P_a(k|\mathbf{P}^t) \tag{4}$$

This means that the expected value is in the range between the smallest and the largest value of $E[P_2(d|k)]$.

3 Diversity in binary space

In the binary space chromosomes are binary vectors of the length n. We will assume that the distance is measured with the Hamming metric which equals the number of bits by which two vectors differ. We will also assume that the mutation consists in changing each bit with equal probability p_m. If we consider two vectors with CCA originated k generations before, then the distance between them will be a random number whose distribution dependends on k. Let us define this distribution.

Consider binary vectors $\mathbf{x}, \mathbf{y}$ and the vector $\mathbf{z}$ being their CCA located k generations before. Consider an i-th bit and the number of changes that have been made on that bit

when getting x_i from z_i and y_i from z_i. According to (2) we can sum up the number of changes and compute its probability as

$$p(l|k) = \left(\begin{array}{c} 2k \\ l \end{array} \right) p_m^l (1-p_m)^{2k-l} \tag{5}$$

where p_m is the probability that a bit is changed in one generation (mutation probability), and l is the sum of the number of chages on the way from z_i to x_i and from z_i to y_i. If the number of changes is odd we get $x_i \neq y_i$, and

$$p_n(k) = Prob(x_i \neq y_i) = \sum_{l=0}^{k-1} \left(\begin{array}{c} 2k \\ 2l+1 \end{array} \right) p_m^{(2l+1)} (1-p_m)^{(2k-2l-1)} \tag{6}$$

Note that

$$2p_n(k) = [p_m + (1-p_m)]^{2k} - [p_m - (1-p_m)]^{2k} \tag{7}$$

so

$$p_n(k) = \frac{1}{2} \left[1 - (2p_m - 1)^{2k} \right] \tag{8}$$

When vectors of n bits are considered, the distance between them is binomially distributed. The distribution of the distance $d = ||\mathbf{x} - \mathbf{y}||$ between two binary vectors with the CCA located k generations before is given by

$$P_2(d|k) = \left(\begin{array}{c} n \\ d \end{array} \right) [p_n(k)]^d [1 - p_n(k)]^{n-d} \tag{9}$$

Note that the expected value of the distance grows with the number of dimensions n.

4 Genealogy trees in practice

Prediction of the population diversity would be possible if we knew the $P_a(k|\mathbf{P}^t)$ distribution. Unfortunately, when proportionate selection is used, P_a depends directly on the fitness function values which are unknown in advance before starting the evolution. Therefore we present the experimental results rather than analytical analysis of the P_a distribution.

We have performed experiments for binary chromosomes. The experiments were aimed at investigating the P_a distribution when the reproduction was fitness proportionate. At the same time we were observing the population diversity and tried to compare the results with the values predicted according to sections 2 and 3.

4.1 Test problem

The problem formulation comes from applying a penalty function to the knapsack problem. Consider n-dimensional vectors $\mathbf{p}$ and $\mathbf{w}$ of positive real numbers, and the positive real scalar W. The fitness function to be maximized equals:

$$f_k(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i - K \max\left\{0, \sum_{i=1}^{n} w_i x_i - W\right\} \tag{10}$$

where:
$K = \max_{i=1,\ldots,n} p_i/w_i$,
$\mathbf{x} = \{0,1\}^n$,
p_i is the profit of the item i drawn with uniform distribution from $(0, 50)$,
w_i is the weight of the item i drawn with uniform distribution from $(0, 50)$,
$W = \frac{1}{2}\sum_{i=1}^{n} w_i$ is the knapsack capacity.

4.2 Empirical results for ancestry histograms

Evolutionary algorithm We performed tests using an evolutionary algorithm with fitness proportionate nonelitist selection, without crossover. Mutation consists in flipping each bit with probability p_m.

Approximation of ancestry histograms with gamma-distribution. In the first experiment we analyzed the ancestry histogram of a base population. The histogram for generation t is obtained by analyzing all chromosome pairs from $\mathbf{P}^t$ and by recording how frequently a pair of chromosomes has its CCA in population $\mathbf{P}^{t-k}$. We recorded cumulated ancestry histograms (CAH) obtained by averaging ancestry histograms for 100 successive EA generations. We observed that CAH can be approximated with the gamma-distribution, and this holds for all fitness functions under consideration.

The probability density function of the gamma-distribution is given by:

$$PDF(x, k, \Theta) = x^{k-1}\frac{e^{-x/\Theta}}{\Theta^k \Gamma(k)} \text{ for } x > 0 \tag{11}$$

where:
$k > 0$ is the shape parameter,
$\Theta > 0$ is the scale parameter.
Note that larger values of Θ result in a flatter shape for the PDF. Mean value of the gamma-distribution is $k\Theta$, and the variance equals $k\Theta^2$.

An example CAH and its approximation with the gamma-distribution (11) are depicted in Fig. 2.

Moreover we observed that the parameters of the distribution to approximate CAH did not change much over generations. To illustrate this effect let us study Fig. 3 where the parameters of the gamma-distribution estimated for CAH are depicted for successive generations for a single run of an EA. The population size was $\mu = 50$, the fitness was f_1, and the mutation probability was $p_m = 0.005$. The cumulated ancestry histograms were also only slightly different for different EA independent runs. To illustrate this 25 independent runs of the EA with the same settings were performed, and for each run

the CAH were approximated with gamma-distributions. We recorded the population of distribution parameter values. Mean value of the parameters were $\Theta = 9.61$, $k = 1.80$, and the standard deviations of these values were 1.33 and 0.23. We conclude that the cumulated ancestry histograms are an fair approximation to the P_a distribution.

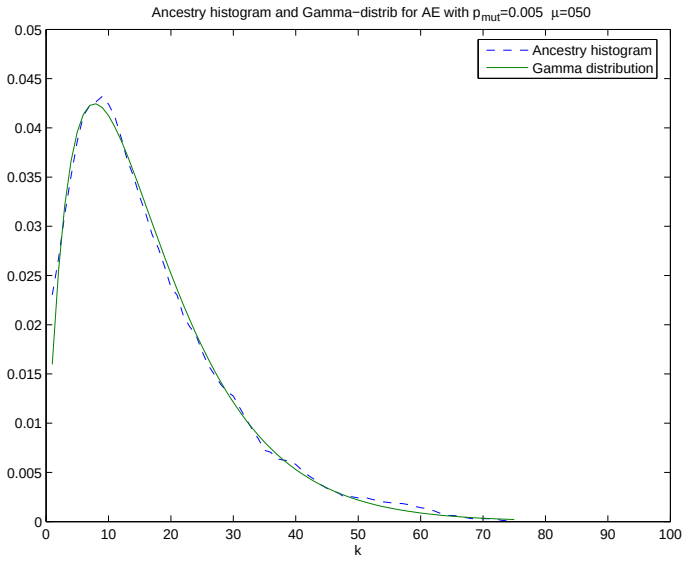

Figure 2. Ancestry histogram and its gamma-distribution approximation

Influence of population size and mutation range. When considering the gamma-distribution approximation to CAH we observed that the distribution parameters depended on various factors such as the fitness function, the mutation probability p_m and the population size. This dependence can be explained by the fact that the aforementioned factors influence the distribution of the fitness over chromosomes in the base population, and this results in various ancestry histograms.

It is well known that smaller populations are more easily dominated. Therefore with increasing population size the ancestry histograms are getting "flatter". This effect can be observed even with the constant fitness function. Larger populations means that chromosomes may have their CCA deeper in the history, that is, they may result from more mutations and the population diversity is larger. This means that the population occupies an area of a larger diameter (in the sense of the number of mutations), so the diameter is also influenced by the mutation range. Therefore, the distribution of fitness values, and in the same time the P_a distribution, depends on the fitness "shape" in the area occupied by the base population. If the fitness is highly variable within that area, the probability for reproducing individuals is highly differentiated and the P_a distribution is relatively narrow-shaped. If the fitness is slightly variable (or is constant) the P_a distribution is "flatter".

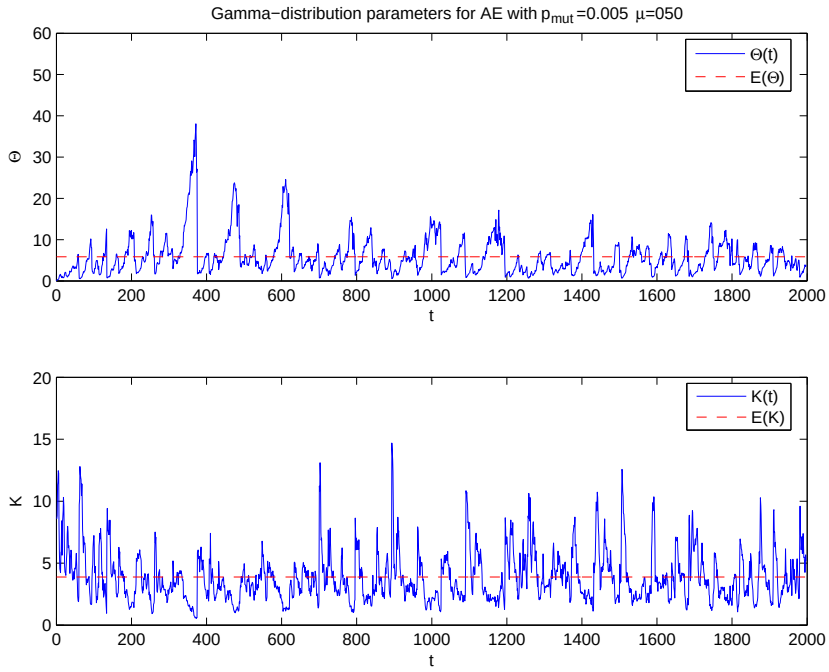

Figure 3. Parameters of gamma-distribution (single run) for different t [solid line]. Mean value of data presented as solid line [dashed line].

Ancestry histograms and takeover time. Note that the expected values of the CAH are the takeover time values. It is reported [2] that the dependence of the takeover time on the population size can be estimated by $o(n \log n)$. In our experiments we observed (Fig. 4) that the dependence population size can be estimated rather by $o(n \log^2 n)$.

4.3 Ancestry histogram and population diversity

Typical dynamics of population diversity is depicted in 5. It can be observed that after a relatively small number of generations the population diversity tends to oscillate around a certain value. This effect can be explained by the fact that the ancestry histograms tend to stay "stable". We can verify if population diversity can be predicted by using genealogy analysis.

As stated in Section 2, when we know the ancestry distribution, it is possible to estimate the population diversity using the formula (3). In Table 1 we give the diversity values computed according to formula (3) and the empirical diversity values observed in 150 independent runs of an EA. The latter value was computed for each EA run as the average diveristy of populations starting from the generation number 500 up to 2000. We also report the estimation percentage error computed as

$$e = 100(d_t - d_a)/d_a \tag{12}$$

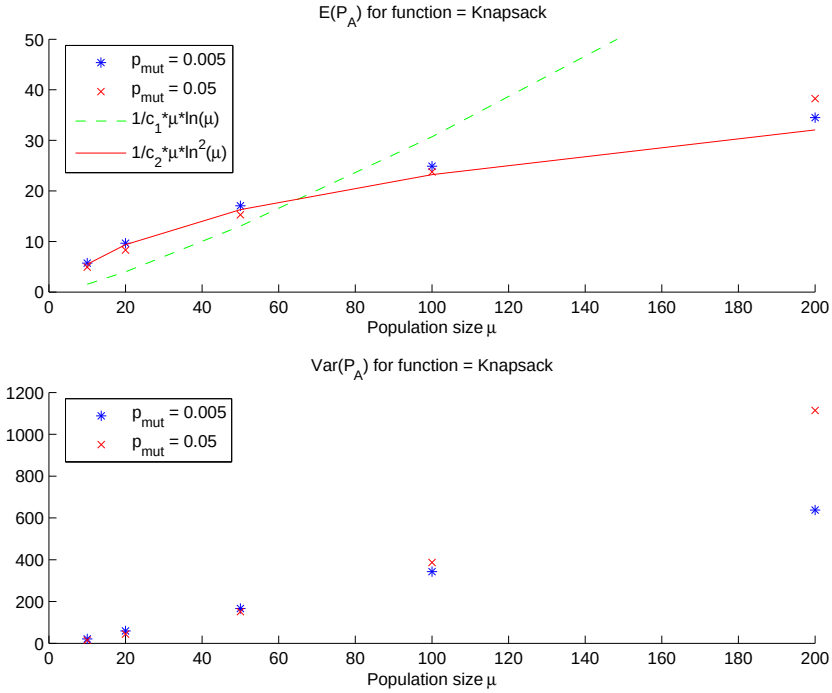

Figure 4. Expected value and variance of P_a for the function f_k and for $p_m \in \{0.005, 0.05\}$. Regression curve parameters are $c_1 = 15$ and $c_2 = 50$.

where e is the estimation percentage error, d_t and d_a are the theoretical and actual values of population diversity.

We can observe that the percentage error in most of the test cases is kept at a reasonable level. For large populations and small mutation probability it grows significantly, and we explain this effect by the fact that we use cumulated ancestry histograms instead of current histograms.

5 Summary and conclusions

In this paper we give formulas for estimating the population diversity. We give the evidence that the diversity is dependent mainly on the parameter values for the evolutionary algorithm. Thus it is possible to control the diversity by setting the population size and the mutation range. We explain the diversity by the genealogy tree and describe the tree properties with ancestry histograms. We observe that these histograms can be approximated well with the gamma-distribution, and the distribution parameters may thus be used to characterize the properties of the genealogy tree.

We expect that this work will give another dimension to the discussion on "what is the optimum population size for an EA". We also hope that it is possible to make steps towards better understanding which optimization problems are well suited for an EA.

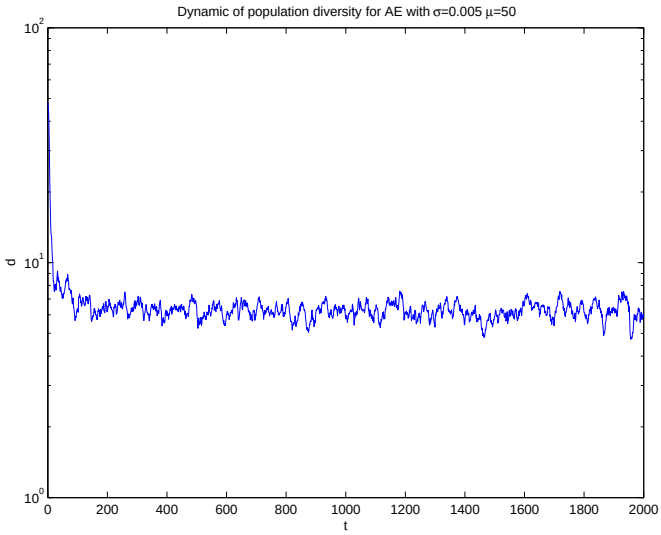

Figure 5. Typical dynamics of the population diversity in subsequent generations.

EA parameters		Diversity values		
p_m	μ	predicted	observed	error(%)
0.005	10	2.74	2.92	-6.61
	20	4.49	4.37	2.82
	50	7.54	6.48	14.02
	100	10.39	8.19	21.22
	200	13.52	9.83	27.32
0.05	10	17.83	26.66	-49.54
	20	24.73	33.23	-34.38
	50	33.02	39.60	-19.94
	100	38.06	42.81	-12.49
	200	41.27	44.94	-8.84

Table 1. Predicted and observed diversity values together with the estimation error for various μ and p_m values.

The answer could be, in the context of the presented results, that between each pair of local maxima there exists a path with its length no greater that a certain characteristic value related to the population diversity. This is our meaning of the Eigen's metaphor of mountain chains in the search space cited in [4]. A somehow "negative" effect of our research is that the diversity *will* be reduced to a certain level implied by the population size and the mutation range unless some tricks like niching are used to keep it on a

higher level. In our opinion the presented results support to some extent the EA model presented in [3] who represented the whole population as a single point in the search space surrounded by a "balloon" which contains the contents of the population. With this paper we give a feeling about the diameter of this balloon, however we do not claim it is symmetrical around a centerpoint.

This paper is an early report from ongoing research. We are working in parallel on the floating point representations. We also recognize that the problems with the estimation of the ancestry histograms, which cannot be avoided for the fitness proportionate selection, can be easily solved by applying e.g. the tournament or rank based selection. In the aforementioned selection methods the distribution of the fitness probabilities is either predetermined by the unique rank values (and thus will not change over generations) or is flattened by the tournament procedure. We expect to get much better diversity estimations for those selection schemes and plan to verify that in the near future.

Acknowledgements The authors gratefully acknowledge the contribution of Mr. Marek Rupniewski to the theoretical part of this paper.

Bibliography

[1] T. Bäck, F. Hoffmeister, and H.-P. Schwefel. A survey of evolution strategies. In *Proc. of the Fourth International Conference on Genetic Algorithms*, pages 2–9, San Diego, CA, 1991.

[2] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, 1996.

[3] Artur Chorążyczewski. On the adaptive provess of the aggregated model of evolution. In *KAEiOG'04*, pages 27–32, 2004.

[4] Roman Galar. *Soft selection in random global adaptation in* R^n, volume 84 of *17*. Wydawnictwo Politechniki Wrocławskiej, Wrocław, 1990.

[5] David Goldberg, Kalyanmoy Deb, and B. Korb. Don't worry, be messy. In *ICGA'91*, pages 24–30, 1991.

[6] Zbigniew Walczak. Graph-based analysis of evolutionary algorithm — prelimiray results. In *KAEiOG'04*, pages 187–192, 2004.