

On fitness distance distributions and correlations, GA performance, and population size of fitness functions with translated optima

Janne Koljonen

Department of Electrical Engineering and Automation, University of Vaasa
P.O. Box 700, FIN-65101, Vaasa, Finland
jako@uwasa.fi

Abstract

Fitness distance correlation has been proposed as an analysis tool for the fitness function landscape of genetic algorithms. It has been shown that it works in some cases, but there are also counter-examples. We studied the fitness distance correlation and the optimal population size by two fitness functions with a translational minimum and with a meta-genetic algorithm. We discovered that the fitness distance distribution and the actual optimization performance depend strongly on the place of the optimum. We also found that fitness distance correlation explains moderately the variations in the performance of the genetic algorithm and also some of the variations in the optimal population size obtained by a meta-genetic algorithm.

Keywords: elitism, fitness distance correlation, meta-GA, optimal GA, population size.

1 Introduction

A major effort has been given to solve when evolutionary computing, i.e. evolutionary strategies (ES), evolutionary programming (EP) and genetic algorithm (GA), will be effective at finding the global optimum. The problem is challenging due to the nonlinear and stochastic nature of the algorithms, and because evolutionary algorithms is actually a family of algorithms, from which the applicator selects a proper one with appropriate parameters and operators according to the application.

There are two separate classes to approach the performance analysis: one that tries to solve the efficiency without actually running the algorithm and without knowing the solution, and another one that analyzes the performance after first obtaining the solution. The first approach would be very useful if it could be done with less effort than the runs of the algorithm would require, but it is probably harder to achieve an effective method to that than to the second case.

The methods to analyze and predict ES, EP and GA performance (and the hardness of the objective function determining, together with the operators etc, the performance of the algorithm) developed so far are based on the building block theory (Forrest and Mitchell, 1993; Goldberg, 1989), multi-

modality (Rana, 1998), fitness distributions (Borenstein and Poli, 2004; Popovici and De Jong, 2003), fitness distance correlation (Jones and Forrest, 1995; Vanneschi and Tomassini, 2002), and several others, too. Each of them can explain some phenomena but, unfortunately, not all. There won't probably be any single method that would explain every case and hence this many are needed.

The main problem of some of them is that they don't take into account the genotype-phenotype mapping of the problem and the genetic operators themselves that actually define the dynamics of the evolutionary algorithm. Modifying or changing them can change the problem completely. The analysis tools should therefore be based on the operators themselves (Altenberg, 1997). There are also analysis methods that take some of the dynamic properties into account (See e.g. Syswerda, 1989). With fixed operators a static analysis of the fitness distance correlation can though be adequate in some cases, as will be shown in this paper.

2 Fitness distance distributions

Genetic algorithms work with the aid of the information given by the fitness function. The usefulness of this information depends on the genotype-phenotype mapping, fitness landscape, and the genetic opera-

tors. The fitness function works on phenotype while the GA itself on genotype. This causes that the fitness function does not directly tell, how close to a solution the GA is. It can even be misleading the algorithm away from the solution, in terms of the Hamming distance of the binary genotype representation to the optimum.

If the global optimum (or optima) is known, the Hamming distance to the optimum (or to the nearest global optimum) and fitness value of a genetic individual can be computed. If possible and sensible, in terms of computational efforts, we can calculate distances and fitness values of all possible genotypes to obtain a fitness distance distribution (FDD); otherwise a random sample should be obtained. Jones and Forrest (1995) have plotted Hamming distances against fitness values as scatter plots. In our opinion, 3-dimensional frequency plots are in many cases more illustrative.

Generally, with some exceptions (Altenberg, 1997), the regular GA is effective if the Hamming distance becomes smaller as the fitness becomes better, be it higher or lower whether we are maximizing or minimizing. In this paper, we are always minimizing and at the only optima of the test problems the fitness function attains 0.

The exceptions in (Altenberg, 1997) are obtained using single-point crossover. Uniform crossover (Syswerda, 1989) is probably the easiest crossover operation to handle theoretically and in many cases the most efficient one. With uniform crossover it is easily seen that crossing two individuals with low Hamming distances with respect to the optimum has a better chance to produce an offspring representing a solution than individuals of higher Hamming distances. Individuals with low distance may not necessarily have low fitness values, as can be seen e.g. in Figure 3. On the other hand, there may be genotypes with low distance and high fitness that are blocked from the elite as well as genotypes with high distance and low fitness i.e. *false friends*. As the GA tries to minimize the fitness in the latter case, it may on the contrary create more individuals false friends thus lowering the chance to obtain a solution.

The properties of the fitness-distance dependence and distribution can be examined with 3D frequency plots or scatter plots. Jones and Forrest have suggested (1995) that this information could be summarized into the Pearson correlation coefficient r , called fitness distance correlation (FDC):

$$r = \frac{c_{FD}}{s_F s_D}, \quad (1)$$

where c_{FD} is the covariance of the samples of the fitness and distance pairs, and s_F and s_D are the stan-

dard deviations of the fitness and distance, respectively.

The fitness distance correlation has been used at least for genetic algorithms (Jones and Forrest, 1995; Altenberg, 1997) and genetic programming (Vanneschi *et al*, 2002; Vanneschi and Tomassini, 2003). It has been shown by some counter-examples that FDC is not always a good estimate of problem difficulty. FDC has been mainly tested on the common benchmark problems and with other not particularly interrelated problems. In this paper, we study families of closely interrelated problems with FDC and compare FDC to the GA performance.

3 Genetic algorithm

Genetic algorithms is a stochastic optimization method effective for difficult and nonlinear problems (Alander, 1994; Alander, 2002). We used a GA with two genetic operators: uniform crossover (Syswerda, 1989) and bitwise mutation. The GA had two populations: one with mature individuals, also known as elite, from which the parents were randomly drawn, and a population of immature individuals into which the offspring are temporarily placed after creation. To refer to the size of the populations we use symbols n_e and n_i , respectively. Hence, the total population size is $n_e + n_i$. By relative elitism we mean $n_e / (n_e + n_i)$. n_{max} denotes for the maximum number of trials, inclusive the original population, allowed for the GA to do.

Each individual in the elite had an equal probability to be selected as a parent, but the parents in the crossover had to be separate to avoid clones. We used natural binary coding in genotype-phenotype mapping.

4 Test problems

To study the FDC as a GA performance predictor we formed families of problems that have a transla-

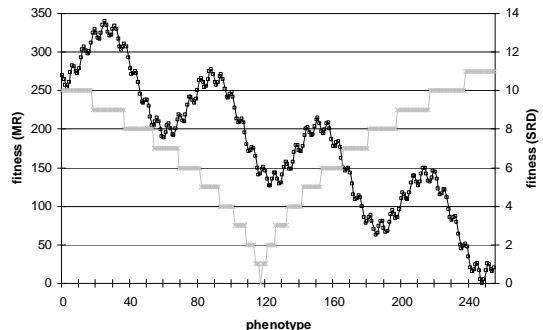


Figure 1: Fitness landscape of the 8-bit MD₂₄₈ (black rectangles) and SRD₁₁₇ (gray crosses).

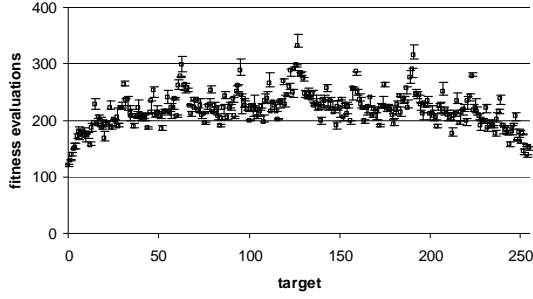


Figure 2: Minima, medians and maxima of numbers of evaluations of three runs that averaged 1000 GA runs ($n_{max} = 800$) on each target of SRD. The population sizes were fixed ($n_e = 32, n_i = 16$).

tional optimum. The first fitness function (FF) was the square-root of the distance (SRD) to the target rounded down to the nearest integer. By changing the target the problems (denoted by subscripts, e.g. SRD₀) the fitness landscape changes. The fitness function is thus:

$$fitness = \lfloor \sqrt{|phenotype - target|} \rfloor. \quad (2)$$

To validate the results another test function (modified Rastrigin, MR) with a multimodal phenotype landscape and a translated optimum was used:

$$fitness = \lfloor |x-t| - 10 \cos(x-t) - 50 \cos((x-t)/10) + 60 \rfloor. \quad (3)$$

where x is the phenotype presentation of the bit vector and t is the target, where $fitness = 0$.

4.1 Fitness landscape

The fitness landscapes (phenotype) of the test functions are shown in Figure 1, where 8 bits are used.

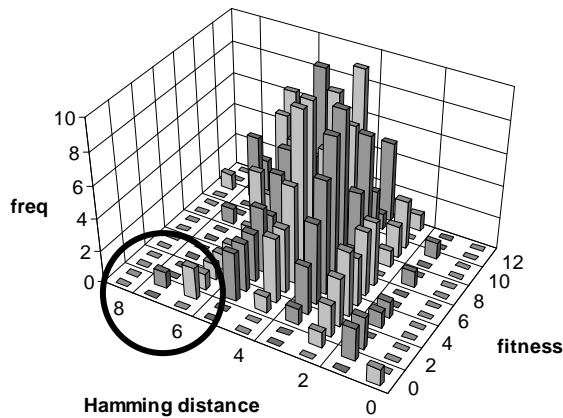


Figure 4: Fitness distance distribution of one of the hardest problems of FF the family (SRD₆₃). There are samples whose fitness is 1 but the Hamming distance large (in the circle).

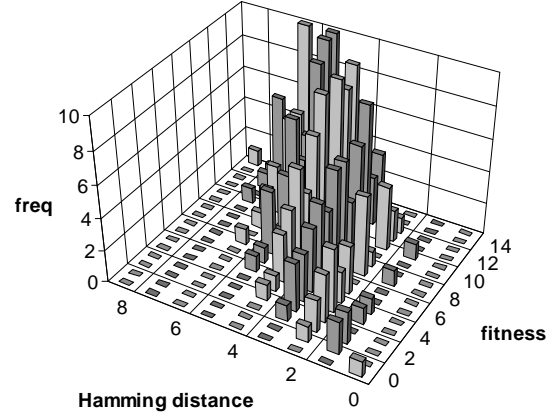


Figure 3: Fitness distance distribution of the SRD₀. This problem was the GA easiest of the fitness function family.

With a phenotype search strategy the SRD problems seems to be quite easy, but the MRs do not.

4.2 Fitness distance distributions

We started to study the fitness functions running the GA with all possible target values; this was considered reasonable with 8-bit fitness functions. The results for SRD are shown in Figure 2. The figures for MRs (not shown) look essentially the same.

With these preliminary runs we found that certain targets differ significantly from others. The easiest ones were near the limits of the range; 0 was the easiest. The most difficult ones were the following targets: 63, 95, 126, 127, and 191. Each of them has false friends. E.g. if the solution is $127_{10} = 01111111_2$, then $128_{10} = 10000000_2$ is clearly a false friend, whose Hamming distance to 127 is 8, but whose fitness is low with continues fitness functions.

Let us examine the fitness distance distributions of the fitness function with some target values. We

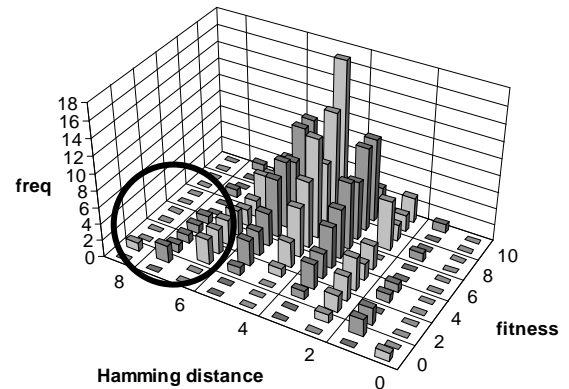


Figure 5: The hardest problem of the ff family (SRD₁₂₇) has a flatter distribution and many false friends (circled) predicting poor evolvability.

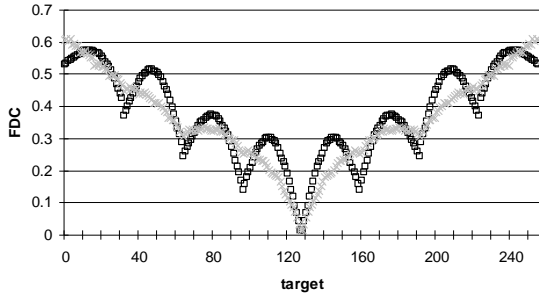


Figure 6: Fitness distance correlation with different targets. Black rectangles = MR, gray crosses = SRD.

can then compare the distributions of the GA hard ones to the easier ones.

The distribution in Figure 3 belongs to the easiest one, as for our GA. It has a clear positive correlation between the Hamming distance and fitness and several low distance low fitness genotypes, from which the solution is easily obtained by crossover.

The FDD in Figure 4 is in turn one of the GA hardest ones. The correlation seems quite strong, may it be less that in Figure 3, but probably the several false friends not existing in Figure 3 cause the GA sometimes to be stuck.

The hardest fitness function of the family was when target was 127. Its fitness distance distribution shows low correlation as well as many false friends. Yet SRD_{63} (Figure 4) was nearly as difficult as SRD_{127} even though its FDD looks much easier compared to the very flat and uncorrelated one in Figure 5.

4.3 Fitness distance correlation

We calculated all FDCs of the fitness function families (Figure 6) to find out how well they predict the GA performance. The results in Figure 6 confirm the conclusions made on the basis of fitness distance distributions, as for variation in FDCs.

In Figure 7 the GA performances of SRD func-

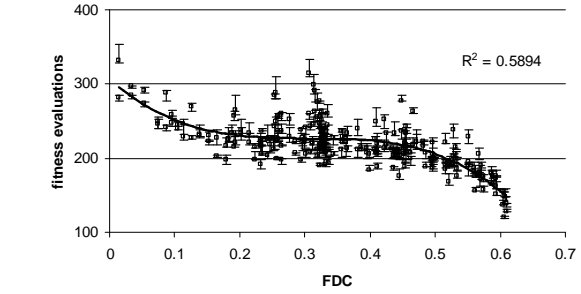
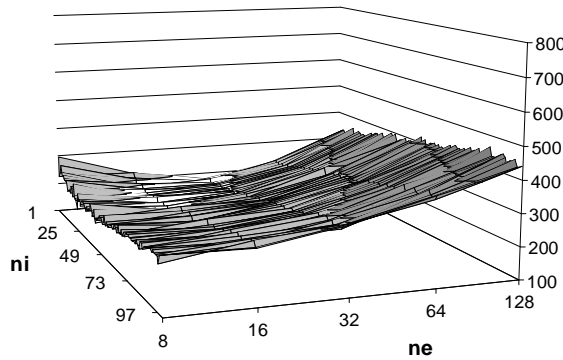


Figure 7: Correlation between the GA hardness and the FDC is moderate: with a 3rd degree polynomial the $R^2 = 59\%$. The population sizes and other parameters were fixed ($n_e = 32$, $n_i = 16$). Minima, medians and maxima of means of three repetitions of 1000 GA runs using SRD functions.

tions (as in Figure 2) are plotted against the FDCs. The plot shows only moderate correlation ($R^2 = 59\%$) and some outliers, so the FDC seems to be inadequate predictor. For MR functions $R^2 = 40\%$. The reason (as will be shown) is that the population sizes (and other parameters, too) are kept fixed when changing the target.

5 Optimization speed and population size

It was presumed that different target values of the fitness function families require different population sizes as the FDC and GA hardness seemed to vary significantly. We tested this hypothesis by changing the population size parameters and running the GA 200 times. This time we used the 10-bit SRD function. The maximum number of new individuals was limited to 800.

5.1 Results and discussion

Running the GA with the easiest and the hardest problem of the SRD family gives a clear indication

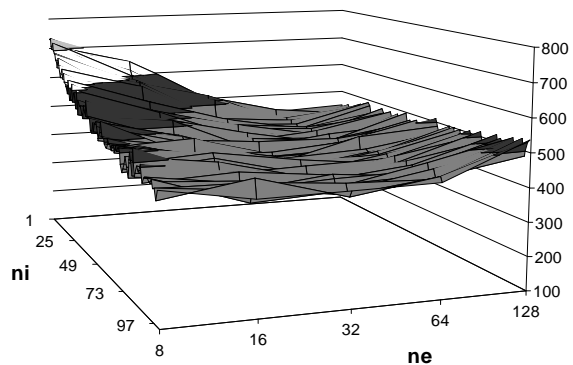


Figure 8: Mean optimization speed vs. elite and immature population sizes with SRD_0 (left) and SRD_{127} (right).

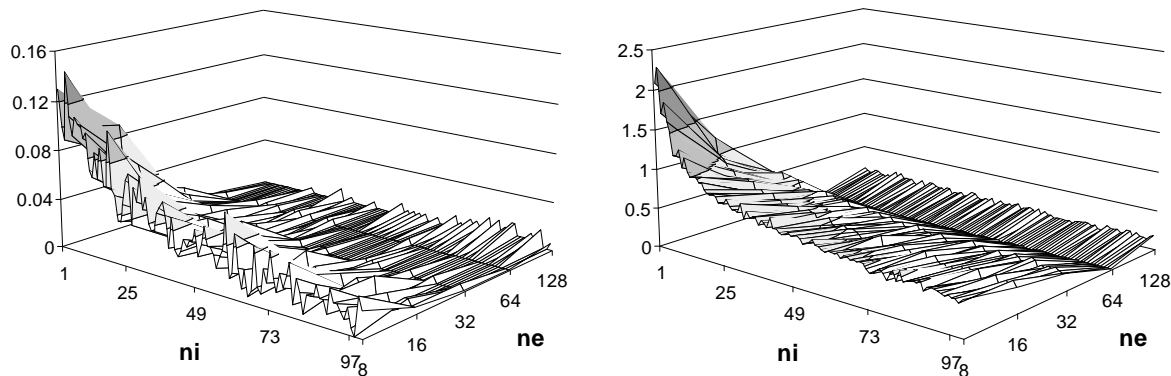


Figure 9: Mean of the best fitness of 200 runs vs. population sizes with SRD_0 (left) and SRD_{127} (right).

how the population size must be tuned individually for each problem to get the optimal GA performance. Figure 8 depicts the mean GA performance vs. population sizes for the easiest and hardest SRDs. By comparison it is evident that the optimal population size increases when having harder problems. Moreover, if the population sizes were kept constant, depending which values were selected, the performance difference between targets varies.

6 Optimization reliability

The optimization speed is not everything but the reliability of finding a solution is also important. We measured the reliability by recording the best fitness of each run.

6.1 Results and discussion

We ran the easiest and hardest SRDs (Figure 9.) and results look very similar: the larger the population the better the reliability. However, the mean of the best fitness is over ten times worse for SRD_{127} than for SRD_0 , particularly with small populations.

7 Meta GA optimization

It seemed that fitness distance correlation was not very efficient in GA performance prediction. Our hypothesis was that this was due to non-optimal GA parameters, particularly population size, whose optima were shown to differ a lot from target to target. To continue studying FDC we optimized the population size parameters with a meta-GA.

7.1 Method

The optimization of GA parameters can be accomplished by another GA. These parameters include crossover type, mutation rate, population size etc. In

other words, the parameter selection problem is formulated as another optimization problem (Mansfield, 1990; Alander, 2002; Bäck, 2002). It has been reported that the GA can optimize its own parameters, too (Hartono, 2004).

In meta-GA the individuals are GAs that solve the actual problem (See Figure 10). The fitness function is defined by measuring the GA performance: it could be for instance the number of fitness function calls required to find a solution. Each fitness function evaluation of a given meta-GA trial may give different results due to the stochastic nature of GAs. Thus the mutual fitness order of the meta-GA individuals in elite may change during optimization.

7.2 Results and discussion

We used a meta-GA to optimize n_e and n_i , whose ranges were determined so that the optimal values obtained from Figure 8 were included. The meta-GA parameters were fixed: $n_{max} = 100$, $n_e = 32$, and $n_i = 4$. The fitness function of the meta-GA was the number fitness function calls averaged over 100 GA runs using the same parameters. The actual GAs used the 8-bit SRDs or MRs and had $n_{max} = 800$.

The meta-GA was run three times with every 256 possible fitness functions. The fitness values of the

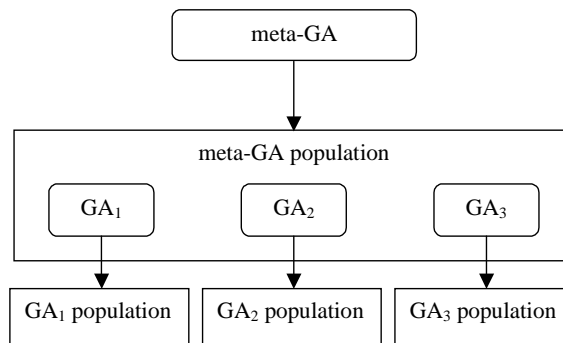


Figure 10: Principle of a meta-GA. Meta-GA optimizes the parameters of the actual GAs.

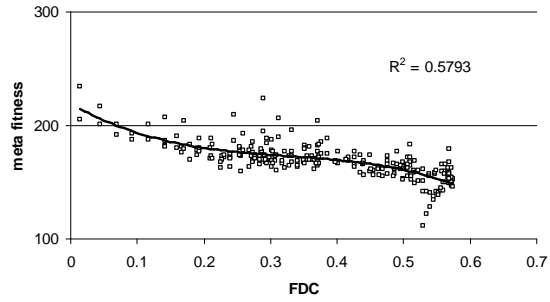
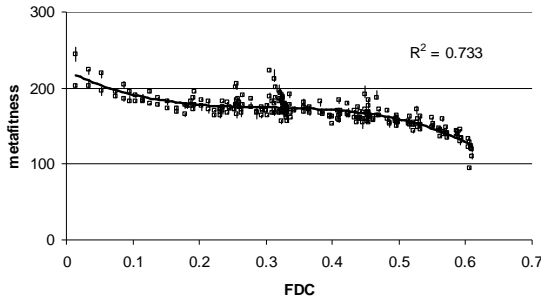


Figure 11: The mean best fitness of the last ten generations of the meta-GA chromosome vs. FDC for SRD (left) and MR (right). The 3rd degree polynomials explain 73 % and 58 % of the variation for SRD and MR, respectively (Compare the SRD result to 59 % with fixed population size in Figure 7). Min, med and max of three runs shown for SRD.

best meta-GA individuals are plotted against the corresponding FDC in Figure 11 (left for SRD, right for MR functions). The deviations of the three runs are also shown. The R^2 has risen from 59 % to 70 % by optimizing the population size.

Because the areas of optimal population sizes are relatively flat (See Figure 8) and the meta-GA fitness function is stochastic, there is a lot of internal variation in the population size. To overcome this we decided to use the average of the best chromosomes of the last ten generations of the meta-GA. By taking averages of the fitness value, elite size and immature population size, the stochastic variation was decreased.

Now the mean of the best fitness of the last ten meta-GA generations correlates a bit stronger with the FDC than when using only the best individual of the last meta-GA generation (Figure 11). The R^2 of the SRD and MR functions have risen from 59 % to 73 % and from 40 % to a reasonable 58 %, respectively.

A bigger improvement is obtained in the correlation between the optimal total population size (Figure 12) and the FDC for SRD function. However,

the deviation of the optimal population size vs. FDC indicates statistical insignificance. Also total population size and meta-GA fitness correlate showing the well-known fact (Alander, 2002) that the smaller the population size, the faster the GA, unless the population is too small. The R^2 for population size vs. FDC are 43 % for SRD and 14 % for MR. The R^2 for population size vs. metafitness are in turn 65 % and 53 %, respectively.

8 Conclusions

One can usually predict the evolvability of a genetic algorithm by studying fitness distance distribution, despite some exceptions. The information can be compressed into fitness distance correlation (FDC). However, the method does not take into account the dynamics of the GA and there are counter-examples that show that low correlation does not necessarily imply poor GA performance.

We studied the ability to predict GA performance by FDC with two test functions. By translating the global minimum of the test functions whole families of similar functions were created. It was

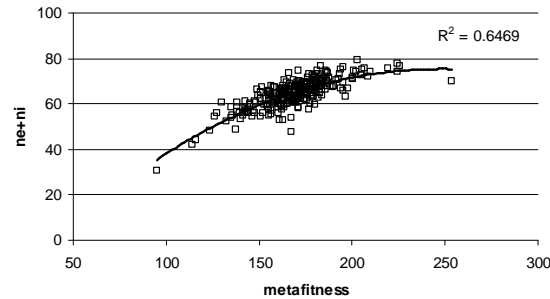
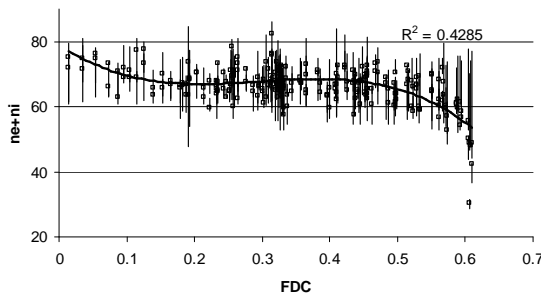


Figure 12: Correlations for SRD: Mean population size of the last ten meta-GA generations vs. FDC (left). R^2 for median runs is 43 %. Also the mean best fitness of the last ten meta-GA generations (right, deviations not shown) is correlated with the population size ($R^2 = 65 %$).

discovered that translation of minimum significantly changes fitness distance correlation and GA hardness. Fitness distance correlation was found a moderate predictor of the GA optimization speed when population size was fixed, but when optimizing the population size with a meta-GA and taking average of the best meta-GA individuals of the last ten generations the predictions improved significantly: FDC explained 73 % and 58 % of the variations in the optimization speed of the test functions. FDC also explains some variations in the optimal population size, a fact that may be useful in practice. FDC measures only linear correlation between fitness and Hamming distance to the solution. In many cases, with strongly nonlinear fitness landscape, other correlation metrics are expected to be more appropriate. Fitness distance analysis could perhaps be useful in development of new genetic operators that were robust to translations of the optima.

Acknowledgements

The author would like to thank the Research Institute for Technology at the University of Vaasa for financial support.

References

- J.T. Alander. An Indexed Bibliography of Genetic Algorithms: Theory and Comparisons. University of Vaasa, 1994. URL: <ftp://ftp.uwasa.fi/cs/report94-1>.
- J.T. Alander. Potentials of Genetic Algorithms. TEKES, 2002. URL: <ftp://ftp.uwasa.fi/cs/report96-1/English.ps.Z>.
- L. Altenberg. Fitness distance correlation analysis: an instructive counterexample. In: *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, 1997.
- T. Bäck, D.B. Fogel, Z. Michalewicz (eds.). *Evolutionary Computation 2*, 2002. ISBN: 0750306653.
- Y. Borenstein, R. Poli. Fitness distributions and GA hardness. In: *Proceedings of PPSN04*, 2004.
- S. Forrest, M. Mitchell. Relative building block fitness and the building-block hypothesis. In: D. Whitley (ed.) *Foundations of Genetic Algorithms 2*, Morgan Kaufmann, San Mateo, 1993.
- D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Morgan Kaufmann, 1989.
- P. Hartono, S. Hashimoto S., M. Wahde. Labeled-GA with adaptive mutation rate. In: *Proc. IEEE CECI*, 2004.
- T. Jones, S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Larry Eshelman (ed.) *Proceedings of the 6th International Conference on Genetic Algorithms*, San Francisco, CA, 1995.
- R.A. Mansfield. *Genetic Algorithms*, University of Wales, 1990.
- E. Popovici, K. De Jong. Understanding EA dynamics via population fitness distributions. In: Alwyn Barry (ed.) *Workshop Program Proceedings of Genetic and Evolutionary Computation Conference (GECCO03)*, 2003.
- .S. Rana. Examining the role of local optima and schema processing in genetic search. PhD thesis, Colorado State University, 1998.
- G. Syswerda. Uniform crossover in genetic algorithms. In: *Proc. of the Third International Conference on Genetic Algorithms*, 1989.
- L.Vanneschi, M. Tomassini. A study on fitness distance correlation and program difficulty for genetic programming. In: *Workshop Program Proceedings of Genetic and Evolutionary Computation Conference*, (GECCO02), 2002.
- L. Vanneschi, M. Tomassini. Pros and cons of fitness distance correlation in genetic programming. In: Alwyn Barry (ed.) *Workshop Program Proceedings of Genetic and Evolutionary Computation Conference (GECCO03)*, 2003.