Fitness Distance Correlation, as statistical measure of Genetic Algorithm difficulty, revisited

P. Collard, A. Gaspar, M. Clergue, C. Escazut[1]

Abstract. This paper revisits past works on fitness distance correlation (FDC) in relation to genetic algorithms (GA) performance, and puts forth evidence that this statistical measure is relevant to predict the performance of a GA.
We propose an interpretation of Hamming-distance based FDC, which takes into account the GA dynamics and the effects of crossover operator. We base this proposition on the notion of predicates developed by M. D. Vose.
The aim of this article is double since, using results obtained with the FDC, it confirms Vose's theory, and starting from this theory, it corroborates FDC as a relevant predictor measure of problem difficulty for GAs.

1 Introduction

The search of factors which have an influence on the ability of the GA to optimize functions represents an attractive subject. Such works are necessary, on the one hand because they allow to determine the context for which the biomimetic approach is relevant, and on the other hand because they allow to better understand how a GA works. Most of these studies are based on the notion of schemata. In his thesis, T. Jones [6] proposes to use a statistical measure of search difficulty: the fitness distance correlation (FDC). He says that the main point of a GA search is the relationship between fitness and distance to the optimum. He suggests to use either the scatter plots of fitness versus distance or to compute the correlation coefficient between fitness and distance. In most cases this last criterion provides relevant information on the search difficulty. This work is amazing because of the use of the Hamming distance. Jones itself admits that these results are unexpected since the distance is defined independently to the crossover operator and the dynamics of GAs. In this paper, we propose an interpretation of Hamming-distance based FDC which takes into account the crossover operator and the dynamics. This interpretation is suggested by M. D. Vose's works on predicates [7]. In other words we propose to answer the question: why is FDC a good measure of problem difficulty for GAs?
The paper is organized as follows: section 2 presents Jones' FDC and Vose's theory. Then, the following section proposes an interpretation of the Hamming distance in order to take into account the effect of crossover. We also present in this section another statistic measure, the fitness unstability correlation, that can be used to predict the performances of a GA. Finally, section 4 uses the GA dynamic to corroborate FDC as a relevant predictor measure for GA dificulty.

2 Related work

2.1 Fitness distance correlation as measure of search difficulty

Jones' idea is to quantifying the relationship between the values taken by the function and the distance to the global optimum. Given a set $\mathrm{F} = \{\mathrm{f}_1, \mathrm{f}_2, ..., \mathrm{f}_n\}$ of n individual fitnesses and the corresponding set $\mathrm{D_h} = \mathrm{d}_1, \mathrm{d}_2, ..., \mathrm{d}_n$ of n Hamming distances to the nearest global optimum, he computes this coefficient FDC, as:

$$\mathrm{FDC} = \frac{C_{\mathrm{FD}}}{\sigma_{\mathrm{F}}\sigma_{\mathrm{D}}}$$

where $C_{\mathrm{FD}} = \frac{1}{n}\sum(\mathrm{f}_i - \overline{\mathrm{F}})(\mathrm{d}_i - \overline{\mathrm{D}})$ is the covariance of F and $\mathrm{D_h}$, σ_{F} and σ_{D} are the standard deviations of F and $\mathrm{D_h}$, $\overline{\mathrm{F}}$ and $\overline{\mathrm{D}}$ are the means of F and $\mathrm{D_h}$.
If we consider maximization problems, what we assume in this paper, we should hope that fitness increases as distance decreases. With an "ideal" fitness function, FDC will therefore be equal to -1.0. The computation of FDC on a collection of representative functions of known difficulty shows that this correlation can be used to predict the performance of a GA. In this way, problems can be classified in one of these three classes Jones calls:

i. *misleading* (FDC $\geq$ 0.15) in which fitness increases with distance,
ii. *difficult* (-0.15<FDC<0.15) in which there is virtually no correlation between fitness and distance,
iii. *straightforward* (FDC$\leq$-0.15) in which fitness increases as the global optimum approaches.

Correlation is only one of the possible ways of studying the relationship between fitness and distance. In some cases FDC reveals itself to be a poor summary statistic. In such situations examining the scatter plots of fitness versus distance to the optimum is more useful.
Recent works from C. Fonlupt and *al.*[5] and M.Clergue and P.Collard[2] confirms previous results about the importance of a good FDC, and exemplify the existence of non artificial problem, as TSP, exhibiting this property.
Altenberg suggests the following interpretations to explain these surprising results [1]. He says that "either Hamming

[1] I3S Laboratory, University of Nice Sophia-Antipolis, France. e-mail:{pc,gaspar,clerguem,escazut}@alto.unice.fr

Submitted to ECAI 98
January 23, 1998

distance is related to GA dynamics, or that this relationship fortuitously holds among the test case Jones examined". The benchmark used by Jones is sufficiently large and various so that we think that one can do away with the second explanation. However this does not exclude the possibility of finding a counterexample as proposed by Altenberg. So, we rather think that the Hamming distance is related to GA dynamics. Following this idea, Jones remarks that Hamming distance is a good measure of distance for a walk based on mutations. Unfortunately as noted by Altenberg "the role of mutation in GA is to prevent premature convergence, and the recombination is the operator most important for GA performance". So we think that the only way to explain the pertinence of the Hamming-distance based FDC as statistical measure of GA difficulty is to find a indirect relationship between Hamming distance and crossover. This interpretation is referred by Altenberg but Jones does not discuss it.

2.2 Vose's predicates

Vose extends the concept of schemata to any subset of the search space, he calls predicates [7]. This is the most general definition of schema. More, he generalizes the schema theorem to predicates. Then, he defines the properties of globality and stability for predicates. The first one concerns the selective pressure and the second one the recombination operator. Globality means that "if a predicate is even rewarded by a population, then it will never be penalized by any population in which it is represented". Global predicates are the invariant predicates of GAs in the sense that they are independent of a population, and therefore thrive from generation to generation to the extend that they are not disrupted by recombination or mutation. Vose shows that the inverse image under the fitness function of interval type $[a, +\infty[$ are global predicates. The collection of increasing global predicates forms a nested sequence of subsets of the search space. Moreover, their intersection is exactly the set of elements which the GA is searching for: those elements which maximize the fitness function. Search is constrained in the appropriate direction by the selective pressure. Vose classes the predicates en three classes according to their sensitivity to recombination:

i. *Stable* predicates are not affected by recombination,
ii. *semi-stable* predicates gracefully degrade under recombination,
iii. *unstable* predicates may be devastated by recombination.

During the GA process, a stable predicate may only suffer from the selective pressure. Inversely, a global predicate may only undergo damage inflicted by recombination. Global and stable predicates confine exploration. Vose thinks that "the genetic paradigm will succeed when global increasing predicates are stable or semi-stable"

3 FDC and crossover

3.1 Crossover based distance

One way to set a link between FDC and crossover is to choose a distance which is in relation with crossover; that is a distance which measures the amount of crossover necessary to go from a bitstring to another one. This is not easy, on the one hand because crossover involves two strings and on the other hand because the variety of choice for mating is strongly governed by the diversity of the population. As a consequence, the effects of crossover are subject to the dynamics of GAs. At worst, after convergence towards an homogenous population, the crossover operator has no effect. In spite of this, Altenberg [1] has conduct a crossover based FDC analysis. This causes the problem to find a crossover based distance. He proposes the number of discontinuities between 0s and 1s in a bitstring x as the distance between x and the string of all 0s. This is the minimum number of 1-point crossovers required to transform x into 0. Another solution is derived by applying crossover in reverse. We wish to give a distance value of 1 to all the strings that can produce 0..0 through a single crossover, a distance of 2 is given to all the strings that can produce distance 1 strings etc. In themselves these results are interesting; but the fact that Jones uses Hamming distance to obtain his results, forces us to consider it to explain the FDC abilities.

3.2 Hamming distance and crossover

Contrarily to previous works, we think that the right track to capture what makes FDC a good forecaster, consists in altering the natural interpretation of the Hamming distance. The key idea is that this distance is in relation with crossover through the notion of stability. In order to give an intuitive idea of this link let us consider a set of two bitstrings x and x'. The more the Hamming distance between x and x' is, the more the offspring will be different from their parents. The less the Hamming distance between x and x' is, the more the offspring will look like their parents. Furthermore, it is obvious that the stability of a sub-population under the effects of the recombination operator is correlated to Hamming distance from offspring to parents: the more the distance is, the less the stability will be. From these properties we can conclude that the stability of a sub-population under the effects of crossover depends on the Hamming distance between its components. More, Hamming distance is invariant under crossover in the sense that distance between the two offspring is the same that the one between parents. As a consequence if we replace parents by offspring from one generation to the next, the stability of the population is not altered.

3.3 Vose's predicates and FDC

In order to compare Vose's predicate with FDC, let us consider the two following functions. These ones are defined on strings of length 3 from the unitation, that is the number of 1s in the chromosome. In both cases, the global optimum is 000 and the deceptive attractor is the complementary string 111. The first function f1 is GA-hard while the second one f2 is GA-easy. These results are predicted by the FDC value: $+0.35$ for f1 and -0.46 for f2. From the point of view of Vose the two functions are equivalent: global predicats have the same disruption probabilities (see tables 1 and 2);

so, it does not allow to predict the GA performance on f2. However, if stabilities of the predicates {000,111,001,010,100}

Table 1. Function f1

unitation	fitness	global predicate	
0	10	{000}	stable
3	9	{000,111}	unstable
2	8	{000,111,011,101,110}	semi-stable
1	1	{...}	stable

Table 2. Function f2

unitation	fitness	global predicate	
0	10	{000}	stable
3	9	{000,111}	unstable
1	8	{000,111,001,010,100}	semi-stable
2	1	{...}	stable

and {000,111,011,101,110} are computed not as a whole but rather relatively to the goal 000, we obtain a different results which allow to predict in both cases the behavior of the GA. We are going to explicit this point in the following section.

3.4 Fitness unstability correlation

The FDC coefficient quantifies the relationship between fitness and distance to the goal. In the same way we define the FUC coefficient to measure the relationship between fitness and unstability. More precisely, let us have $X = \{x_1, x_2, ..., x_n\}$ a set of n points in the search space, the corresponding set $F = \{f_1, f_2, ..., f_n\}$ of n fitnesses, the global optimum x', and the set $U = \{u_1, u_2, ..., u_n\}$ where the element u_k represents the unstability of the couple $\{x', x_k\}$ relating to recombination. The unstability u_k is computed as the mean on all the offspring y (from x' and x_k) of the $min(d_h(x', y), d_h(x_k, y))$. Finally, the FUC value is computed as the correlation coefficient between the two sets F and U. If we are maximizing, we should hope that fitness increases as unstability decreases. With an "ideal" fitness function, FUC, will therefore be close to -1.0.

For example, we give the FUC for some functions of known difficulty, along with the FDC and the ratio between them (see table 3). These results show that there is a strong link between Hamming distance and crossover disruption. So, the FUC correlation can be used to predict the performance of a GA.

Table 3.

function	λ	FDC	FUC	FDC/FUC
unitation	16	-1.000	-0.877	0.877
Withley F2 [8]	16	0.249	0.293	0.877
Trap (z=12) [4]	16	0.963	0.844	0.877

4 FDC and GA dynamics

Vose's point of view allows GAs to be regarded as a constrained random walk. We are going to base the GA dynamics explanation on this suggestion.

4.1 Scatter plots of fitness-distance and stability

In light of the FDC paradigm, one can consider a single criterion optimization problem as a multi-criteria one. The first criterion is the distance to the optimal fitness value and the second one is the distance to the goal bitstring (which have an optimal fitness value). These two criteria have to be minimized. Jones' idea can be reformulated by asserting that convergence will be easier if these two criteria are compatible. During the GA process, successive populations are under the influence of two strengths: the first one is the selective pressure and the second one tends to homogenize the population in order to get a steady state. The selective pressure is an explicit force pushing towards the fitter points at the risk of distancing the goal bitstring. Analysis of the scatter plot according to stability, can suggest that there are strengths which drive the populations either to approach or to distance the optimal bitstring. The scatter plots can be useful to specify areas more or less stable under crossover operator. Intuitively, a subset which contains points close to the goal (in words of Hamming distance) is slightly unstable. For instance the set of points which differ from the goal by only one bit is a semi-stable subset because the Hamming distance between parent and offspring is always equal to one. Likewise, a subset which contains points distant from the goal is nearly stable. For example the set of points which differ from the goal by $\lambda - 1$ bits, where λ is the length of the bitstring, is a semi-stable subset, too. On the other hand, nothing can be deduced concerning the stability of a subset of points situated at a middle distance to the goal. For instance, let us assume that the goal is the string 0000000000, the subset {0000011111, 1111100000} is unstable while {0000011111, 0000111110} is nearly stable. Finally, a subset is greatly unstable if it contains points at once close to the goal and distant from the goal. Likewise, a subset is unstable if it contains points at once close or distant to the goal and at a middle distance from the goal.

In conclusion the only areas where stability is sure are the ones which are distant from the central zone. So the crossover operator induces two implicit strengths which can drive the successive populations either towards the right rime or the left rime of the scatter plot of fitness vs. distance to the optimum. If the scatter plots contains points in such areas and more, if some of these points have a high fitness, there are paths for a GA. Such a way can be misleading if it drives towards the right rime; in this case the population converges towards a deceptive attractor. On the other hand, if the area is localized on the left side a GA is able to converge to the optimal solution. This interpretation of the dynamics is consistent with the fact that a deceptive attractor is near the complementary string of the optimum bitstring [8]. This model of the dynamics highlights two paths for a GA: the ones which drive towards the solution or a deceptive attractor; but it does not exclude the possibility that other ways through the central zone can be followed.

4.2 Stability analysis as GA's difficulty predictor

In order to validate our hypothesis about the relation between Hamming distance and crossover, we are going to examine the scatter plots of fitness vs. distance for the following function, constructed to be difficult for a GA with normal crossover Assuming an even length λ for the chromosome, we define the so-called ZeroOne function by: for each bitstring x, if the number of 0s is even, ZeroOne(x) is the number of 0s, otherwise it is equal to the number of 1s. The optimum is the string of all 0's with a fitness set to λ and there are λ deceptive attractors at unitation $(\lambda - 1)$ whose fitnesses are set to $(\lambda - 1)$ (see figure 1).
More, we consider the following operator defined by Vose

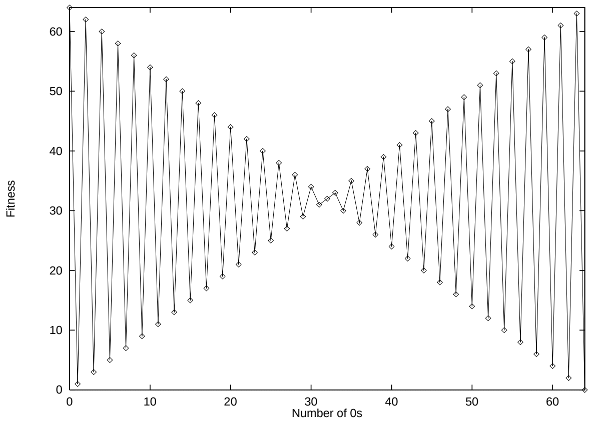

Figure 1. ZeroOne function with 64 bits.This plot may be viewed as the scatter plot fitness vs. distance to optimum

[7] for $\lambda = 3$, and called *mixing* operator. It can be extended to any length; by example, for λ=4, it is defined by $\mathcal{M} = M \circ \chi \circ M^{-1}$, where χ denotes the normal crossover, M is the matrix:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

and where bitstrings are identified with raw vectors over the field of integers modulo 2. One can note that this operator does not verify the properties of normal crossover relating to stability: for instance if the bitstrings x and y are close relatively to the Hamming distance, the subset {x,y} is unstable under $\mathcal{M}$ and inversely any subset $\{x, \overline{x}\}$, where $\overline{x}$ stands for the complementary string, is stable under $\mathcal{M}$. Experiments driven with a simple GA without mutation show that the ZeroOne function is GA-hard with the normal crossover while it becomes easy with the alternate mixing operator (see figure 2). In both case, the recombination rate is 1.0 and the chromosome length is set to 16 bits. Due to the *quasi* symmetry of the scatter plots the FDC value is close to zero and so, does not permit to predict these results. However examination of the scatter plots reveals the structure of the relation between fitness and distance. If the FDC ic calculated just for the points with an even (respectively odd), it is close to -1.0 (respectively $+1.0$). So, there are two paths, one leads towards the global optimum, the other leads towards the local optimum (see figure 1). With the normal crossover, the only areas where stability is sure are the ones which are close to the right or the left rime of the scatter plots. Unfortunately, for the ZeroOne function, more a string moves from the central zone more the difference between fitnesses grows. So there are a conflict between the selective pressure and the strength which drives the population towards the rimes of the scatter plots. This property allows to explain why the GA stays during a long epoch (from the first generations to generation 70) in the vicinity of the central zone (unitation close to 8) to finally converges to a deceptive attractor at unitation 15. This last phase of convergence is probably due to the genetic drift (population size is set to 200). The unstability near the rimes of the scatter plot, induced by the crossover, prevents the population from coming near the rimes. This explanation is confirmed by the fact that higher the crossover rate is, longer the stability period in the central area (fitnesses around 8) is (see figure 3). On the other hand, if we use

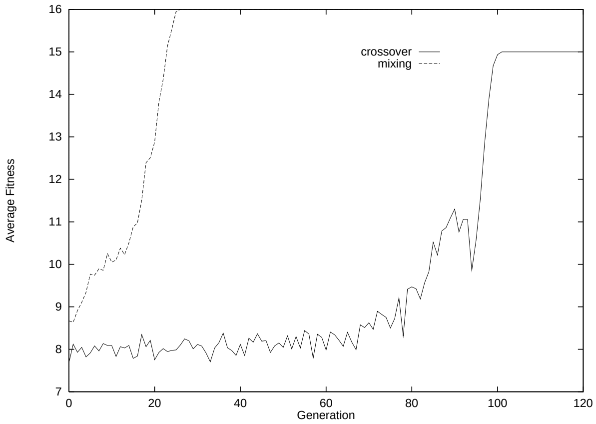

Figure 2. Average fitness vs. generation for the ZeroOne function on 16 bits

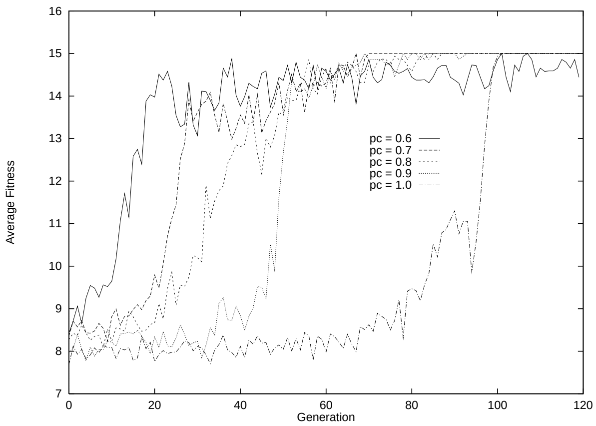

Figure 3. Average fitness vs. generation for the ZeroOne function on 16 bits with the normal crossover for 5 different crossover rate values

the alternate mixing operator, the areas closed to the rimes of the scatter plots are unstable while subpopulations containing complementary strings become stable. More, for the ZeroOne function, points closed to the rimes may have very different fitnesses while complementary strings have very close fitnesses. These properties suppress the conflict between the selective pressure and the strength which drives the population towards stable area of the scatter plots. However, in this case, a more precise tool should be introduced, in order to better analyse the GA dynamic. We do so in the next section. All this shows that the Hamming distance is firmly connected with the normal crossover operator. More generally, one can think that the Hamming-FDC is linked to any mixing operator which verify the properties worded in section 3.2. At this point, it could be interesting to investigate mixing operators other than crossover in terms of their properties in relation to Hamming-distance. Another point of view should be to consider one particular mixing operator and to look for the "right" distance such as Hamming distance stands for crossover.

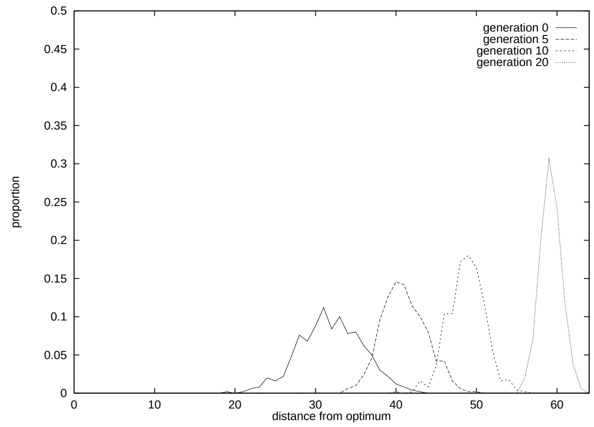

Figure 4. Population frequencies vs. distance for the Trap Function (64 bits, r=0.9, z=0.4) with crossover. The population size is 500

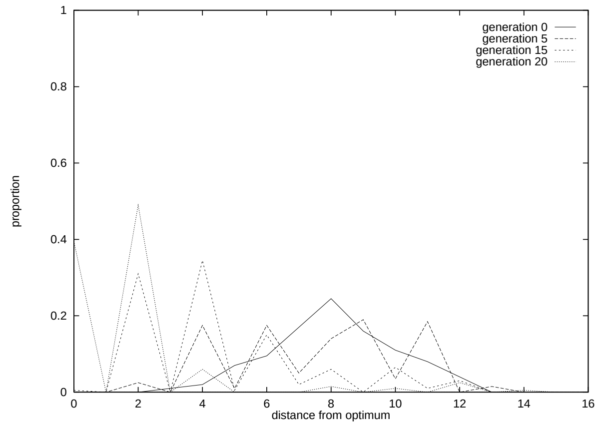

Figure 5. Population frequencies vs. distance for the ZeroOne Function (16 bits) with mixing. The population size is 200

4.3 Population frequencies vs. distance

The dynamics of a GA can be visualized by plotting the successive states of the population. However snapshots of scatter plots of fitness vs. distance at different stages during the test run of a GA provide defective information because they do not permit a visualization of the distribution of individuals (all the strings at the same distance are represented by only one point). Plotting the population frequencies vs. distance to the optimum is a more accurate way to visualize the dynamics. This kind of plotting provides an useful tool to better understand how a GA converge. First, let us consider a deceptive trap function defined on 64 bits with a population of 500 individuals (figure 4). The dynamics induced by the normal crossover drives the population towards the deceptive attractor. The distribution slides regularly towards the right rime while remaining homogenate. This confirms that the GA can evolve stable populations which mean fitness raises from generation to generation. Finally, let us consider the ZeroOne function. With the mixing operator, during the first populations, the distribution spread while remaining symmetrical (figure 5). This can be explain by the fact that complementary pairs are stables. More, populations contains especially points for which the number of 0s is even. The set of such points constitutes a way to the solution. This unexpected behavior can be induce by the unstability of strings close to each others. On the others hands, with the normal crossover, the examination of the evolution of distribution confirms that populations stay during a long time in the vicinity of the central zone; and so, points out the difficulty for the GA to approach the rimes of the scatter plots (figure 5).

5 Conclusion

We have presented an explanation of why the FDC is a good measure of problem difficulty for GAs despite of a lack of relationship between this measure and crossover or dynamics. The key idea is to reinterpreted the FDC in terms of fitness and stability. On the one hand, Vose's theory about predicates can be refine by results obtained on the FDC, and on the other hand this theory brings FDC to lights and permits to explain why it is a good predicator. The main conclusion is that FDC values are relevant because the Hamming distance is firmly connected with the normal crossover operator. So, it could be interesting to investigate mixing operators other than crossover in terms of their properties in relation to Hamming-distance. Another point of view should be to consider one particular mixing operator and to look for the "right" distance such as Hamming distance stands for crossover. Concerning the dynamics, significant insight can be obtain resulting from shifting emphasis away from the notion of distance to a goal towards the concept of stability under recombination; this allows GAs to be regarded as a constrained random walk rather than a walk in a topological space. Finally, we have proposed to analyze the dynamics of a GA by plotting the successive states of the distribution of individuals relatively to the distance to the goal. This kind of plotting provides an useful tool to study the relationship between the properties of the recombination operator regarding to stability and the dynamics.

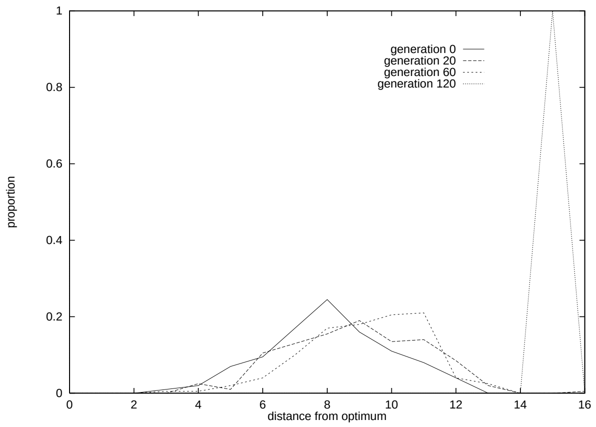

Figure 6. Population frequencies vs. distance for the ZeroOne Function (16 bits) with crossover. The population size is 200

REFERENCES

[1] L. Altenberg, 'Fitness Distance Correlation Analysis: An Instructive Counterexample', in *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, ed., Thomas Bäck. Morgan Kaufmann, (1997).

[2] M. Clergue and P. Collard, 'Genetic algorithm for neurogenesis', in *1998 International Conference on Evolutionary Computation (ICEC'98)*, (1998). to appear.

[3] P. Collard and C. Escazut, 'Fitness Distance Correlation in a Dual Genetic Algorithm', in *ECAI 96: 12th European Conference on Artificial Intelligence*, ed., W. Wahlster, pp. 218–222. Wiley & Son, (1996).

[4] K. Deb and D. E. Goldberg, 'Analysing deception in trap functions', in *Foundations of Genetic Algorithms 2*, ed., L. D. Whitley, pp. 93–108. Morgan Kaufmann, (1993).

[5] C. Fonlupt, D. Robillard, and P. Preux, 'Fitness landscape and the behavior of heuristics', in *Evolution Artificielle 97 (EA'97)*, (1997). To appear.

[6] T. Jones, *Evolutionary Algorithms, Fitness Landscape and Search*, Ph.D. dissertation, University of New Mexico, 1995.

[7] M. Vose, 'Generalizing the notion of schema in genetic algorithms', *Artificial Intelligence*, **50**, 385–396, (1991).

[8] L. D. Whitley, 'Fundamental principles of deception in genetic search', in *Foundations of Genetic Algorithms*, ed., G. Rawlins, 221–241, Morgan Kaufmann, San Mateo, CA, (1991).