Fitness Distance Correlation for Variable Length Representations

Leila Kallel Marc Schoenauer

Centre de Mathématiques Appliquées, CNRS URA 756,
École Polytechnique, 91128 Palaiseau Cedex, France.

Draft version - ()

Abstract

The choice of a representation for the problem at hand is the first cornerstone encountered when designing an Evolutionary Algorithm. The study of the correlation between fitness and distance to the global optimum (Fitness Distance Correlation), recently introduced as a measure of problem difficulty, has been demonstrated to be useful in the framework of bitstring representation (e.g. to compare between standard and Gray binary coding for real numbers). The first goal of this paper is to extend the FDC to the general framework of variable length representation involving continuous parameters. General guidelines are given to build a distance over such genotype spaces. But some other limitations of the original FDC concept, that are raised when applying FDC to variable length representations, are addressed here. The choice of the size of the random sample is adjusted to the problem at hand. The need to know the global optimum, unrealistic in real problems, is relaxed by the introduction of the Local FDC. Moreover, the stability of FDC results with respect to different samplings, or different choices of "local" optimum brings in some more information regarding not only the problem at hand, but also the tools used to derive these FDC results: It allows us in turn to choose among several available distances on the genotype space. Finally, all ideas presented in this paper are validated on some Structural Mechanics problems, where the target space is that of partitions of a two dimensional domain into two polygonal subsets, and for which different representations have been proposed, from (fixed-size) bitstrings to variable length Voronoï diagrams. FDC outputs turn out to fit well previous experimental results: FDC could hence be used in the future as a predictive tool, escaping endless trial-and-error processes usually performed to choose one representation for a given problem.

1 Introduction

Evolutionary Algorithms (EAs) involve two spaces during their search for optimality: the phenotype space, where the fitness values can be computed, and the genotype space, where the evolution operators are applied. The choice of the phenotype space is usually directed by the problem at hand. On the other hand, the choice of the *representation* of phenotypes, that is the choice of the genotype space and of the decoding function between a genotype and the corresponding phenotype, is left to the EA practitioner. Far up-stream from the usual choice of parameters (population size, operator probabilities, ...), the choice of representation is now unanimously recognized among the Evolutionary Computation community as being of vital importance for the success of the whole optimization process [31, 13, 3]. Furthermore, recent theoretical results [44, 35] suggest that black-box evolutionary optimization is bound to failure, hence emphasizing the need for a problem-dependent algorithm: the first possible problem-dependent choice is that of representation.

However, very few – if any – guidelines are available to help to choose the best representation for a given problem. The final choice mostly depends on the programmer's experience and/or faith, on a few theoretical or heuristical studies, and on the large body of experimental comparisons performed by researchers on other problems. Two aspects of such studies make it even more difficult when facing a new problem. First, their conclusions are either contradictory – as witnessed by the endless debate "binary vs non-binary coding" [16, 1] – or drawn in tow different frameworks, as for the coding of real parameters [42, 6, 33, 22, 11]. Second, it is very unlikely that any conclusion drawn from experimental results on a given problem applies to a brand new problem.

Among the few theoretical studies devoted to the study of representations, Radcliffe's theory of *formae* (which are generalized schemata) [33] offers a theoretical framework that gives design principles for operators on a given representation. One interesting consequence is the definition of representation-independent operators and evolutionary algorithms [43]. But it also allowed *a priori* comparisons of different representations for the same problem that were confirmed by numerical experiments on representations for the TSP problem [34]. The basic notion for these comparisons is that of "related fitness" of formae, and the tool was statistical evaluations of the variance of the fitness among formae of prescribed order. However, the definition of formae (or, equivalently, of equivalence relations on the genotype space) for variable length representations is not straightforward. For instance, the notion of "sharing an allele" should be relaxed to "having some similar allele" – which is exactly what a distance tries to describe. Further, the concept of related fitness can then be partially captured by studying the correlation between distance and fitness.

Some other approaches have been proposed, using statistical estimations over the search space to characterize problem difficulty [18, 27, 19, 36]. Their pros and cons will be discussed in Section 7, in the light of the results based presented in this paper.

The central argument of this paper relies on yet another tentative statistical estimation of problem difficulty, the notion of Fitness Distance Correlation, recently introduced by Jones [24]. The basic idea of the Fitness Distance Correlation (FDC) is that the higher the correlation between the fitness and the distance to optimum, the easier the optimization problem. In Jones'work, many examples on well-studied problems in the bitstring framework witness for that statement. Moreover, this approach brings a new understanding to some surprising previous results, e.g. on respective performances of standard versus Gray binary codings for real parameters. However, only bitstring representations using the classical Hamming distance are studied in [24].

The aims of this paper are to address further issues regarding the FDC concept, in particular in the case of variable length representations involving continuous parameters. General procedures will be proposed, and tested on variable length representations of subsets of a given domain of $\mathbb{R}^2$. The target problems are identification and optimization problems in the domain of Structural Mechanics (e.g. Topological Shape Optimization [26], Mechanical Inclusion Identification [40]) whose common feature is a very costly fitness evaluation procedure. Variable length representations for such problems have been proposed in [38, 39, 40]. But the only comparisons between these representations are based on experimental runs of Evolutionary Algorithms on chosen instances of the problems. The idea here is to try to *a priori* predict relative behaviors of the different representations by Fitness Distance Correlation analysis.

The paper is organized the following way:

Section 2 is devoted to the discussion of the original work on FDC. First, the basic definitions are recalled, illustrated by some simple results in the bitstring case, taken from [24]. General issues to extend FDC to other representations are addressed in section 3. The first major difficulty one has to face is to define a distance in the genotype space. Some guidelines are given to design useful distances on spaces described by variable length representations with continuous parameters at a reasonable computational cost. Another limitation of the original approach of Jones is the requirement that the global optimum should be known. A "Local Fitness Distance Correlation" is introduced and advocated in section 2.4, relaxing this unrealistic prerequisite. But uniform random sampling of the search space is critical to derive useful results. In large continuous spaces, an *a priori* fixed number of samples (e.g. 4000 in [24] for genotype spaces larger than 2^{12}) is not likely to give enough information on all problems. Procedures to derive problem dependent sampling are proposed in section 2.3.

Section 4 introduces the variable length representations that will be used in the experimental parts of the paper. The target problems, presented in section 4.1, are identification and optimization problems in Structural Mechanics, where the phenotype space is that of two-dimensional shapes [38, 40]. Based on the general guidelines, different distances are developed for these representations.

Section 5 is devoted to the experimental tuning of the different components of FDC analyses developed in section 2: the random sampling procedures are validated, the utility of the Local FDC is demonstrated, and finally FDC stability with respect to the choice of the local optimum provides some hints on the relative reliability of different distances for given genotype space and problem.

Last, but not least, the results of FDC studies should help in the *a priori* choice of a representation. Section 6 demonstrates that they can partially replace the "run-and-see" experimental procedure, resulting in actual savings of CPU resources. Previous experimental results of different representations are so "*a posteriori* confirmed".

Finally, the FDC concept, the proposed methodology and its results are discussed in Section 7, in the light of related works of the literature. Moreover, further directions are proposed to correct some weaknesses of the approach. A summary of results build up the conclusion section.

2 Fitness Distance Correlation

This section recalls the definition of the Fitness Distance Correlations (FDC) to measure problem difficulty, as stated by Jones [24], and discusses some important issues not mentioned by Jones: the extension of FDC to variable length representations with the consequences on computational cost, the random sampling of the search space and the impossibility in real problems to suppose that the global optimum is known.

2.1 Definition

Consider the problem of finding the minimum of a real fitness function $\mathcal{F}$ defined on a metric search space $\mathcal{E}$ with distance Δ and suppose that there exists one global minimum x_m of $\mathcal{F}$. A high correlation between the fitness of an individual and its distance to that global optimum characterizes optimization problems that should be easy to solve. Hence, the study of the correlation between the fitness of individuals and their distance to the global optimum can be a useful tool to improve our understanding of optimization problems.

Given a sample of n individuals x_i, $i \in [1, n]$, with associated fitnesses (f_i) and distances (d_i) to global optimum x_m, the *FDC coefficient* τ is defined by Jones [24] as:

$$\tau = \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})(d_i - \bar{d})}{\sigma_F \sigma_d}, \tag{1}$$

where $\bar{f}, \bar{d}, \sigma_F$ and σ_d are respectively the means and standard deviations of fitnesses and distances over the sample.

The expected link between FDC and problem difficulty holds for any optimization algorithm, but should be even more pertinent for Evolutionary Algorithms: the selective pressure globally drives the individuals toward regions of high fitness, the ergodicity prevents the traps led by local optima, and the population effect should allow to find the paths of increasing fitness and decreasing distance to optimum that are witnessed by the high FDC rate, even though FDC only gives a global view of the fitness landscape, regardless of local accidents.

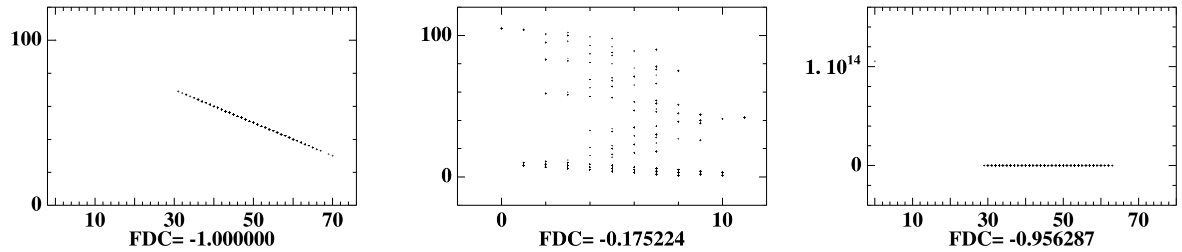
Furthermore, problems have been grouped by Jones in three classes (Jones actually considered maximization problems, while this paper is concerned with minimization problems, see Section 4.1; Hence, the sign of all FDC coefficients is reversed here compared to Jones').

1. Misleading if $\tau \leq -\tau_0 \leq 0$, as the fitness globally increases when individuals get closer to the global optimum; A typical value for τ_0 is 0.7;

2. Difficult if $-\tau_0 \leq \tau \leq \tau_0$ as no preferred path toward the global optimum exists to drive the population;

3. Straightforward if $\tau \geq \tau_0 \geq 0$, as the global tendency if the fitness increases is to get nearer the global optimum.

However, a single correlation coefficient cannot catch all the flavor of a problem difficulty. Further information is given by the repartition of the sample in the fitness $\times$ distance-to-optimum space, as shown in Figure 1 below in the bitstring framework. Such clouds of (fitness,distance) points will be termed *FDC plots* in the following.



(a) 100-bits onemax (b) 11-bits LongPath (c) 91-bits LongPath

Figure 1: *FDC plots for three bitstring problems. Each dot is a point from the random sample in the fitness $\times$ distance space.*

2.2 The Bitstring Case

All results of the seminal paper [24] deal with bitstring representations ($\mathcal{E} = \{0,1\}^N$ for some integer N), using the standard Hamming distance, defined by:

$$\Delta_H(x, x') \quad = \textstyle\sum_1^N d_a(a_i, a'_i) \tag{2}$$

where a_i are the alleles of x, and d_a is the trivial allelic distance taking value 0 if both alleles are equal, 1 if they differ. Another point of view on the Hamming distance, that will be used later to define variable length distances, is to see it as

$$\Delta_H(x, x') \quad = N - \text{number of common alleles} \tag{3}$$

In this framework, the well-known "count the ones" problem (also called "onemax") has a fitness that is defined as the Hamming distance to the global optimum $(111\ldots111)$. Hence, the FDC for that problem is perfect: the FDC coefficient is -1 (for a maximization problem), and the FDC plot of any large enough sample is a straight line, as can be seen on Figure 1-a.

Figure 1-b and -c are the FDC plots of the LongPath problem [20] with respectively 11 bits (as in [24]) and 91 bits. The fitness landscape of the LongPath problem consists in two different parts. First, a low-fitness hill gently leads the population to the entrance of the path, using the "count-the-ones" fitness; Second, the path itself is spiraling in that hill, steeply climbing toward the unique global maximum. On the path, the fitness of two adjacent chromosomes differs by 1: this makes the path clearly visible on the 11-bits FDC plots of Figure 1-b, and modifies the FDC coefficient from that of the onemax problem ($\tau = -0.17$).

However, the length of the path increases exponentially with the size of the chromosomes, but slower than the size of the whole space: the path is less and less "visible" in the fitness landscape. Hence, all the random sampling used to draw Figure 1-c happened to be on the gentle hill outside the path, giving both a false impression on the plot itself and a misleading FDC coefficient (the almost perfect $\tau = -0.96$). And as the length of the bitstring increases, the FDC coefficient tends to 1. Note that if the global maximum is included in the sample used in Equation 1, due to its very high fitness, the FDC coefficient becomes almost 0: and it's true that there is no correlation at all between the fitness and the distance to optimum. One possible remedy for that is to increase the size of the random sample: this issue will be addressed in next section.

2.3 Size of the random sampling

The importance of the random sampling procedure in any statistical study, as well as for the initialization of any EA, needs not to be emphasized. Most of the problems presented in [24] involve small bitstrings, and the FDC can be computed enumerating the whole search space. For all problems involving more than 12 bits, though, a sample of 4000 individuals is randomly chosen, uniformly in the search space. Unfortunately, this arbitrary choice of the size of the sample might not be suited to very large spaces.

Moreover, even for the same search space, different problems might require very different sample sizes. Caricature examples are given by the onemax and the LongPath problems in $\{0,1\}^N$: For the onemax problem, as the fitness of all individuals is the Hamming distance to the optimum, a sample containing exactly one point with k ones, for each value of $k \in [1, N]$ is enough to get the exact value of the FDC coefficient, as well as the exact FDC plot. On the opposite side, for the LongPath problem with large bitstring length, the path itself will hardly appear in small samples, and the FDC will only "see" the low-fitness hill leading to the entrance of the path: FDC results are not representative of the actual fitness landscape. Obviously, the size of the sample should depend on the ruggedness of the fitness landscape to get an approximation of the FDC coefficient with good confidence.

Yet another important factor influencing on the "quality" of the sample is the random initialization procedure. It is usually acknowledged to be uniform over the search space – as the choice of bits with equal probability for 0s and 1s over $\{0,1\}^N$. However, it might be difficult to design an unbiased initialization procedure on search spaces such as those defined in Section 4, as witnessed by results of section 5.2. Nevertheless, assuming a uniform initialization procedure, the rest of this section proposes procedures designed to automatically adapt the size of the random sample to both the size of the search space and the ruggedness of the fitness landscape.

2.3.1 Search for Stability

With little hope for any *a priori* estimation of a sufficient size for the random sample, *a posteriori* tests for stability were the only choice: If the addition of a significant number of example to an existing sample does not significantly modify some statistical measurement on that sample, it can be considered stable for this measure.

The most obvious measure that can be used to estimate that the stability is reached is the target value itself (here the FDC coefficients), though it is always dangerous to test the statistical stability of a sample using the very same measure to be estimated on that sample, due to possible over-fitting.

Other possible measurements rely on estimations of the diversity in the sample. Diversity is difficult to assess, but can be measured considering the repartition over the sample of known quantities such as: the fitness values, some phenotypic or genotypic distance to some known optimum.

These descriptors are nevertheless related: One fitness value is generally associated to different phenotypes; One phenotype can be associated to different genotypes if the representation exhibits some degeneracy. So it first seems easier to look for stability in the phenotype space than in the genotype one, the latter being more diversified: Experimental results of section 2.3 will nevertheless contradict that.

2.3.2 The Sizing Algorithm

After choosing a measure to test the stability of the sample, the algorithm proceeds as follows.

1. Generate N_{epoch} (typically 2000) random individuals.
2. Sort the sample according to the chosen measure.
3. Define fitness intervals containing the same proportion ρ (typically 5%) of individuals of the sample.
4. Add N_{epoch} random individuals to the sample.
5. Compute the new fitness intervals as in step 3.
6. Compare to the previous interval boundaries.
7. Repeat steps 4-5 until the average difference of the interval boundaries becomes smaller than some proportion ε (typically 0.1) of the average interval length.

New individuals are added to the sample until the repartition of the measure values becomes stable. Note that many experiments were conducted after the stability test succeeded, without ever experiencing any loss of stability. Of course, the three parameters N_{epoch}, ρ and ε have to be adjusted. But – and this is confirmed by experiments – the resulting algorithm is hopefully robust with respect to these parameters, and the values listed above were used for all following experiments.

Practically, two measures were used concurrently to test the stability: the fitness and the distance to global optimum – while only the fitness could be used on new problems with unknown solution. The experimental results of Section 5.1 will confirm that impression.

2.4 Local FDC

One of the major limitations of FDC for real problems is that the computations require the knowledge of the global optimum. i.e. the solution of the problem at hand. Though FDC based on global optimum allowed Jones to explain and understand previous results on well-known problems of GA literature, it will be of poor use when facing a new problem whose solution is actually unknown.

However, it is possible to compute the correlation coefficient between a sample and the best individual *in that sample*. Such a coefficient, termed τ_{loc}, is computed as in Equation 1, with the difference that all distances are now distances to the "local" optimum (the optimum in the sample). Of course, this coefficient will probably depend on the choice of that local optimum, and its stability with respect to different samplings should be checked carefully (see section 5.4). Nevertheless, both this stability and, when possible, the comparison of the values (or the average value) of τ_{loc} with the global FDC coefficient τ computed using Equation 1 can give some hints about the fitness landscape ... and the distance itself, as will be demonstrated on instances of variable length distances in section 5.5.

A related issue addresses the case where the representation exhibits a fairly high degree of degeneracy (different genotypes encode the same phenotype). In such situations, even if the global optimum is known, it is generally known in the phenotype space. Hence, different distinct genotypes representing the optimum might exist. The choice of a genotype for the optimum phenotype certainly can influence the FDC results. Here again, the study of the stability with respect to the optimum genotype is mandatory, and new insights on the fitness landscape and the distance at hand can be obtained (see section 5.3).

3 FDC for Arbitrary Representations

In the preceding section, the FDC concept, as introduced by Jones [24] has been recalled in the original bitstring framework, and new approaches have been proposed to overcome some of its limitations. This section will now address the case of more general representations.

The basic problem when dealing with more general representations is the definition of the (genotypic) distance. Three issues can raise some difficulty: The number of alleles can be variable; No canonical ordering of alleles might exist; When alleles are continuous parameters, there is no point in looking for common alleles between two individuals (as for instance in the corner-representation for the TSP [34]). Some guidelines will be proposed to account for any combination of these three difficulties.

3.1 Prerequisites for a distance

Suppose the genotype space $\mathcal{G}$ is that of unordered lists of *alleles*, all alleles taken from a space $\mathcal{A}$. Moreover, suppose there exist a *similarity measure* Sim from $\mathcal{A}$ into $\mathbb{R}^+$ and define for all alleles a, b in $\mathcal{A} \times \mathcal{A}$ the quantities $|a - b|$ and δ by

$$\begin{aligned} |a-b| &= Sim(a,a) - Sim(a,b), \\ \delta(a,b) &= |a-b| + |b-a| \end{aligned} \tag{4}$$

Suppose finally that δ is a distance over the space $\mathcal{A}$ (see Appendix 1, Lemma 1 for some sufficient conditions on Sim to ensure that δ actually is a measure).

Equation 2 (or 3) must be modified to a handle variable length number of continuous alleles. The concept of common alleles must be replaced by some "minimal distance between alleles", if continuous parameters are involved in the representation. Moreover, the lack of canonical ordering makes it necessary to perform some "allele-matching" before deciding which alleles are to be paired in the distance evaluation. Finally, unmatched alleles should be taken into account and increase the distance. The end of this subsection formalizes there considerations.

Consider two individuals $x = (a_i)_{i\in[1,n]}$ and $x' = (a'_i)_{i\in[1,n']}$ of $\mathcal{G}$. If for instance $n \leq n'$ a *matching* between x and x' is defined as an injective application from $[1, n]$ in $[1, n']$. The set of all matchings between x and x' will be denoted $M(x, x')$. The *best matching* $\mu(x, x')$ between x and x' is the matching minimizing the sum of the distances between matched pairs of alleles $\sum_1^n \delta(a_i, a'_{\mu(i)})$.

The tentative distance between x and x' in $\mathcal{G}$ is defined as

$$\Delta(x,x') = \sum_1^n \delta(a_i, a'_{\mu(x,x')(i)}) + \sum_{\mu(x,x')^{-1}(i)=\emptyset} \min_{j\in[1,n]} |a'_i - a_j| \tag{5}$$

Function Δ obviously satisfies
$\Delta(x,x) = 0\, \forall x \in \mathcal{G}$, and
$\Delta(x,x') = \Delta(x',x)\, \forall (x,x') \in \mathcal{G}^2$.
Moreover, the triangular inequality can be proved (see Appendix 1) in the case where all individuals have the same number of alleles. However, it does not hold in the general case: this issue will be addressed in section 3.3.

Once the best matching is known, the evaluation of the distance Δ is straightforward. Unfortunately, finding the best matching is itself a NP problem, and cannot be solved exactly. To cope with this difficulty, some heuristic matching was designed.

3.2 Heuristic Matching

To approximate the best matching between two lists of alleles $(a_i)_{i\in[1,n]}$ and $(a'_i)_{i\in[1,n']}$, with $n \leq n'$, the following deterministic algorithm was used:

1. Set a high matching level $L \in [0,1]$ (typically 0.90).

2. For all alleles a_{i1} in x yet unmatched
   find the closest allele a'_{i2} in x' (for distance δ),
   match the pair if $Sim(a_{i1}, a'_{i2}) > LSim(a_{i1}, a_{i1})$;
   otherwise, match the pair if $\delta(a_{i1}, a'_{i2}) = \min_{a_k unmatched} \delta(a_k, a'_{i2})$;
   otherwise a_{i1} stays unmatched.

3. If all alleles a_i are paired, Stop
   otherwise, decrease L (typically by subtracting 10%) and goto 2.

The complexity of this algorithm is in $n^2 \times n'$, as it requires the computation of all similarities between pairs of alleles. Experimental results not presented here with some of the distances defined in section 4 demonstrate the efficiency of this algorithm: results are very similar to exact results for small values of n and n', and consistently outperform the choice of the best matching among $n^2 \times n'$ randomly chosen matchings. Moreover, experimental tests introduced in next subsection enforce the presumption that replacing the exact best matching by the heuristic matching does not significantly degrade the utility of the distance at hand. Hence, all experiments throughout this paper will use the heuristic matching to approximate the best matching, together with Equation 5 to compute the distance between individuals of variable length genotype spaces.

3.3 Violations of the triangular inequality

Because the "distance" Δ defined in Equation 5 is not actually a distance, because in any case the best matching is approximated in the actual implementations (section 3.2 above), and because of technical representation-dependent reasons that will be detailed in turn, the distances actually used for the representations defined in Section 4 do violate the triangular inequality. In order to nevertheless justify their use and as a first attempt to compare their respective quality, the amounts of violation were experimentally computed the following way:

1. Generate t random triples of individuals;

2. For all three possible orderings A_1, A_2, A_3, compute the "violation of metricity"
   $$\omega(A_1, A_2, A_3) = \delta(A_1, A_2) + \delta(A_2, A_3) - \delta(A_1, A_3) \quad ;$$

3. Report the percentage ρ of violating examples (exhibiting strictly positive violations), and the average $\bar{\omega}$ over those examples of the relative amounts of violation $\frac{\omega(A_1,A_2,A_3)}{\delta(A_1,A_3)}$.

For each instance of a distance presented in section 4, these quantities will be computed as one measure of *a priori* quality of the distance at hand.

4 Representations and distances for 2D partitions

This section introduces actual instances of variable length representations on which all ideas presented in the preceding section will be experimented. The common phenotype space is the space $\mathcal{P}$ of partitions into polygonal subsets of a given subset of $\mathbb{R}^2$, called the *design domain*. Among the actual problems fitting in this description, are the topological Optimum Design, where the goal is to find the shape of a mechanical structure with minimum weight and sufficient strength [23, 8, 25], the inverse problem of geological soil properties [30], and the identification of mechanical inclusions [40], which will be the test problem in the present paper, and will be detailed in next section.

However, different representations are available for the phenotype space of partitions of a design domain of $\mathbb{R}^2$, and the choice of the best representation for a given problem instance is still an open problem, whatever the target application. This paper intends to partially answer that issue, using the FDC-based tools.

4.1 The mechanical problem

In structural mechanics, the non-destructive identification of inclusions is a difficult problem, mathematically ill-posed, and resisting up-to-date numerical methods [28, 9] In steel manufacturing plants, for instance, it is of vital importance to check if coal scories are included in steel parts, and if their repartition does not dangerously weaken the whole part.

In the simplest instance of mechanical inclusion identification problem, a structure is made of two different known materials, but their respective repartition in the structure is unknown. The available data consist of reference records of the mechanical behavior of the structure under known loadings (e.g. displacements of some points on the structure when given point-wise forces are applied). The goal is to find the geometrical repartition of both materials from these experimental data.

The identification can then be viewed as an inverse problem: For a given repartition of both materials, the computation of the simulated mechanical behavior of the structure is straightforward using any Finite Element package, and allows comparision of the simulated behavior to the experimental reference data. The goal is then to find the optimal repartition i.e. whose mechanical behavior is the most similar to the reference one. For instance, when reference data are displacements, the least-square error on the displacements will be the fitness function to minimize. The first results presented in [40] on simple artificial instances of mechanical inclusion identification problems are outstanding with respect to the deterministic state-of-the-art numerical results [9].

As stated above, the fitness values are computed using a Finite Element Analysis of the phenotype: this step requires a mesh of the structure, i.e. a discretization of the domain. It is well known that the discretization step introduces numerical noise in the resulting simulated mechanical behavior of the structure. But the Evolutionary Algorithm relies heavily on comparisons of such fitness values. Hence, a crucial point it that all Finite Element Analyses of structures in the same population should be performed using the same mesh.

For the purpose of this paper, some comparisons will be made between FDC predictions and sample runs of Evolutionary Algorithms on simple problems of mechanical inclusion identification: the reference data will actually be the simulated behavior of a known structure. One global solution is thus known, allowing more precise insight in the results.

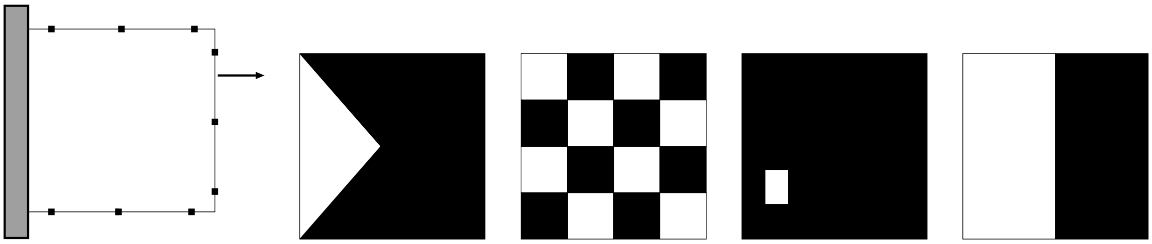


Figure 2: *The experimental structure, and the four test cases, respectively termed corner, checkerboard, asymmetric and symmetric.*

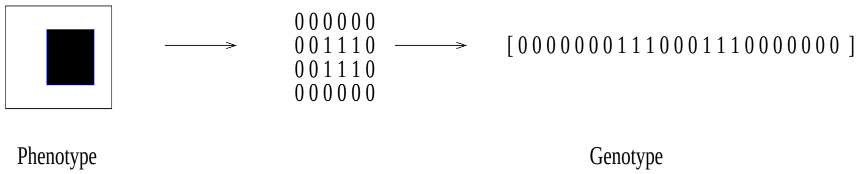


Figure 3: *A sample bitarray representation for a* 6×4 *mesh.*

Figure 2 shows the four reference structures that will be used throughout this paper: the fairly easy *corner* problem and the *checkerboard* problem, difficult in the eyes of mechanical engineers, are taken from [9], while the *asymmetric* and *symmetric* problems are purposely designed to test different behaviors of the representations introduced further in this section.

4.2 Bit Array Representation

Due to the mandatory Finite Element Analyses used to compute the fitness (as discussed in detail in [25, 39]) or perhaps simply because it is a straightforward representation from the GA point of view, the most commonly used representation for the phenotype space $\mathcal{P}$ is an array of bits (termed *bitarray*), based on a regular discretization (*mesh*) of the design domain into $l \times m$ *elements* (see also [23, 8]). Each bit corresponds to one element, and indicates to which of the two subsets this element belongs. Figure 4.2 shows an example of a bitarray representation for a simple 6×4 mesh.

The genotype space can thus be viewed as the regular space of bitstrings $\{0,1\}^{l \times m}$. Note however that, such a point of view can be misleading when designing the crossover operator: the standard n-point crossover operators are geometrically biased, and specific two-dimension operators must be designed, as demonstrated in [26].

4.3 Distances on bitarray space

The standard Hamming distance can be used on the bitarray representation, though suffering from the same one-dimensional bias, as bitarrays are not bitstrings. Hence, another distance was tried, based on the distance transformation.

$$
\begin{array}{cccc}
\begin{array}{c}
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\\
0\,1\,1\,1\,0\,0\\
0\,1\,1\,1\,0\,0\\
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\\
I_1
\end{array}
&
\begin{array}{c}
2\,2\,2\,2\,2\,2\\
1\,1\,1\,1\,1\,2\\
1\,0\,0\,0\,1\,2\\
1\,0\,0\,0\,1\,2\\
1\,1\,1\,1\,1\,2\\
2\,2\,2\,2\,2\,2\\
3\,3\,3\,3\,3\,3\\
DI_1
\end{array}
&
\begin{array}{c}
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\\
0\,0\,1\,1\,1\,1\\
0\,0\,0\,1\,1\,1\\
0\,0\,0\,0\,1\,1\\
0\,0\,0\,0\,0\,1\\
0\,0\,0\,0\,0\,0\\
I_2
\end{array}
&
\begin{array}{c}
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,0\,0\\
0\,0\,0\,0\,1\,2\\
0\,0\,0\,0\,1\,2\\
0\,0\,0\,0\,1\,2\\
0\,0\,0\,0\,0\,2\\
0\,0\,0\,0\,0\,0\\
|I_2 - I_1| = 11/10
\end{array}
\end{array}
$$

Figure 4: *The $DT1$-distance, based on the 1/1 Distance Transformation*

4.3.1 Distance Transformation

The distance transformation (DT) has been used in several applications of digital picture processing to measure the distance between features. The DT converts a binary picture consisting of feature and non feature elements (e.g. black and white parts), into an image where each element has a value that approximates the distance to the nearest feature element.
Good integer approximations of Euclidean distance can be computed by the process known as Shamfer distance. All details about distance transformation are given in [5].

DT-Distance:
Given two binary images I_1 and I_2, their feature dissimilarity is defined the following way: first, I_1 (Figure 4-a) is converted by the Shamfer 1/1 DT described above into distance image DI_1 (Figure 4-b). The dissimilarity $|I_1 - I_2|$ is then the average of the distances to features of I_1 of all feature-points of I_2, given by the corresponding value of the distance image DI_1 (Figure 4-d). The dissimilarity $|I_1 - I_2|$ is zero as soon as the features of I_2 are contained in the features of I_1.

Finally, the $DT1$-distance between images is defined by:
$d_{DT1}(I_1, I_2) = 0.5(|I_1 - I_2| + |I_2 - I_1|)$.
This distance will be used directly in Section 4.7.

4.3.2 Symmetric DT-distance

As for bitarray representation, the distance between bitarrays can be computed using the DT concept if both the ones and the zeros are considered in turn as "features" for the DT. The sum of dissimilarities between the zeros and the ones of both bitarrays is a distance over the genotype space of bitarrays:

$$\Delta_{DT}(B_1, B_2) = d_{DT1}(B_1, B_2) + d_{DT0}(B_1, B_2). \tag{6}$$

4.4 Limits of the bitarray representation

Anyway, the bitarray representation encounters very severe restrictions from the mechanical and numerical point of view. The accuracy of the Finite Element Analyses is dictated by the size of the mesh. Increasing the size of the mesh would not only increase the cost of the fitness computation (whose complexity is roughly quadratic in the size of the mesh) — but also the size of the bitarrays. This in turn would require increasing not only the size of the population, as heuristically derived in [17] and theoretically proved in [7, **?**], but also probably the number of generations needed to reach the same level of convergence. Hence, the bit-array representation scales up poorly when refining the mesh — not to speak of handling 3-D problems.

Therefore other representations have been designed in order to overcome this limitation, and disassociate the complexity of the representation and the accuracy of the fitness evaluation [38, 39].

4.5 Holes representation

A first variable length representation for partitions is based on an old-time heuristic method in Topological Optimum Design: from the initial design domain, remove material where the mechanical stress is minimal, until the constraints are violated. However, the lack of backtracking makes this method useless in practical applications. Nevertheless, this idea gave birth to the "holes" representation[1].

The design domain receives a default value (e.g. subset 0), and a list of "holes" describes the topology of the other subset (e.g. subset 1). These holes have an elementary shape taken from a library of possible simple shapes, and variable dimensions. Note however that the choice of the default value is arbitrary, and that another representation is defined by assigning value 1 as the default value. Both representations will be distinguished when necessary as the 1-Holes (respectively the 0-Holes) representations. Another important point is that the holes must be allowed to extend out of the initial domain Ω. Otherwise, the boundary regions of Ω would have a very high probability to stay at the default value, imposing a strong bias on the resulting phenotypes.

4.5.1 Rectangles Representation

The first trials used only rectangles as elementary shapes [38, 39]. Figure 5 gives an example of a genotype for the *Rectangle* representation, and the associated phenotype (partition of the initial domain). A genotype can be written as $(R_1, R_2, \ldots, R_l)$, where each rectangle R is defined by the coordinates of its center, and both dimensions, four real numbers with prescribed bounds.

All details of evolution operators for the rectangle representation can be found in [38, 39], but will not be discussed here, being out of the scope of this paper. Briefly, the crossover proceeds by exchange of some rectangles between the parents on the basis of two-dimensional grouping. The mutation operators involve Gaussian mutations inspired from Evolutions Strategies [42] with standard deviation inversely proportional to fitness (for a minimization problem) as in [12].

4.6 Triangles Representation

The triangle representation is another instance of a "hole" representation in which the only allowed elementary shape for the holes is the triangle. An individual is thus a list of triangles $(T_1, T_2, \ldots, T_l)$, each triangle being defined by the coordinates of its three summits.

While the basic ideas to design a distance for the triangles were the same than for the rectangles, it proved however to be slightly more difficult to actually implement, merely for technical reasons.

However, distances for the Holes representations can be designed regardless of the actual possible elementary shapes for the holes: this is the scope of next section.

[1] Another approach in the domain of dynamical image processing led to a similar representation in [29].

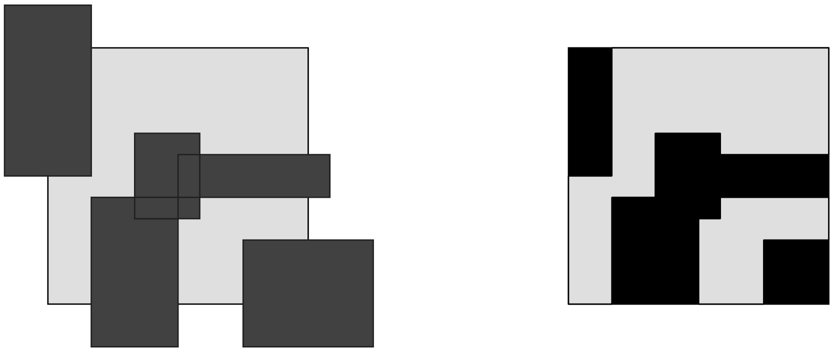


Figure 5: *A sample rectangle genotype and the associated phenotype.*

4.6.1 Area based distance

Following the general guidelines of Section 3, a distance of the genotype space can be defined from any similarity measure between holes. The general idea to design such similarity measure is to define $Sim(R,S)$ for two holes R and S as $A(R \cap S)$, $A(X)$ denoting the area of subset X of $\mathbb{R}^2$.

However, such a definition leads to undesirable results whenever the two holes have empty intersection: the similarity is the same, whatever the distance between the two disjoint holes. That's why the previous definition was modified into:

$$Sim_A(R,S) = \min(A(R), A(S))e^{\frac{-d(R,S)}{C(R,S)}}, \tag{7}$$

where $d(R,S)$ is the Euclidean distance between the centroids of the holes R and S, and $C(R,S)$ is a scaling parameter computed such that the quantity $e^{-d(R,S)/C(R,S)}$ for two identical holes that are exactly half way overlapping is equal to 0.5. This is achieved by the formula $C = (l_R + l_S)/(4 * ln(2))$, where l_R is a characteristic dimension of hole R, calculated according to its shapes and position relatively to S. Note that $Sim_S(R,R) = A(R)$, the area of hole R.

Following now the general ideas of Section 3, a distance Δ_A can be designed, using Equations 4 to define a distance δ_A over the space of holes from the similarity measure given by Equation 7, and Equation 5 to define the distance over the genotype space.

If parameters $C(R,S)$ are considered independent of holes R and S, it can be proved that δ_A is a distance over the space of holes (see Appendix 1, Lemma 1). However, when $C(R,S)_A$ does depend on the holes, as described above, δ is no longer a distance. But numerical results show that it differs from a distance by very small amounts: using the measures defined in section 3.3 over a sample of 1000 random rectangles, the percentage ρ of violating examples is 0.04% while the average of violations $\bar{\omega}$ is 4.5%.

To sum up the area-based distance between $x = (R_i)_{i \in [1,n]}$ and $x' = (R'_i)_{i \in [1,n']}$ in the genotypic space is the following:

$$
\begin{aligned}
\Delta_A(x,x') = \quad & \sum_1^n A(R_i) + \sum_1^{n'} A(R_i') - 2\left(\sum_{i=1}^n \min(A(R_i), A(R_{\mu(i)})')e^{\frac{-d(R_i,R'_{\mu(i)})}{C(R_i,R'_{\mu(i)})}}\right) \\
& - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(\min(A(R_j), A(R_i'))e^{\frac{-d(R_k,R_i')}{C(R_k,R_i')}}\right),
\end{aligned} \tag{8}
$$

where $\mu = \mu(x, x')$ is the *heuristical best matching* between x and x' computed following Section 3.2.

The numerical test regarding the triangular inequality for Δ_A gave the following results (see Section 3.3): over a sample of 1000 random holes, the percentage ρ of violating examples is 0.9% while the average of violations $\bar{\omega}$ is 5.8%.

4.6.2 Hole based distance

Another approach to the definition of a distance over the space of variable length list of holes comes from starting from Equation 3 instead of Equation 2. Back in the general formalism of above Section 4.6.1, this amounts to consider all holes having "size" 1, and leads to the following definition for the similarity measure between holes:

$$
Sim_H(R,S) = \frac{2\min(A(R), A(S))}{A(R)+A(S)} e^{\frac{-d(R,S)}{C(R,S)}}, \tag{9}
$$

Note that here, $Sim_H(R, R) = 1$ for all holes R. If parameters $C(R, S)$ are considered independent of holes R and S, it can be proved that δ_H is a distance over the space of holes (see Appendix 1, Lemma 2). However, when $C(R, S)_H$ does depend on the holes, numerical results show that it still remains a distance. There have been no violating examples over a sample of 1000 random rectangles. Following the same process than in the preceding section leads to define the distance Δ_H over the genotype space as:

$$
\begin{aligned}
\Delta_H(x,x') = \quad & (n+n') - 2\left(\sum_{i=1}^n \frac{2\min(A(R_i), A(R'_{\mu(i)}))}{A(R_i)+A(R'_{\mu(i)})} e^{\frac{-d(R_i,R'_{\mu(i)})}{C(R_i,R'_{\mu(i)})}}\right) \\
& - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(\frac{2\min(A(R_j), A(R_i'))}{A(R_j)+A(R_i')} e^{\frac{-d(R_k,R_i')}{C(R_k,R_i')}}\right),
\end{aligned} \tag{10}
$$

where $d(R_i, R_j)$ is the Euclidean distance of the centers of holes R_i and R_j.

This function Δ_H is a distance if C does not depend on the holes and $n = n'$ (see Appendix 1, Lemma 2). The numerical test regarding the triangular inequality for Δ_H gave the following results (see Section 3.3): over a sample of 1000 random holes, the percentage ρ of violating examples is 0.03% while the average of violations $\bar{\omega}$ is 1.6%.

4.7 Voronoï Representation

Another possible way of representing polygonal shapes comes from computational geometry, more precisely from the Voronoï diagram theory (see [32, 4] for a detailed introduction to Voronoï diagrams, and a general presentation of algorithmic geometry).

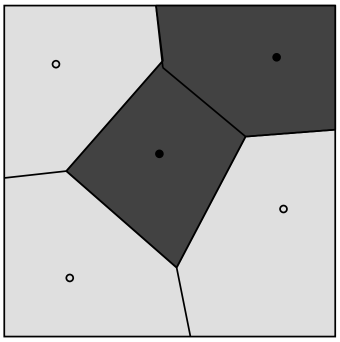


Figure 6: *A sample Voronoï representation for a partition of the design domain.*

4.7.1 Voronoï diagrams

Consider a finite number of points $V_0, \ldots, V_N$ (the *Voronoï sites*) of the design domain Ω. To each site V_i is associated the set of all points of the design domain for which the closest Voronoï site is V_i, termed *Voronoï cell.* The *Voronoï diagram* is the partition of the design domain defined by the Voronoï cells. Each cell is a polygonal subset of Ω, and any partition of Ω into polygonal subsets is the Voronoï diagram of at least one set of Voronoï sites [32, 4].

Consider now a list of Voronoï sites, each site being labeled 0 or 1. The corresponding Voronoï diagram represents a partition of the design domain into two subsets if each Voronoï cell is labeled as the associated site (here the Voronoï diagram is supposed regular, i.e. to each cell corresponds exactly one site). Example of Voronoï representations can be seen in Figure 6. The Voronoï sites are the dots in the center of the cells.

Here again, the definition of a distance on the Voronoï genotypes will follow the ideas of Section 3. But different solutions are available: From the purely genotypic point of view, an individual is completely defined by the list of Voronoï sites. However, and experimental results of section 5.5 will account for that, considering only the list of sites does not bring much information about the resulting partition of the domain, and it might be useful to consider also the Voronoï cells in the distance, in a similar way than the holes ino section 4.5. Along these lines, four different attempts to design a distance have been made.

4.7.2 Site based distance

As said above, the first approach is to consider a Voronoï representation as a list of sites $(V_i)_{i\in[1,n]}$. Each site is labeled 0 or 1, and the label of site V is denoted $l(V)$. The Euclidean distance of $\mathbb{R}^2$ is denoted d. A similarity measure between two sites is given by

$$\begin{aligned} Sim_S(V_1, V_2) &= e^{\frac{-d(V_1,V_2)}{C}} \text{ if } l(V_1) = l(V_2), \\ &= 0 \text{ otherwise,} \end{aligned} \tag{11}$$

where C is a constant parameter. Note that $Sim_S(V, V) = 1$ for all sites V.

Following the guidelines of Section 3, namely Equations 4 and 5, a distance Δ_S between $x = (V_i)_{i\in[1,n]}$ and $x' = (V_i')_{i\in[1,n']}$ is constructed as:

$$
\begin{aligned}
\Delta_S(x,x') = \quad & (n+n') - 2\sum_1^n e^{\frac{-d(V_i,V'_{\mu(i)})}{C}} \\
& - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(e^{\frac{-d(V_k,V'_i)}{C}} \right),
\end{aligned}
\tag{12}
$$

4.7.3 Site/Cell based distance

A second approach is to introduce the size (area) of the cells in the weight of the similarity measure. The "alleles" are now denoted $(V_i, C_i)_{i\in[1,n]}$, where C_i is the Voronoï cell associated to the site V_i, of area $A(C_i)$. In the same way as Section 4.7.2, the similarity measure is defined by:

$$
\begin{aligned}
Sim_{SC}((V_1,C_1),(V_2,C_2)) \quad & = \frac{2\min(A(C_1),A(C_2))}{A(C_1)+A(C_2)} e^{\frac{-d(V_1,V_2)}{C(C_1,C_2)}} \text{ if } l(V_1)=l(V_2), \\
& = 0 \text{ otherwise.}
\end{aligned}
\tag{13}
$$

where C scales the distance according to the size of cells:

$$
C(C_i,C_j) = \frac{\sqrt{2A(C_i)} + \sqrt{2A(C_j)}}{6ln(2)} \tag{14}
$$

The general idea when matching two cells is to approximate their average dimension with C, so that the amount of penalization is dynamically adapted to the dimensions of the couples involved. (The same issue is addressed in Section 4.7.5).

Note that the self-similarity $Sim_{SC}((C,V),(C,V))$ is still 1 for all sites (V,C).

Using again Equations 4 and 5, the distance Δ_C between $x=(V_i)_{i\in[1,n]}$ and $x'=(V'_i)_{i\in[1,n']}$ is:

$$
\begin{aligned}
\Delta_{SC}(x,x') = \quad & (n+n') - 2\sum_1^n \frac{2\min(A(C_i),A(C'_{\mu(i)}))}{A(C_i)+A(C'_{\mu(i)})} e^{\frac{-d(V_i,V'_{\mu(i)})}{C(C_i,C'_{\mu(i)})}} \\
& - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(\frac{2\min(A(C_k),A(C_i)')}{A(C_k)+A(C_i)'} e^{\frac{-d(V_k,V'_{\mu(i)})}{C(C_k,C'_i)}} \right)
\end{aligned}
\tag{15}
$$

4.7.4 Cell/Site based distance

In some sense, the distance Δ_{SC} above can be considered similar to the holes distance of section 4.6.2. Hence, going further toward phenotypic distances, a distance similar to the area-distance for holes (section 4.6.1) can be designed. First define the similarity measure as:

$$
\begin{aligned}
Sim_{CS}((V_1,C_1)-(V_2,C_2)) \quad & = \min(A(C_1),A(C_2)) e^{\frac{-d(V_1,V_2)}{C(C_1,C_2)}} \text{ if } l(V_1)=l(V_2), \\
& = 0 \text{ otherwise,}
\end{aligned}
\tag{16}
$$

where $C(C_i,C_j)$ is defined by Equation 14. The self-similarity $Sim_{CS}((V,C),(V,C))$ is now the area of cell C.

Then, using Equations 4 and 5, the distance Δ_{CS} is:

$$\begin{aligned}\Delta_{CS}(x,x') = \; & 2A(\Omega) - \sum_1^n 2\min(A(C_i), A(C'_{\mu(i)}))e^{\frac{-d(V_i,V'_{\mu(i)})}{C(C_i,C'_{\mu(i)})}} \\ & - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(\min(A(C_k), A(C'_i))e^{\frac{-d(V_k,V'_i)}{C(C_k,C'_i)}} \right),\end{aligned} \tag{17}$$

where $A(\Omega)$ is the area of the whole domain, sum of areas of all cells of an individual.

4.7.5 Cell based distance

The ultimate step in the same direction is to consider as alleles only Voronoï cells. The matching between "alleles" can also be made according to a distance between cells – namely the DT-distance defined in section 4.3.2 by Equation 6. The similarity measure is so defined as:

$$\begin{aligned} Sim_C(C_1,C_2) & = \min(A(C_1), A(C_2))e^{\frac{-d_{DT_l}(C_1,C_2)}{C(C_1,C_2)}} \text{ if } l(V_1) = l(V_2), \\ & = 0 \text{ otherwise,}\end{aligned} \tag{18}$$

where $d_{DT_l}(C_1,C_2)$, the *DTl-distance* between cells defined in Section 4.2. The self-similarity of all cells $Sim_C(C,C)$ is again the area of cell C.

The distance Δ_C is then defined by:

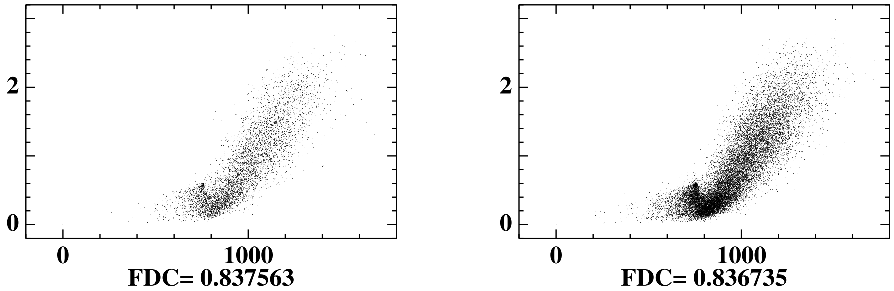
$$\begin{aligned}\Delta_C(x,x') = \; & 2A(\Omega) - \sum_1^n 2\min(A(C_i), A(C'_{\mu(i)}))e^{\frac{-\delta_{DT}(C_i,C_\mu(i))}{C(C_i,C_\mu(i))}} \\ & - \sum_{\mu^{-1}(i)=\emptyset} \max_{k\in[1,n]} \left(\min(A(C_k), A(C'_i))e^{\frac{-\delta_{DT}(C_k,C'_i)}{C(C_k,C'_i)}} \right),\end{aligned} \tag{19}$$

Of course, having defined these distances, the first step before actually using FDC should be to choose the best distance, if any. Some results of stability sketched in section 2 will be used in forthcoming section 5.5.

5 Tuning FDC components

Before using FDC as an estimator of problem difficulty, or representation-problem adequacy, it is necessary to validate the different components of FDC tool box introduced in preceding sections. The goal of this section is to experimentally study the new features introduced earlier: The problem-dependent sizing of the random sample is experimented in section 5.1; an unexpected by-product is the comparison of different random initialization procedure when no obvious uniform procedure exists, sketched in section 5.2; The behavior of FDC for different genotypes of the global optimum (section 5.3), as well as the Local FDC approach (section 5.4) are then tested. Finally, these different results build up a global picture of the distance at hand, and turns out to allow for a comparison between different distances on the same genotype space, as shown in section 5.5.

All results presented below deal with one of the test problems of Section 4.1 (corner, checkerboard, asymmetric, symmetric), using one of the representations of section 4 (bitarray, 0- or



(a) 6000 sample (genotypic stability) (b) 20000 sample (fitness stability)

Figure 7: *FDC plots using two different stability criterion (corner problem, 1-Rectangle representation). The FDC coefficients and the global outlook of the plots are almost identical.*

	Genotypic stability		Phenotype stability	
problem	sample size	τ	sample size	τ
corner (Rect. in)	6000	0.86	20000	0.86
corner (Rect.out)	6000	-0.66	14000	-0.63
asymmetric (Rect. in)	6000	0.93	18000	0.94
asymmetric (Rect. out)	6000	-0.62	6000	-0.7
corner (Voronoï)	6000	0.36	6000	0.36

Table 1: *The minimal stability sizes for both the distance-to-optimum and the fitness criteria.*

1-rectangles or triangles, Voronoï) together with one of the specific distances for that representation. Unless otherwise specified, the distances will be the area-based distance of Section 4.7.5 for the holes representations, the Cell/Site based distance of Section 4.7.4 for the Voronoï representation and the DT-distance of section 4.3.2 for the bitarray representation, following the – forthcoming – results of section 5.5.

5.1 Random sampling procedures

In Section 2.3, a general procedure has been introduced to adapt the size of the random sample to the problem at hand. This procedure is based on the stability of some statistical measure with respect to addition of epochs of examples to the actual sample. Two measures have been used: the fitness values (phenotype stability), and the distance-to-optimum values (genotype stability). The most significant results regarding the minimal sizes of the stable sample in both cases, with epoch size of 2000, are given in Table 1, together with the subsequent FDC coefficients. Slightly different behaviors can be observed, depending on the representation. But apart from the Voronoï representation, the samples generated using the genotype distance are generally smaller than those generated using the fitness, while the FDC coefficients (and the fitness-distance plots) are very similar.

Further, a more detailed comparison is presented for the corner problem with 1-Rectangle representation and area-based distance, for which a large difference in sample size was observed. On the FDC plots in Figure 7, the general outlook of the samples after stability was reached are identical.

	Genotypic stability		Phenotype stability	
Problem	Genotypic space	τ	Objective space	τ
corner (arbitrary size)	6000	0.83	20000	0.83
corner (small size)	8000	-0.752	8000	-0.754

Table 2: Stable sample sizes for different initialization procedures.

5.2 Optimizing the Initialization Procedure

One of the important questions which is generally omitted when running an Evolutionary Algorithm is that of the initialization procedure: most of the time, a uniform sampling procedure is assumed, and easily implemented, be it based on bitstrings or real-valued parameters with prescribed bounds. However, when dealing with complex representations (e.g. variable length), there might exist no obvious uniform initialization procedure.

5.2.1 Rectangle representation

Consider for instance the rectangle representation, with initial domain $\Omega = [0, L_x] \times [0, l_y]$. The initialization procedure first draws the number of rectangles, uniformly in $[1, Max_R]$, for some user-defined parameter MAX_R (50 in all experiments of this paper), then chooses the coordinates of the center of each rectangle, uniformly in Ω, then the width and length uniformly in $[0, x_{Max}]$ and $[0, y_{Max}]$, for some user-defined x_{Max} and y_{Max}. However, different choices of x_{Max} and y_{Max} do lead to different sampling of the genotype space, hence resulting in different performances of the evolutionary algorithm. This phenomenon can be captured by the FDC analysis, as witnessed by the stability results of Table 2 and the FDC coefficients and plots of Figure 8-a and -b. Moreover, actual runs of the Evolutionary Algorithm (described in more details in section 6) demonstrate similar influence of the initialization procedure in the early generations (see Figure 11).

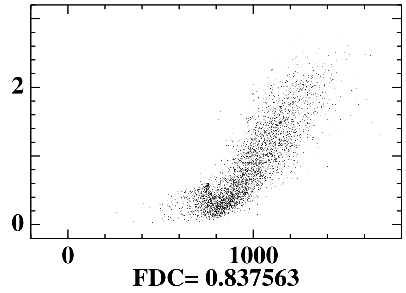
5.2.2 Triangle representation

Another spectacular example is given using the Triangle representation, for which three initialization procedures were tested:

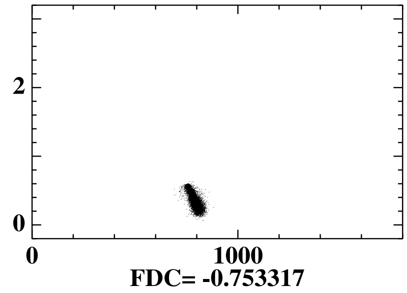
1. Generate three points uniformly in a domain larger than Ω.
2. Generate one point uniformly in Ω, then two other points by gaussian perturbations of mean deviation $x_{Max}/2$ of its coordinates.
3. Generate one point uniformly in Ω, then two other points by gaussian perturbations of mean deviation $x_{Max}/4$ of its coordinates.

The results of correlation are quite different as shown in Figure 9. Procedures 1 and 2 (Figure 9-a and -b) do not generate enough examples on the left lower side of the plot (genotype made out of small triangles and possibly high fitness). On the other hand, procedure 3 (Figure 9-c) doesn't generate enough chromosomes on the right side of the plot (genotypes with large triangles).

The question of the choice of the initialization procedure remains however open:
In both cases, the sample isn't representative of the search space. All these results put together will likely build up a representative sample, meaning that this problem is probably not well

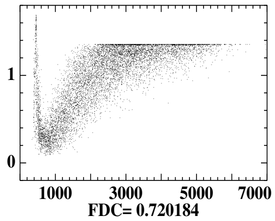


(a) Large size rectangles

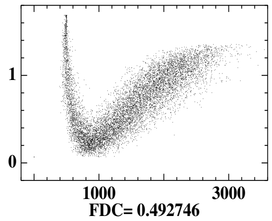


(b) Small size rectangles

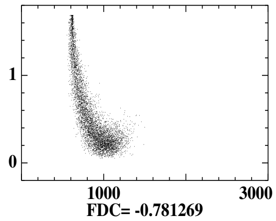
Figure 8: *FDC plots for different initialization procedures for the 1-rectangle representation on the corner problem: The FDC coefficients highly depend on the initialization.*



(a) Procedure 1

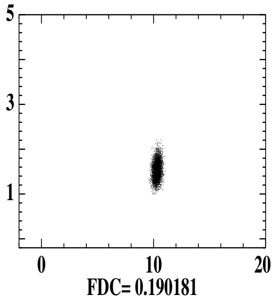


(b) Procedure 2

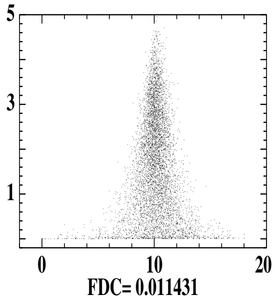


(c) Procedure 3

Figure 9: *FDC plots for three different initialization procedures for the Triangle representation on the checker-board problem. The upper bound in plot (a) corresponds to the genotypes made of a few degenerate triangles, having empty phenotypes. Note how the FDC coefficients degrade from (a) to (c).*

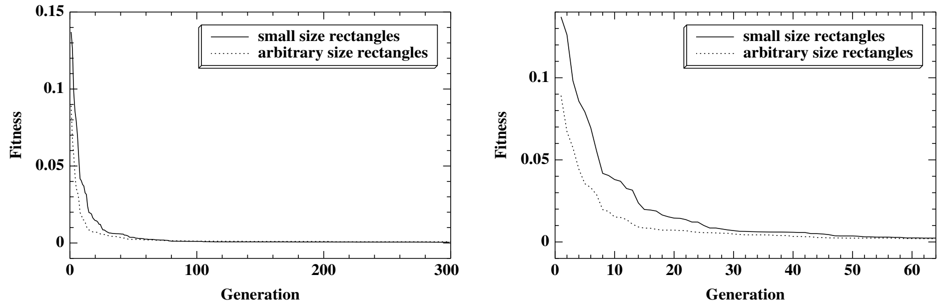


(a)



(b)

Figure 10: *FDC plots for two different initialization procedures for the bitarray representation for the asymmetric problem. Plot (a) represents a uniform random sample of the bitstring space, while plot (b) is a random sample generated by the Rectangle representation and translated in the bitstring space. Both samples are of size 6000.*



(a) Complete runs (b) Zoom on the first generations

Figure 11: *Median plots out of 21 runs of a GA on the corner problem and Rectangle representation. The EA runs reflect the difference in initialization procedure, at least in the early generations.*

correlated. In fact, when the individuals get genotypically close to the optimum, the fitness gets worse: The correlation result over the hole space is not always significant, as witnessed by the first and the second plots. The increasing difficulty when getting closer to the optimum hardly influence the global result.

5.2.3 Bitarray representation

A caricature instance of that phenomenon can be seen for the bitarray representation on Figure 9: The plot of Figure 9-a is a stable sample uniformly chosen in hte bitstring space (each bit is chosen randomly with probability 0.5 for each label 0 or 1); The sample plotted on Figure 9 was built using the Rectangle initialization procedure with large rectangles (see Section 5.2.1 above), then translated into the bitstring space. The distance is the DT-distance (see section 4.3.2), chosen after the failure of the Hamming distance reported later in Section 5.5. The peculiarity of the bitarray representation appears here clearly: a random sample in the bitstring space cannot bring any information on the fitness landscape for the asymmetric problem, as all samples, having about half bits to 1, have a very high error when compared to the almost-all-zeros solution. Obviously, the FDC coefficient alone is not enough here. However, it can be easily argued that the initialization procedure should take into account the specificity of the problem: mechanical inclusions are made of large subsets of either material.

5.3 Genotype of the Global Optimum

In non orthogonal representations, the same phenotype can be represented by different genotypes. Moreover, when the global optimum is known, it is generally known in the phenotype space. Hence, the results of the FDC analyses depend on the choice of a genotype for the known optimum phenotype.

Experimental comparisons have been performed on the problems and for the representations presented in Section 4. For each of the known phenotypes of the global optima (Figure 2), different genotypes were computed. In the simplest cases (e.g. one rectangle for the symmetric problem), the genotypes were designed by hand. In order to obtain more complex genotypes, a simple Evolutionary Algorithm was run, using the phenotypic distance to the optimum as fitness.

The best solutions of different runs of the algorithm were then eventually adjusted slightly to give the perfect phenotype.

Globally, and whatever the representation (Rectangles, Triangles or Voronoï), the distance and the difficulty of the problem for that representation (see Section 6.1), very small variations have been observed for the FDC analyses using different genotypes for the same phenotype. The most significant results are presented in Figure 12: for the Rectangles representation and the area-based distance, where no significant difference can be observed. Slightly greater differences appear for the Voronoï representation and the cell/site based distance, as can be seen in Figure 13. However, the general outlook of the FDC plots are the same, hence giving close values for the FDC coefficients. In particular, the estimation of problem difficulty is not modified by the choice of a genotype for the optimum phenotype.

5.4 Global vs Local FDC

As mentioned in section 2.4, when handling a new problem, the global optimum is generally unknown, and only Local FDC analyses can be performed.

For 0-Rectangles representation on the corner problem, the optimum in the sample (Figure 14-c) resembles the actual solution of the corner problem: the FDC plots and coefficients (Figures 14-a and -b) do not differ much. The same representation on the checker-board problem exhibits an optimum quite different from the actual solution (Figure 15-c), which might explain the difference between the corresponding FDC coefficients, though the FDC plots look much alike (Figures 15-a abd -b). An intermediate situation is achieved by the Voronoï representation using the cell/site based distance on the asymmetric problem: Though the optimum in the sample is not much different from the actual optimum (Figure 16-c), the global FDC plot (Figure 16-a) is more compact than the local one (Figure 16-b). Nevertheless, their general slope is the same, resulting in close values for the FDC coefficients.

However, the optimum of a given random sample is arbitrary, and the stability of the local FDC results should be checked carefully. Table 3 gives experimental results for almost all representations and distances for each one of the four test problems of Section 4.1: the averages and standard deviations of FDC coefficients with respect to the 20 best individuals of a random sample are given, together with the corresponding global FDC coefficients. Some obvious different situations can be observed among the distance-representations couples, regarding both the stability of the local FDC coefficients (as given by the standard deviations) and the comparison between the global and the local results.

The influence of the choice of the local optimum on the FDC coefficient thus seems to be another indicator of the reliability of FDC results. However, it can also be interpreted *a posteriori* as a measure of the adequacy of the distance to the representation. The Next subsection is devoted to the comparisons of different distances for the same representation.

5.5 Choice of a Distance

5.5.1 Holes Representations

For the Holes representations, two distances have been defined in section 4.5, one based on weighting each hole by its area (given by Equation 8), and the other one having all holes equivalently weighted by 1 (given by Equation 10). Both distances give roughly similar results for FDC coefficients and plots as shown in Figure 17.

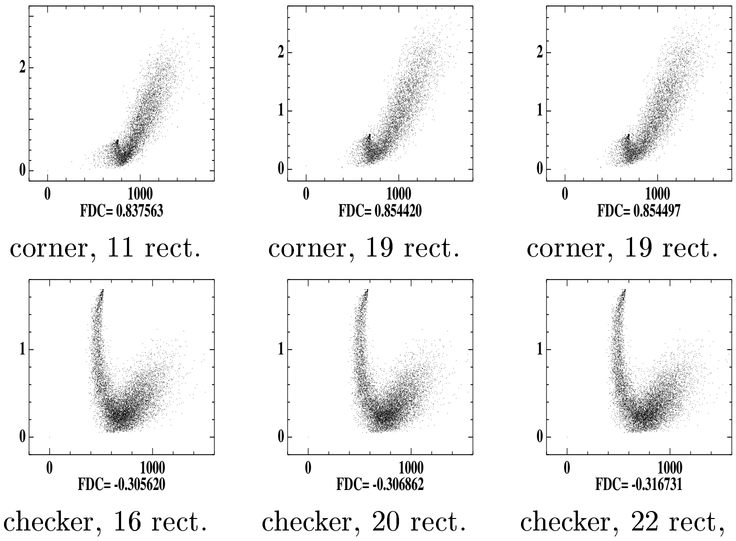


Figure 12: *FDC plots for different genotypes of the global optimum for the Rectangle representations. Both the plots and the coefficients are very similar, for the well-correlated corner problem as well as for the weakly-correlated checker-board problem.*

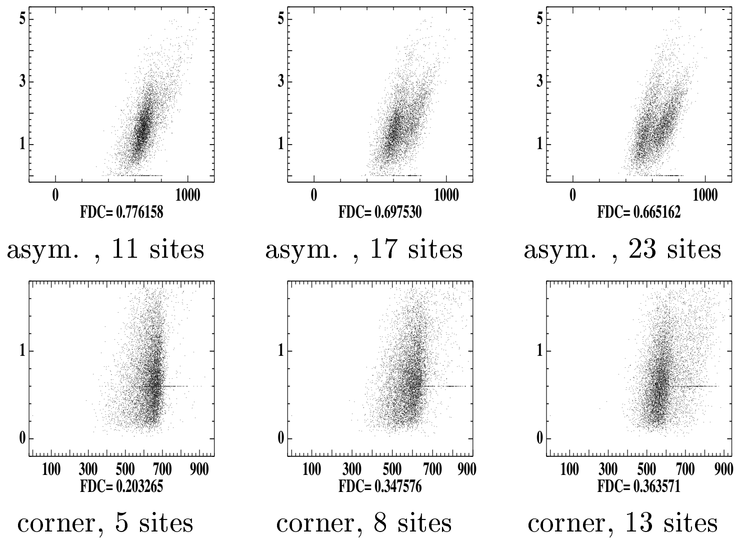
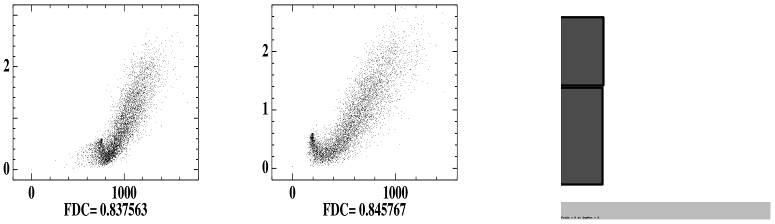
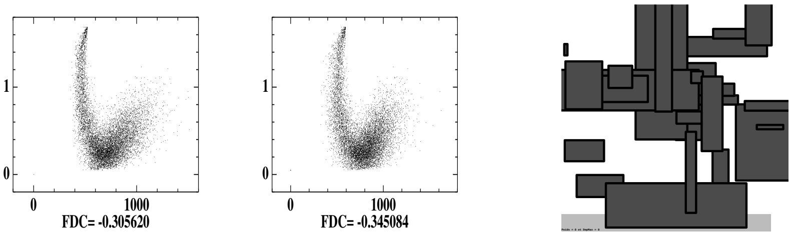


Figure 13: *FDC plots for different genotypes of the global optimum for the Voronoï representations. Slight variations can be observed on the plots and the coefficients, for the well-correlated asymmetric problem as well as for the weakly-correlated corner problem.*



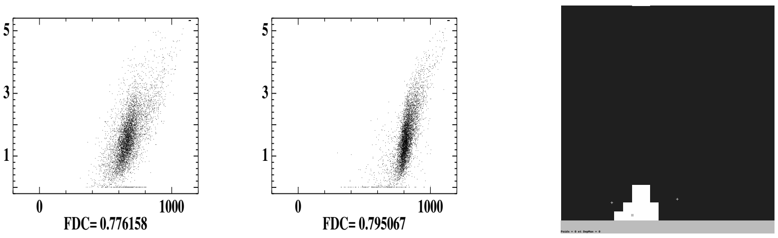
(a) Global FDC (b) Local FDC (c) Local optimum genotype

Figure 14: *Comparison of the global (a) and one local (b) FDC plots and coefficients for the 1-Rectangles representation on the corner problem. The actual optima of the sample is shown in (c).*



(a) Global FDC (b) Local FDC (c) Local optimum genotype

Figure 15: *Comparison of the global (a) and one local (b) FDC plots and coefficients for the 0-Rectangles representation on the checker-board problem. The actual optima of the sample is shown in (c).*



(a) Global FDC (b) Local FDC (c) Local optimum genotype

Figure 16: *Comparison of the global (a) and one local (b) FDC plots and coefficients for the Voronoï representation on the asymmetric problem. The actual optima of the sample is shown in (c).*

	Average (std dev.) of Local FDC coefficient Global FDC coefficient			
Distance / Problem	asymmetric	symmetric	corner	checker-board
Rectangles : holes based	0.9(0.002) 0.87	-0.3(0.05) -0.4	0.8(0.004) 0.75	-0.4(0.04) -0.43
Rectangles : area based	0.9(0.002) 0.93	-0.3(0.04) -0.23	0.8(0.01) 0.83	-0.3(0.02) -0.3
Rectangles : area based, with a constant $C(R_1, R_2) = C$	0.9(0.02) 0.91	-0.2(0.1) 0.06	0.4(0.26) 0.09	-0.3(0.04) -0.17
Voronoï : site based	-0.1(0.02) -0.03	-0.3(0.05) -0.33	-0.06(0.007) 0.08	0.18(0.33) -0.3
Voronoï : cell/site based	0.7(0.03) 0.77	0.4(0.19) 0.37	0.3(0.1) 0.25	0.7(0.02) 0.76
Bitarray : hamming	0.9(0.003) 0.93	0.4(0.2) 0.37	0.7(0.04) 0.76	0.3(0.2) 0.08
Bitarray : symmetric DT	0.03(0.03) 0.01	0.3(0.05) 0.35	0.12(0.03) 0.15	0.13(0.08) 0.26

Table 3: *Experimental results of local FDC stability, and comparisons with the global coefficients for all 4 test problems and different representation-distances.*

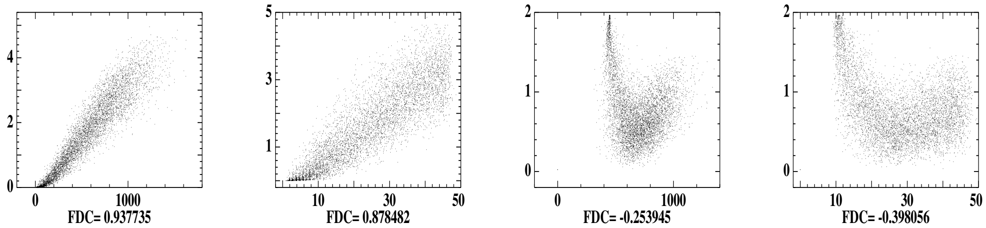
Special attention should be paid to the Area-based distance with constant parameter C ($C(R, S)$ is the scaling parameter of Equation 7 of Section 4.6.1). With constant C, the allelic distance δ_A is metric (see Lemma 1 of Appendix), whereas it does violate the triangular inequality when C is computed as described in section 7. However, Table 3 shows that the distance using constant value for C is not stable with respect to the Local FDC, exhibiting large deviations for the symmetric and corner problems: in the present case, heuristic tuning of the allelic distance seems to perform better than strict respect of mathematical properties.

Further, Table 3 shows that slightly greater differences are observed between local and global coefficients for the H-distance than for the A-distance. Moreover, the FDC plots of Figure 17 show higher dispersion of the same sample points for the Hole-distance (Figure 17-b and -d) than for the Area-distance (Figure 17-a and -c). Hence, though any of the two distances could be used, probably leading to the same conclusions, the A-distance will be used for all further results using one of the Hole-based representations (Rectangles or Triangles).

5.5.2 Voronoï representation

The situation is quite different for the Voronoï representation than for the holes representation. Four different distances have been designed for the Voronoï representation: the Site-based distances perform site-matching to compute the best matching between two individuals, and are based either on sites distances (S-distance of Equation 12) or on relative cell-areas (SA-distance of Equation 17); all alleles have the same weight 1. The Cell-based distances weight the alleles by their cell-areas, after matching them either by the site distances (AS-distance of Equation 15) or by cell DT-distance (C-distance of Equation 19).

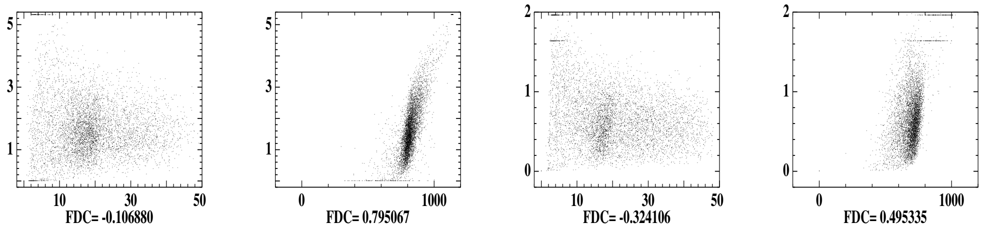
The first obvious conclusion is that site-based distances, and especially the S-distance, are quite unstable with respect to both the choice of a genotype for the global optimum (results not detailed here) and the local optimum, as can be seen in Table 3. Moreover, the FDC plots for the S-distance shown in Figure 18-a and -c indicates that this distance fails to capture any relevant



Area-distance Hole-distance Area-distance Hole-distance

Asymmetric problem Symmetric problem

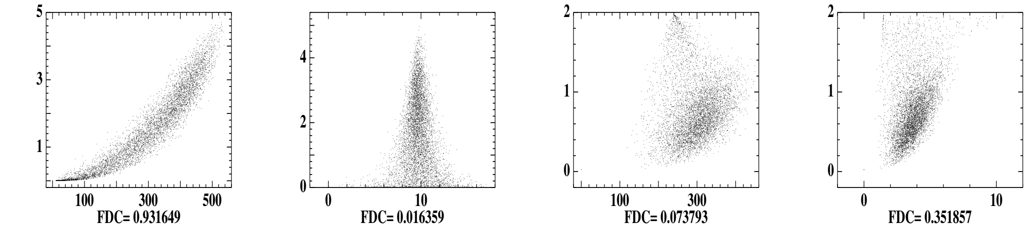
Figure 17: *Local FDC plots for 1-Rectangles representation on the asymmetric and the symmetric problems for the two distances.*



Site distance Cell/Site distance Site distance Cell/Site distance

Asymmetric problem Symmetric problem

Figure 18: *Local FDC plots for Voronoï Representation on the asymmetric and the symmetric problems for two different distances: the site-based distance does not give any useful information.*



Hamming distance DT-distance Hamming distance DT-distance

Asymmetric problem Symmetric problem

Figure 19: *Local FDC plots for bitarray representation on the asymmetric and the symmetric problems for two different distances: The Hamming distance for the asymmetric problem behaves like a 'count the ones' distance and completely biased the FDC coefficient.*

information for the given problem. Though the definition based on site distances are simple and intuitive, they cannot take into account the important features of the fitness landscape.

Meanwhile, the cell-based distances demonstrate much more efficiency and stability in describing correlation, though none of them does give results as good as those of both the A-distance and the H-distance of the Holes representation: they are more sensitive to a change in the genotype, either global (see Figure 13) or local (see Figure 16). No significant difference was observed between the Cell/Site distance or the pure Cell distance. Nevertheless, the CS-distance wil be used for all further results involving the Voronoï representation, as being less costly to compute.

A tentative explanation for these differences in performances of the different distances for the Voronoï representation is related to the epistasis of that representation, i.e. the way the expression of an allele in the phenotype depends on the other alleles: it appears that the more the cells are taken into account in the distance, the better the results. But the shapes and areas of the Voronoï cells are only defined by the relative positions of the neighboring Voronoï sites, and are thus directly connected to the epistasis.

5.5.3 Bitarray representation

It has already been demonstrated in Section 5.2.3 that uniform random sampling in the bitstring space was hopeless for the problems at hand. Therefore, all results of this section were obtained on random samples generated by the Rectangle representation. Results of Table 3 show a clear advantage in terms of stability for the DT-distance (see Section 4.3.2) over the standard Hamming distance. This is acknowledged by the plots of Figure 19: the plot for the asymmetric problem (Figure 19-a) indicates a FDC almost perfect, because the asymmetric problem is not far from a onemax problem. On the opposite, the uniform cloud of Figure 19-c only reflects the symmetry of the problem, for which the Hamming distance does not bring any useful information. On the other hand, the DT distance seems to capture something of the problem difficulty, but this still needs to be confirmed by actual runs (see Section 6.3). Nevertheless, for further use of the bitarray representation, only the DT-distance will be considered.

6 Validating FDC

In this section, the expected outcomes of FDC will at last be harvested: Section 6.1 presents results of FDC used as a measure of problem difficulty, in the line of Jones'work [24], but on variable length representations. But another important motivation for this study was the choice of a representation for a given problem, addressed in Section 6.2. First, all partial results regarding the bitarray representation will be presented together, showing that the poor performances of that representation for the problems at hand, *a priori* predicted by numerical arguments given in section 4.4, might have been foreseen by a detailed FDC study. Then the variable length representations, namely both Rectangle representations (Section 4.5.1), the triangle representation (Section 4.6) and the Voronoï representation (Section 4.7), will be compared with respect to FDC predictions on each of the four test problems.

To make these points clear, FDC "predictions" have to be compared to actual runs of an Evolutionary Algorithm. The instance of Evolutionary Algorithm used thereafter can be described as standard GA: it uses the generational GA scheme with 2-tournament selection, one offspring per selected parent(s), 0.6 crossover rate, 0.1 mutation rate per individual, plus EP-like tuning of the standard deviations of the Gaussian mutations of the continuous variables [12], all

	Asymmetric		Symmetric		Corner		Checker	
	best	median	best	median	best	median	best	median
1-Rect.	4e-06	4.8e-05	0	2.8e-05	0.00024	0.00081	0.000355	0.00107
0-Rect.	8.5e-05	0.00027			0.00039	0.0031		
Triangle	8.33e-06	3.5e-05	0	1.39e-05	4e-06	2.64e-05	0.0028	0.0084
Voronoï	0	0.000232	0	3e-06	2e-06	3.6e-05	7.1e-05	0.000282

Table 4: *Off-line performances of GAs for all test problems and variable length representations. The best and median result from 21 runs of 300 generations each are presented. The null best means that the actual global solution was found.*

offspring replacing all parents. Though some experiments [39] suggested that strong epistasis (as in the Voronoï representation) might rather favor Evolution Strategy scheme – μ parents give λ offspring with 1.0 mutation rate, the best μ among offspring, or parents + offspring, deterministically replacing the parents – the study of the influence of the evolution paradigm, and its choice with respect to epistasis or other criterion, is beyond the scope of this paper.

For each setting of experimental conditions (problem, representation, initialization procedure, ...), 21 independent runs have been performed. However, the forthcoming plots presenting the online performances of these runs are the *median* curves of fitness vs number of generations, rather than the usual averages: At each generation, the best individuals of all runs are sorted according to fitness, and the median value (here, the 11^{th} out of 21) is retained. As discussed in [15], whereas using the average values (together with standard deviation and/or T-test results) make the implicit assumption of normal repartitions for the sampled values, choosing the median values seems more adapted to non-parametric statistics. Moreover, when dealing with fitness values that can take very small values (as the mean square errors of the mechanical behaviors for the mechanical problems at hand), averages hardy reflect actual behavior of the algorithm: Error values for different runs can be of fairly different orders of magnitude. Hence, a few large values might hide a larger number of very small errors.

All FDC plots and coefficients useful for the present section are presented in Figure 20. Some of these plots were presented separately before, to illustrate different aspects of FDC behavior. It seemed however useful to group together the ones resulting from the different choices made in section 5. In that Figure 20, all plots in the same column address the same problem, and have been scaled identically on the fitness axis, while all plots on the same line use the same representation/distance, with same order of magnitude on the horizontal (distance) axis. Note that the plots are all Local FDC plots (see section 2.4), in a tentative to simulate as much as possible a realistic situation where the global optimum is unknown.

Further, Table 4 presents the corresponding off-line performances of the actual GA runs (best and median out of 21 runs) that wil be used in conjunction with standard plots of online evolution of fitness along generations in the following.

6.1 Problem difficulty

This section uses FDC to compare problem difficulties for a given representation. Figure 20 should be read line by line.

Comparing the FDC coefficients and plots of Figure 20 with the results of actual runs, in Figure 21 for the 1-rectangle representation (first line of Figure 20), and in Figure 22 for the

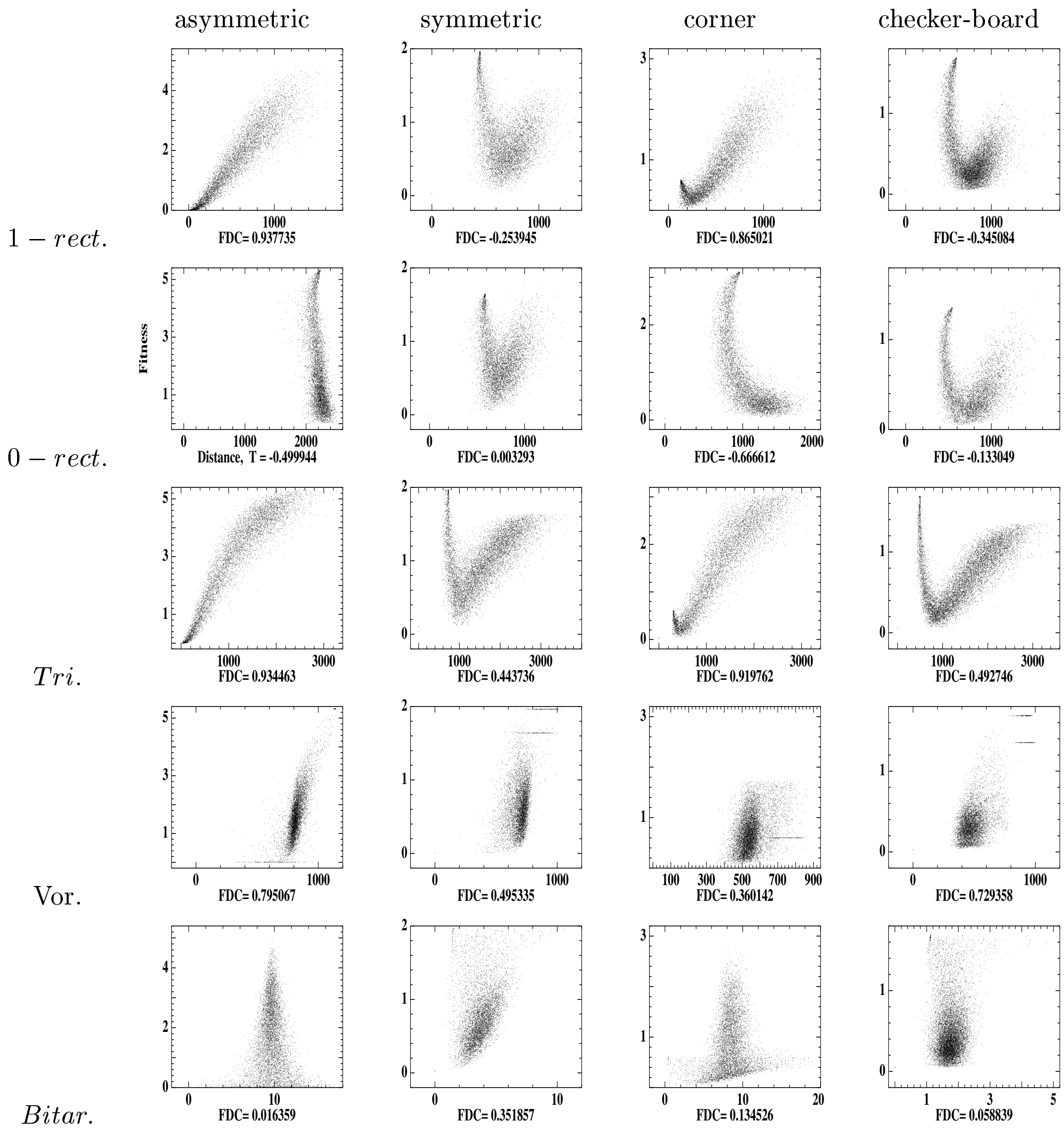
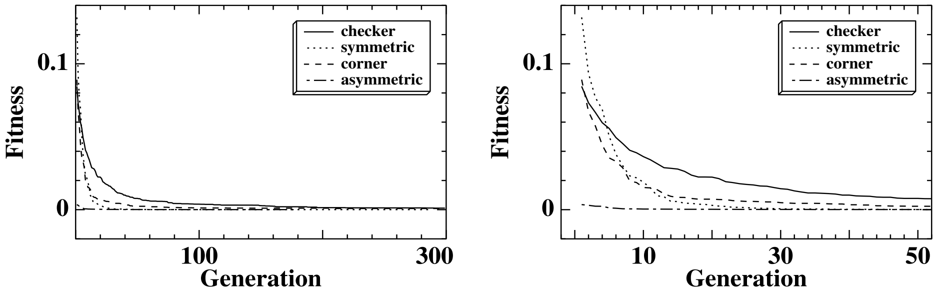


Figure 20: *Local FDC plots for different representations on the four test problems: rows give hints on relative problem difficulty for a given representation, columns allow to choose between representations for a given problem.*

	whole sample	best 3000	best 2000	best 1000
asymmetric	0.93	0.9	0.87	0.81
corner	0.86	0.5	0.3	0.3
symmetric	-0.2	0.2	0.2	0.18
checker-board	-0.3	0.05	0.06	0.08

Table 5: *FDC coefficient when only parts of the sample are considered, for the 1-Rectangle representation.*



1-Rectangle Representation — Zoom on the first generations

Figure 21: *Actual runs of GAs with the 1-Rectangle representation on the four test problems.*

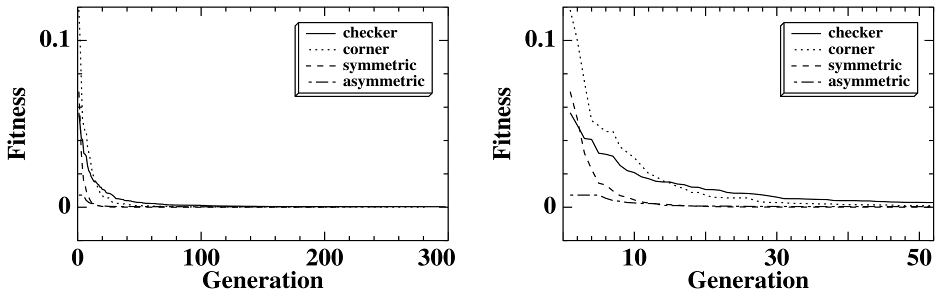
Voronoï representation (fourth line of Figure 20), it seems that FDC coefficients only represent what is happening at the beginning of evolution: Making a zoom on the first generations, the problem difficulty is almost correctly described by the correlation measure. However, on later generations, comparative performance order of the GA might be totally different, as witnessed by the off-line performances given in Table 4: for both the 1-rectangle and the Voronoï representations, the best result is achieved on the symmetric problem, where the actual solution is found at least once.

In order to try to make improve the predictive accuracy of the FDC coefficients, even only in the first generations, the FDC coefficients were computed using only the best individuals in the whole sample (Table 5). One can see that indeed the problem difficulty is better described by such partial correlation measures.

Nevertheless, it is quite difficult to compare errors on different problems, that might be of different orders of magnitude even when at the same distance from the global optimum: some pre-scaling (yet to be defined) might help in that task, both on the fitness axis for different problems and for the FDC coefficients for different representations.

6.2 Choice of a representation

In this section, FDC is studied from the point of view of *a priori* choice of representation on a given problem. Figure 20 should be read column by column. The first subsection will handle the special case of the – fixed length – bitarray representation, while the variable length representations will be compared in section 6.4.



Voronoï Representation   Zoom on the first generations

Figure 22: *Actual runs of GAs, with the Voronoï Representation, on the four test problems.*

6.3 Forget about Bitarrays

It has already been argued in Section 4.4 that the bitarray representation need not to be considered for the problems of shape identification or optimization, due to its high dependency on the underlying mesh. But such considerations are very specific to the class of problems considered here. This subsection will try to determine whether the same conclusions could have been reached using only FDC-based arguments.

Two distances have been defined on the bitarray representation (see Section 4.2): The standard Hamming distance, and a distance based on the Distance Transformation concept. Following the results of section 5.5.3, only the DT-distance will be used now.

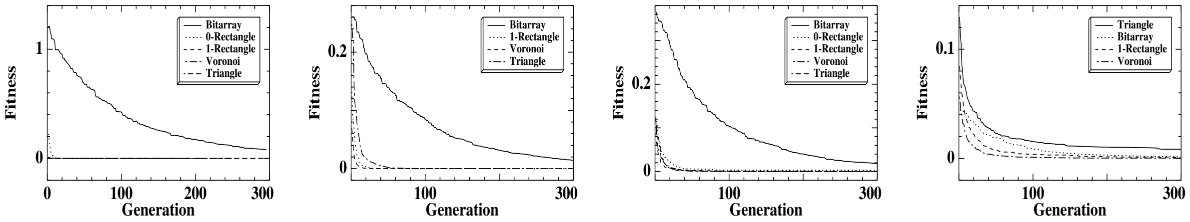
The first hint about the inadequacy of the bitarray representation-distance is given by the disastrous results of the uniform initialization procedure (section 5.2). Even if mechanical considerations favor the initialization by large homogeneous zones, it is somewhat worrying for the representation that its "natural" initialization procedure gives such horrible results. But the second bad news come from the last line of Figure 20: though obtained using the same samples that the 1-rectangle line, these FDC plots also seem to predict poor performances.

It is nevertheless interesting to have a look at the actual runs performed using the bitarray representation (Figure 23). The bitarray representation is outperformed by all the variable length representations presented in this paper, with the noticeable exception of the checker-board problem, where it seems to catch up most of them, even outperforming the triangle representation. Note however that even on the checker-board problem, the off-line performances of the bitarray representation are still behind those of the Voronoï representation.

However, it would probably be interesting to do the same comparison on a problem whose solution would be a uniform random choice of all elements of the underlying mesh. On such academic problem, all other representations might perform as bad as the bitarrays. The *a priori* mechanical argument would still make other representations preferable, but independently of any FDC results.

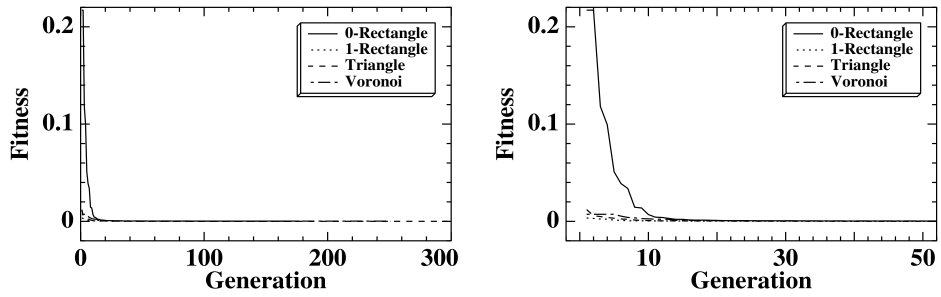
6.4 Choice of a variable length representation

This subsection compares the four variable length representations whose FDC plots are presented on Figure 20, namely the 1- and 0-rectangles, the (1-)triangles and the Voronoï representations. The goal here is to see if the results of FDC plots could allow to choose between these represen-



(a) (b) (c) (d)

Figure 23: *Comparative runs involving the bitarray representation. The bitarray runs use a random sample generated by the Rectangle representation, as described in section 5.2: bitarrays performances are lousy, except for the checker-board problem.*



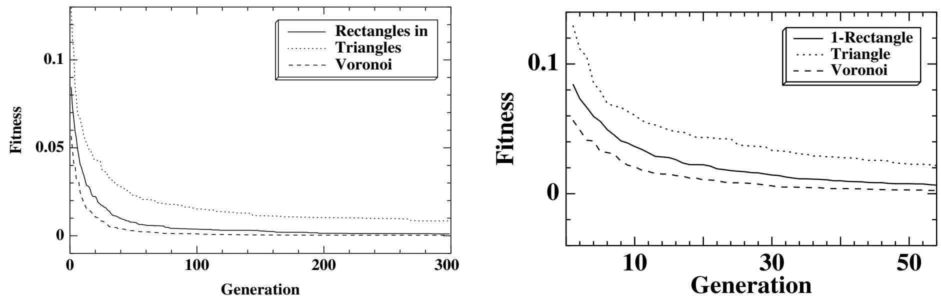
Asymmetric problem   Zoom on the first generations

Figure 24: *Actual runs of GAs, with the four variable length representations on the Asymmetric problem: The 0-Rectangle representation is largely outperformed by the three others.*

tations. Of course, for the purpose of validation, such FDC-based conclusions will be checked against actual GA runs.

The most striking difference in the FDC plots of Figure 20 is the poor FDC coefficient of the 0-rectangle representation on both the asymmetric problem and the corner problem. In both cases, all other representations show good or very good FDC coefficients (the 0-triangle, not presented here, had the same defect than the 0-rectangle). Actual GA runs on the asymmetric problem (Figure 24) confirm the poor performance of the 0-rectangle representation, in the early generations as well as in the long run (see Table 4). The situation is different for the symmetric and the checker-board problems as the optimum topology of both problems is symmetric. Due to the different mechanical properties of materials 1 and 0, the mechanical problem is not symmetric which explains the slight difference in the plots. Note that on both asymmetric and corner problems, online performance of the Voronoï representation is almost as good as that of the 1-rectangle, in spite of lower FDC coefficients, while being even better when off-line performances are considered (Table 4).

On the opposite side, the Voronoï representation is the best choice for both the symmetric and the checker-board problems: The FDC coefficients are the highest, and the plots of Figure 25 attest that Voronoï representation performs indeed better. This is confirmed by the off-line



Checker-board problem Zoom on the first generations

Figure 25: *Actual runs of GAs, with different representations, on the Checker-board problem: The Triangle representation is outperformed by the two others.*

results of Table 4. However, what was not predictable from the plots of Figure 20 is that the 1-rectangles would perform better than the triangles. But this should be put together with the outlook of the FDC plot itself, as already mentioned in section 5.2 about the initialization procedure: Procedure 2 was used for the triangles, but the evolution probably brings rapidly the population in regions corresponding to the left part of the FDC plot, where the correlation between distance and fitness is strongly negative as shown in figure 9 when procedure 3 is used.

Some partial conclusions can be drawn from these results. From the FDC point of view, it is clear that the FDC coefficient alone is not sufficient. First, when comparing different representations, it needs some pre-scaling – the Voronoï coefficient seem under-evaluated when comparison with actual runs is later performed. Second, the shape of the cloud of points of the FDC plot can help to better understand the underlying fitness landscape. But over all, FDC analysis does not help much to choose based on the off-line performance.

From the user point of view, regarding the possible representations for problems of two dimensional partitions identification, there are clear indications that the 1-holes representations – where the holes describe the topology of material 1 – should be preferred when the solution contains more of material 0 than material 1. Moreover, FDC analyses do reflect that without any a priori knowledge. Further, the Voronoï representation seems more adapted to problems where both materials are equally distributed. Note however, that the FDC results also strongly depends on the initialization procedure, as discussed in section 5.2.

7 Discussion and Related Work

This section discusses the pros and cons of the FDC approach that has been proposed, in the light of related work from EC literature. Meanwhile, some further directions of research are sketched to overcome present limitations.

7.1 Random sampling

The first lesson from the experiments presented in this paper is that great attention must be paid to the random sampling of the genotype space: The stability-based procedure was designed to adjust the size of the random sample with respect to both the complexity of the search space and the ruggedness of the fitness landscape. This procedure demonstrated a large influence on the FDC results compared to the fixed-size sampling used in [24] for large problems.

But being based on static random sampling of the search space, FDC encounters the limitation described in [19] on for the extended LongPath problem of [20], also detailed in Section 2.2 for the same problem. Such double characteristics of the fitness landscape certainly fools the FDC analyses (see the plots of Section 2.2). And even the problem-dependent procedure proposed in Section 2.3 in such an ugly situation cannot be of any help when the "interesting" part of the fitness landscape almost surely escapes any random sampling.

Note that this limitation is met by all other statistics-based *a priori* estimators developed in the Evolutionary Optimization framework. The Virtual Genetic Algorithm of Grefenstette [18], for instance, uses statistics on the fitness of random samples to derive a simple statistical model for the fitness landscape. This model then allows performing the trial-and-error process of designing the most efficient algorithm at low cost, replacing the exact computation of the fitness value by the estimation based on that model. It is an elegant alternative to the standard brute-force parameter estimation process, as first described systematically in [37]. However, in the case of large search spaces, the simple statistical model generally proves of poor help, as the fitness of all individuals in the random sample is many orders of magnitude worse than that of the target optimum: the same situation has been encountered as one of the bitarray representation defects (section 6.3).

7.2 Initialization procedures

When no obvious uniform procedure is available, FDC seems a good tool for comparing different initialization procedure, experimentally confirmed in the early stage of the evolutionary optimization (Section5.2).

Little attention has been paid to initialization procedures in EA literature. One of the main reason is probably that uniformly distributed individuals are fairly easy to draw in the most common representations frames (i.e. bitstrings and real-valued parameters with prescribed bounds). It has been recognized, though, that a biased initialization using a preliminary EA-run could greatly improve the overall results [10, 41]. Moreover, recent work in the GP framework [21], as well as the preliminary experiments of this paper, show that a careful choice of initialization procedure can enhance the performance of the further EA. It seems however, from the actual runs presented here, that such enhancement does not last very long during evolutionary runs: On-going work will address this important issue in a more detailed way.

7.3 FDC and problem difficulty

The – unrealistic – requirement of the knowledge of the global optimum of the problem at hand was relaxed by the introduction of the Local FDC analysis (Section 2.4). Moreover, where the global FDC gave information on the outlook of the fitness landscape seen from the global optimum, local FDC gives information about other regions of the fitness landscape. The stability of the results (FDC coefficients and aspects of the fitness $\times$ distance plots) is another hint for the reliability of FDC as analysis tool for the problem and representation at hand.

This approach is somehow related to the study of landscape modality in [20], where problem difficulty is thought in term of number and repartition of local optima. The stability of the local FDC (coefficients and fitness $\times$ distance plots) with respect to the local optimum, and its similarity with its global counterparts is a sign of homogeneity of the fitness landscape, and hence should enforce the confidence in FDC results and predictions.

Another closely related work, already mentioned in the introduction, is the formae theory and the study of the variance of the fitness versus the precision of the formae [34]. The notions of formae and precision could be replaced by those of sphere of given radius once a distance is defined. The Local FDC analysis is a first step in that direction.

But this also suggests that other measures could be computed beside the FDC coefficient to make the picture more precise. This includes variances coefficients for given subsets, but also results of some clustering analysis. Further, the comparison of FDC plots, now calibrated by the eye of the user, could be made more precise using image analysis tools, e.g. contour and edge detection and local wavelet decomposition.

7.4 FDC and Evolution

Probably the most severe limitation of the importance of FDC results comes from the total separation between FDC analyses and the actual optimization algorithm to be used further. Due to that separation, and to the obvious limitation of the part of the search space that is actually sampled by the static random sampling procedure, FDC results can only predict accurately the very beginning of the evolution process. Two directions can be thought of to remedy that situation:

First, FDC coefficients can be computed all along evolution, providing insight in the regions of the search space where the Evolutionary Algorithm actually searches, and maybe allowing it to intelligently modify the algorithm itself. Such an approach is somewhat similar to the density-of-state concept [2, 36], though with a different estimator. The density-of-state uses the number of genotypes having a given fitness to estimate problem difficulty, but the estimation of the density-of-state is carried on during the course of some Metropolis algorithm. The exploration is hence not limited to blind random sampling of the genotype space. But FDC plots contains the same information: the number of dots in any horizontal slice of the FDC plots is an estimation of the density-of-state. An interesting experiment would be to compare the results of the Metropolis algorithm to those of dynamical FDC – yet to be run.

Second, the effects of the genetic operators could be incorporated in the FDC framework, either in the line of [14], by studying the correlation between the fitness of the parent with that of the offsprings, or by applying the FDC technique described in this paper to some other distances defined through the genetic operators themselves. For instance, a distance can be defined between two points of the genotype space as the minimum "length" of mutation paths between these points. The careful study of the stability of such distances with respect to the different criteria experimented in Section 5 could indeed bring new light on the interaction between genetic operators and fitness landscapes.

8 Summary and Conclusion

The FDC tools introduced by Jones in the bitstring framework [24] have been extended to variable length representation involving continuous parameters. General hints have been given to design genotypic distances from allelic similarity measures, including a heuristic allele matching algorithm. Moreover, two limitations of the seminal work of Jones were addressed: The size of the random sample was automatically adjusted to the problem characteristics; A Local FDC analysis was proposed to handle the realistic cases where the global optimum is actually unknown.

All these ideas were experimentally validated on some representations of two-dimensional shapes [39] applied to the problem of mechanical inclusion identification [40]: the traditional fixed-length discrete bitarray representation and the variable length continuous Holes and Voronoï representations.

Following the general guidelines, different distances were designed for each of the test representations. Though the resulting functions are not formal mathematical distances, all implementations were found numerically "close" to actual distance, and were applied to perform FDC analyses on the test problems.

Further, the stability of FDC results with respect to both the genotype of the global optimum – when available – and the random value of the "local" optimum used for the Local FDC analysis allowed to choose among different possible distances for the same representation, avoiding the false results due to non–adapted distance: The brute Hamming distance for the bitarray representation gives lousy results when compared to the two-dimensional DT-distance, and due to the high degree of epistasis of the Voronoï representation, purely genotypic distances based on Voronoï sites only were outperformed by the cell-based distances, taking implicitly into account some phenotypic expression of all alleles.

Keeping these limitations in mind, FDC demonstrated nevertheless to be quite accurate in predicting the behavior of further GA runs, at least in the early generations. General conclusions regarding the choice of representation can be drawn for problem classes: Based on FDC results, a broad conclusion could be to use Voronoï representation rather than some Holes representation when the solution contains equal quantities of both materials, and the Hole representation corresponding to the less present material in the other case. However, the reliability of FDC predictions degrades after the early generations: the off-line performances seem to indicate that Voronoï representation actually slightly outperform Holes representation on all problems except the asymmetric problem, which could not be seen on FDC plots. Nevertheless, FDC results do detect weird situations, as demonstrated on the caricature asymmetric problem for the 0-rectangle representation.

So FDC cannot – and does not – pretend to be the ultimate *a priori* analysis tool. The example of the bitarray representation shows that *a priori* considerations about the adequacy of the representation to the problem at hand can still be useful before blindly giving credit to the brute FDC results. On the other hand, though, the study of the stability of different FDC coefficients with respect to variations of experimental conditions (choice of the local optimum, of the genotype of the global optimum, of the cutoff in the fitness space) gradually build up a picture that should lead to question the representation at hand.

Finally, the main limitation of FDC analysis – the decorrelation with the Evolutionary Algorithm – must be addressed in order to reliably and accurately link FDC results to actual evolutionary runs.

9 Appendix

Lemma 1. Let δ be defined on the space of alleles by:

$$\delta(a_1, a_2) = A(a_1) + A(a_2) - 2g(A(a_1), A(a_2))e^{\frac{-d(a_1,a_2)}{C}},$$

where $A(a_1)$ is the area of allele a_1, d is a metric distance, and $g : \mathbb{R}_+^2 \to \mathbb{R}_+$ satisfies

$$g(kX, kY) = kg(X, Y) \; forall \; (X, Y) \in \mathbb{R}^2 \; and \, k \in \mathbb{R}_+$$

$$g(X, Y) \leq min(X, Y)$$

and,

$$\forall(\alpha,\beta) \in \mathbb{R}_+^2 \; 1 + g(\alpha,\beta) \geq g(\alpha,1)) + g(1,\beta).$$

Then δ is a metric distance on the space of alleles.

Proof. To prove the triangular inequality, consider 3 alleles a_1, a_2, a_3. Then:

$$\delta(a_1,a_2) = A(a_1) + A(a_2) - 2g((A(a_1),(A(a_2))exp(-d(a_1,a_2)/C)$$
$$\delta(a_2,a_3) = A(a_2) + A(a_3) - 2g((A(a_2),(A(a_3))exp(-d(a_2,a_3)/C)$$
$$\delta(a_1,a_3) = A(a_1) + A(a_3) - 2g((A(a_1),(A(a_3))exp(-d(a_1,a_3)/C)$$

Let T, α and β be defined by:

$$T = \delta(a_1,a_2) + \delta(a_2,a_3) - \delta(a_1,a_3)$$
$$A(a_1) = \alpha A(a_2)$$
$$A(a_3) = \beta A(a_2)$$

Then T has the same sign than

$$T_1 = 1 - g(\alpha,1)exp(-d_{12}/C) - g(1,\beta)exp(-d_{23}/C) + g(\alpha,\beta)exp(-d_{13}/C)$$

Knowing that d is a metric distance:
$exp(-d(a_1,a_3)/C) > exp(-d(a_1,a_2)/C)exp(-d(a_2,a_3)/C)$, so $T_1 \geq 0$ holds if

$$g(\alpha,\beta)exp(-d(a_1,a_2)/C)exp(-d(a_2,a_3)/C) \geq g(\alpha,1)exp(-d(a_1,a_2)/C)+g(1,\beta)exp(-d(a_2,a_3)/C)-1$$

Since $0 \leq exp(-d/C) \leq 1$ for all values of d, and from the inequality $1 + g(\alpha,\beta) \geq g(\alpha,1)) + g(1,\beta)$, the latter inequality holds, and hence δ is a distance.

Remark: The following functions obviously satisfy the conditions of Lemma 1:

$$\begin{aligned} g(X,Y) &= 2min(X,Y) \\ g(X,Y) &= 2min(X,Y) \text{ if X and Y have the same label} \\ &= 0 \text{ otherwise} \end{aligned}$$

Lemma 2. Let δ be defined on the space of alleles by:

$$\delta(a_1,a_2) = 2 - 2g(A(a_1),A(a_2))e^{\frac{-d(a_1,a_2)}{C}},$$

where $A(a_1)$ is the area of allele a_1, d is a metric distance, and $g : \mathbb{R}_+^2 \to \mathbb{R}_+$ satisfies

$$g(kX,kY) = g(X,Y) \, forall \, (X,Y) \, \in \mathbb{R}^2 \, and \, k \in \mathbb{R}_+$$
$$g(X,Y) \leq 1$$

and,

$$\forall(\alpha,\beta) \in \mathbb{R}_+^2 \; 1 + g(\alpha,\beta) \geq g(\alpha,1)) + g(1,\beta).$$

Then δ is a metric distance on the space of alleles.

Proof. The proof is a straightforward modification of the proof of lemma 1.

Remark: The following functions used in the paper satisfy the conditions of Lemma 2:

$$
\begin{aligned}
g(X,Y) &= 1 \\
g(X,Y) &= \tfrac{2min(X,Y)}{X+Y} \\
g(X,Y) &= \tfrac{2min(X,Y)}{X+Y} \text{ if X and Y have the same label} \\
&= 0 \text{ otherwise}
\end{aligned}
$$

Lemma 3. In the genotype space of unordered lists of **fixed length** n of alleles, consider the function Δ between two chromosomes $x = (a_i)_{i \in [1,n]}$ and $x' = (a'_i)_{i \in [1,n]}$ defined by

$$\Delta(x, x') = \sum_{i=1}^{n} \delta(a_i, a'_{\mu(x,x')(i)}),$$

where δ is a metric distance on the space of alleles, and $\mu(x, x')$ is the "best" matching between the alleles of x and those of x', i.e. that minimizes $\Delta(x, x')$.
Then Δ is a metric distance on the genotype space.

Proof. To prove the triangular inequality, consider three lists of n alleles x_1, x_2, x_3. Then, by definition,

$$\Delta(x_1, x_2) = \sum_{i=1}^{n} \delta(a_i^1, a^2_{\mu(x_1,x_2)(i)})$$

$$\Delta(x_2, x_3) = \sum_{i=1}^{n} \delta(a_i^2, a^3_{\mu(x_2,x_3)(i)})$$

$$\Delta(x_1, x_3) = \sum_{i=1}^{n} \delta(a_i^1, a^3_{\mu(x_1,x_3)(i)})$$

Let $d_{aux}(x_1, x_3) = \sum_{i=1}^{n} \delta(a_i^1, a^3_{\mu(x_2,x_3)(\mu(x_1,x_2)(i))})$
Since δ is a metric distance,

$$\Delta_{aux}(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$$

So because $\mu(x, x')$ is the matching that minimizes the expression $\Delta(x, x')$,

$$\Delta(x_1, x_3) \leq \Delta_{aux}(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$$

References

[1] J. Antonisse. A new interpretation of schema notation that overturns the binary encoding constraint. In J. D. Schaffer, editor, *Proceedings of the 3^{rd} International Conference on Genetic Algorithms*, pages 86–91. Morgan Kaufmann, June 1989.

[2] T. Asselmeyer, W. Ebeling, and H. Rosé. Smoothing representation of fitness landscape – the genotype-phenotype map of evolution. *Biosystems*, 39(1):63–76, 1996.

[3] T. Bäck. *Evolutionary Algorithms in theory and practice.* New-York:Oxford University Press, 1995.

[4] J.-D. Boissonnat and M. Yvinec. *Géométrie algorithmique.* Ediscience International, 1995.

[5] G. Borgefors. Distance transformation in arbitrary dimention. *Computer Vision, Graphics, and Image Processing*, 27:321–345, 1984.

[6] R. A. Caruna and J. D. Schaffer. Representation and hidden bias : Gray vs binary coding for genetic algorithms. In *Proceedings of ICML-88, International Conference on Machine Learning.* Morgan Kaufmann, 1988.

[7] R. Cerf. *Une théorie assymptotique des algorithmes génétiques.* PhD thesis, Université de Montpellier II, March 1994.

[8] C. D. Chapman, K. Saitou, and M. J. Jakiela. Genetic algorithms as an approach to configuration and topology design. *Journal of Mechanical Design*, 116:1005–1012, 1994.

[9] A. Constantinescu. *Sur l'identification des modules élastiques.* PhD thesis, Ecole Polytechnique, June 1994.

[10] H. de Garis. Genetic programming : building artificial nervous systems using genetically programmed neural networks modules. In R. Porter and B. Mooney, editors, *Proceedings of the 7^{th} International Conference on Machine Learning*, pages 132–139. Morgan Kaufmann, 1990.

[11] L. Eshelman and J. D. Schaffer. Real-coded genetic algorithms and interval-schemata. In L. D. Whitley, editor, *Foundations of Genetic Algorithms 2*, pages 187–202, Los Altos, CA, 1993. Morgan Kaufmann.

[12] D. B. Fogel. An analysis of evolutionary programming. In D. B. Fogel and W. Atmar, editors, *Proceedings of the 1^{st} Annual Conference on Evolutionary Programming*, pages 43–51. Evolutionary Programming Society, 1992.

[13] D. B. Fogel. *Evolutionary Computation. Toward a New Philosophy of Machine Intelligence.* IEEE Press, Piscataway, NJ, 1995.

[14] D. B. Fogel. Phenotypes, genotypes and operators in evolutionary computation. In D. B. Fogel, editor, *Proceedings of the Second IEEE International Conference on Evolutionary Computation.* IEEE, 1995.

[15] C. M. Fonseca and P. J. Fleming. On the performance assessment and comparison of stochastic multiobjective optimizers. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the 4^{th} Conference on Parallel Problems Solving from Nature*, number 1141 in LNCS, pages 584–593. Springer Verlag, Sept. 1996.

[16] D. E. Goldberg. A theory of virtual alphabets. In Hans-Paul Schwefel and Reinhard Männer, editors, *Proceedings of the 1^{st} Parallel Problem Solving from Nature*, pages 13–22, 1991.

[17] D. E. Goldberg, K. Deb, and J. H. Clark. Genetic algorithms, noise and the sizing of populations. *Complex Systems*, 6:333–362, 1992.

[18] J. J. Grefenstette. Virtual genetic algorithms: First results. Technical Report AIC-95-013, Navy Center for Applied Research in Artificial Intelligence, February 1995.

[19] C. Höhn and C. Reeves. Are long path problems hard for genetic algorithms? In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the 4^{th} Conference on Parallel Problems Solving from Nature*, volume 1141 of *LNCS*, pages 134–143. Springer Verlag, 1996.

[20] J. Horn and D.E. Goldberg. Genetic algorithms difficulty and the modality of fitness landscapes. In L. D. Whitley and M. D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 243–269. Morgan Kaufmann, 1995.

[21] H. Iba. Random tree generation for genetic programming. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the* 4^{th} *Conference on Parallel Problems Solving from Nature*, volume 1141 of *LNCS*, pages 144–153. Springer Verlag, 1996.

[22] C. Z. Janikow and Z. Michalewicz. An experimental comparison of binary and floating point representations in genetic algorithms. In R. K. Belew and L. B. Booker, editors, *Proceedings of the* 4^{th} *International Conference on Genetic Algorithms*, pages 31–36. Morgan Kaufmann, 1991.

[23] E. Jensen. *Topological Structural Design using Genetic Algorithms.* PhD thesis, Purdue University, November 1992.

[24] T. Jones and S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In L. J. Eshelman, editor, *Proceedings of the* 6^{th} *International Conference on Genetic Algorithms*, pages 184–192. Morgan Kaufmann, 1995.

[25] C. Kane. *Algorithmes génétiques et Optimisation topologique.* PhD thesis, Université de Paris VI, July 1996.

[26] C. Kane and M. Schoenauer. Genetic operators for two-dimensional shape optimization. In J.-M. Alliot, E. Lutton, E. Ronald, M. Schoenauer, and D. Snyers, editors, *Artificial Evolution*, number 1063 in LNCS. Springer Verlag, Septembre 1995.

[27] H. Kargupta. Signal-to-noise, crosstalk, and long range problem difficulty in genetic algorithms. In L. J. Eshelman, editor, *Proceedings of the* 6^{th} *International Conference on Genetic Algorithms*, pages 193–200. Morgan Kaufmann, 1995.

[28] P. Ladeveze, N. Reynier, and D. Nedjar. Parametric corrections of finite element models using modal tests. In M. Tanaka and H. D. Bui, editors, *IUTAM Symposium on Inverse Problems in Engeneering Mechanics.* Springer Verlag, May 1993.

[29] P. Lee and T. Nagao. Hierarchical description of two dimensional shapes using a genetic algorithm. In D. B. Fogel, editor, *Proceedings of the Second IEEE International Conference on Evolutionary Computation*, 1995.

[30] D. Luu. Algorithmes génétiques et diagrammes de voronoî : Optimisation et identification de modèles géologiques, Sept. 1996. Rapport de DEA d'Analyse Numérique de l'Université de Paris VI.

[31] Z. Michalewicz. *Genetic Algorithms+Data Structures=Evolution Programs.* Springer Verlag, New-York, 1996. 3rd edition.

[32] F. P. Preparata and M. I. Shamos. *Computational Geometry: an introduction.* Springer Verlag, 1985.

[33] N. J. Radcliffe. Equivalence class analysis of genetic algorithms. *Complex Systems*, 5:183–20, 1991.

[34] N. J. Radcliffe and P. D. Surry. Fitness variance of formae and performance prediction. In L. D. Whitley and M. D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 51–72. Morgan Kaufmann, 1995.

[35] N. J. Radcliffe and P. D. Surry. Fundamental limitations on search algorithms: Evolutionary computating in perspective. In J. van Leeuwen, editor, *Computer Science Today: Recent Trends and Developements*, number 1000 in LNCS, pages 275–291. Springer Verlag, 1995.

[36] H. Rosé, W. Ebeling, and T. Asselmeyer. The density of states - a measure of the difficulty of optimization problems. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the 4^{th} Conference on Parallel Problems Solving from Nature*, volume 1141 of *LNCS*, pages 208–217. Springer Verlag, 1996.

[37] J. D. Schaffer, R. A. Caruana, L. Eshelman, and R. Das. A study of control parameters affecting on-line performance of genetic algorithms for function optimization. In J. D. Schaffer, editor, *Proceedings of the 3^{rd} International Conference on Genetic Algorithms*, pages 51–60. Morgan Kaufmann, 1989.

[38] M. Schoenauer. Representations for evolutionary optimization and identification in structural mechanics. In J. Périaux and G. Winter, editors, *Genetic Algorithms in Engineering and Computer Sciences*, pages 443–464. John Wiley, 1995.

[39] M. Schoenauer. Shape representations and evolution schemes. In L. J. Fogel, P. J. Angeline, and T. Bäck, editors, *Proceedings of the 5^{th} Annual Conference on Evolutionary Programming*, pages 121,129. MIT Press, 1996.

[40] M. Schoenauer, F. Jouve, and L. Kallel. Identification of mechanical inclusions. In D. Dasgupta and Z. Michalewicz, editors, *Evolutionary Computation in Engeneering*, pages 477–494. Springer Verlag, 1997.

[41] M. Schoenauer and S. Xanthakis. Constrained GA optimization. In S. Forrest, editor, *Proceedings of the 5^{th} International Conference on Genetic Algorithms*, pages 573–580. Morgan Kaufmann, 1993.

[42] H.-P. Schwefel. *Numerical Optimization of Computer Models.* John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[43] P.D. Surry and N.J. Radcliffe. Formal algorithms + formal representations = search strategies. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Proceedings of the 4^{th} Conference on Parallel Problems Solving from Nature*, number 1141 in LNCS, pages 366–375. Springer Verlag, 1996.

[44] D.H. Wolpert and W.G. Macready. No free lunch theorems for search. Technical report, Santa Fe Institute, 1995.