*Complex Systems* 4 (1990) 369–383

# Epistasis Variance: Suitability of a Representation to Genetic Algorithms

**Yuval Davidor**
*Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel*

**Abstract.** The most problematic aspect in the application of a genetic algorithm (GA) is the coding of the problem. In superficial applications, choosing a representation may appear simple. Yet it is really an art because the theory provides only partial directives and is not always fully applicable. Different representations incorporate varying degrees of nonlinearity among the representation elements. This interwoven nonlinearity is directly coupled with the representation and considerably affects the efficiency of a GA search. Both too much and too little nonlinearity detract from the relative efficiency of a GA.

This paper suggests that measures to qualify the suitability of a representation to a GA search can be developed with the concept of *epistasis* (a biological term that states the amount of intrachromosome gene interaction). By viewing the representation as a whole, being more than the sum of its composing parts, the discussion on epistasis in GAs reveals several fundamental features of GAs and leads to a unique mechanism for "spying" on the suitability of a representation to a GA.

## 1. Background

The schema theory [6,12] implicitly lists prerequisite features that a representation should exhibit in order to utilize a GA search, namely that with an above average probability, short, low-order schemata will combine and form a higher-order co-adapted schemata. The schema theorem shows that above average schemata will proliferate, but it does not indicate whether this proliferation will occur at the optimum rate. In that respect, it is self-evident that the representation is the primary aspect of a GA application and determines its performance. The importance of the representation was recognized, attention was given to the issue of building blocks (their size and number), but the effect of interdependency among the representation elements did not receive sufficient attention [3,5,8]. Only certain degrees of nonlinearity enable a GA search to exhibit a relative efficiency, while others diminish this efficiency. Therefore, the amount of interdependency among the representation elements is an important ingredient in the GAs' cookbook and constitutes an essential source of information.

Gene interaction is a central issue in natural genetics, where genes not only are dependent on each other in order to jointly express phenotypical characteristics but also suppress and activate the expression of other genes [16]. The term that has become synonymous with almost any type of gene interaction is epistasis [14]. Derived from the Greek words *epis* and *stasis* ("stand" and "behind"), epistasis is therefore equated with *stoppage* or *masking*. Epistasis is used to describe the situation where one gene pair masks or modifies the expression of another gene pair. When the epistasis of a chromosome is said to be high, it means that many genes are strongly linked to other genes. It is helpful to remember that GAs, like many natural systems, assume a certain holistic structure, a structure where the whole is different from the sum of its parts [7,13,15,17,18]. The frequent characteristic of such information systems is that knowing the value of the parts does not necessarily enable the calculation of their effect together. In the GAs coterie, epistasis is used to indicate the extent of nonlinearity and interdependency among the elements composing the representation.

GA literature emphasizes that GAs do not "see" the problem domain directly because the latter is obstructed by the representation. Accordingly, the question "Which problem domains are amenable for a GA search?" should be replaced by the question "Does the representation (of the given problem) promote the most efficient GA search?" By shifting the question of suitability from the problem domain to the representation, one focuses on the core issue of GA applications, thus asking a question that is not only more consistent with the schema theorem but also easier to answer (and at least its meaning is more clear).

## 2. Notional epistasis in GAs

Tracing epistasis is an elusive occupation because the presence of epistatic elements can be traced only at the phenotypic level away from their scene of interaction (genotypic level). Furthermore, even if the amount of epistasis is known, the question remains "Can this be put to use?" Section 2.1 addresses the latter question and section 2.2 expands on string epistasis in contrast to fitness.

### 2.1 Epistasis on a scale

If a representation contains very little or no epistasis, any individual string element is affected by the value of the other elements, and therefore optimization becomes a bit-wise maximization. At the other end of the epistatic scale, when a representation is highly epistatic, too many elements are dependent on other elements and the building blocks become long and of high

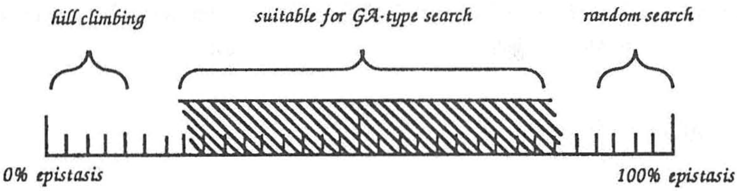

Figure 1: The region on the epistasis scale suitable for GAs, hill climbing, and random search.

order. When the epistasis is extremely high, the elements are so dependent on each other that unless a complete set of unique element values is found simultaneously, no substantial fitness improvements can be noticed (such as in the parity problem). Under such extreme circumstances, nonlinearity has exacerbated to the extent that the performance space does not contain significant regularities (or the mapping function from string to fitness becomes too global).

This leads to the conclusion that a representation should be constructed in a manner incorporating mild epistasis (neither too high nor too low). In figure 1, the three typical search strategies — hill climbing, GA, and random search — are plotted on a percentage epistasis scale according to their zones of relative efficiency: low, mild, and high epistasis respectively.

What effect does epistasis have on relative efficiency? A representation with low epistasis means that co-adaptation is not prominent, and therefore a hill-climbing algorithm is likely to be the most efficient. A representation with high epistasis implies that co-adaptation is too strong, above-average fitness schemata are of too high order, and therefore the efficiency of a GA will decrease significantly. A representation with mild epistasis is suitable for a GA (figure 1). If the epistasis can be calculated for a given representation, it is likely to offer an important yardstick of its suitability to a GA.

## 2.2 The linear assumption

Another aspect of GAs and their coding paradigm is that any fitness function can ultimately be reduced to a set of linearly independent partial fitness functions [8] so that for any string $j$ it is possible to write its fitness as the following sum:

$$v(S^j) = \sum_{i=1}^{2^l} f(S^j)\delta_{ij}, \qquad \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{2.1}$$

In other words, theoretically a fitness space can always be reduced into a table of fitness values for each of the phenotypes. This approach is adopted here, but in a different way. Instead of decomposing the fitness space according to strings as equation 2.1 implies, the fitness space is decomposed according to the coding elements (genes' value or alleles). Assuming such a decomposition is possible, the fitness of any string $j$ may be calculated by summing the values of its genes:

$$A(S^j) = \sum_{i=1}^{2l} A(S_i^j) \tag{2.2}$$

This means that instead of the $2^l$ fitness values required to compute any fitness according to equation 2.1, only $2l$ values are needed when considering equation 2.2. Furthermore, the discussion in this paper focuses on the relationship between the two values and its use as a suitability criterion for GAs efficiency.

The objective for applying the above linear decomposition is to develop a method for the prediction of the amount of nonlinearity (in terms of gene interaction) embedded in a given representation. To this end, fitness has to be associated with the representation elements. If a linear decomposition proves to be inaccurate, then it implies that the representation incorporates nonlinearities. Quantifying the amount of nonlinearity will provide an estimate for the suitability of a given representation to a GA processing. From a GA perspective, a coding format in which the effect of any individual parameter on the total fitness is independent of other parameters suggests that there is little co-adaptation.

On the other hand, a high degree of nonlinearity indicates that above-average schemata are too long. The whole GA ideology is based on the assumption that one can only say something about the whole by knowing its parts. What neither the schema theorem nor population genetics indicate is exactly how much of the whole the parts should indicate.

It is possible to detect nonlinearity by measuring the discrepancy between the real fitness and the recomposed fitness according to equation 2.2. The arguments for estimating the degree of nonlinearity of a coding function by estimating the applicability of the linear assumption are sound and are implicitly founded in the schema theorem. What is less clear is how these ideas can be formulated in a practical way. This issue is discussed in the next section.

## 3. The basic elements of epistasis

It was already emphasized above that the effect of epistasis lies in the ability to predict the value of a whole from the value of its parts. One possible method of calculating epistasis is based on the linear assumption and is loosely connected to Fisher's theorem (see Crow and Kimora [4] for a detailed discussion of Fisher's theorem). The following definitions are adopted for the preliminary analysis.

A string $S$ is composed from $l$ elements $s_i$ (without loss of generality, $l$ is fixed),

$$S = (s_1, s_2, \ldots, s_l) \tag{3.1}$$

Without loss of generality, only a binary alphabet is considered. The allele of the $i$th gene in a string is denoted by

$$s_i = a \qquad a \in \{0,1\}, \qquad i = 1, 2, \ldots, l \tag{3.2}$$

The *Grand Population*, $\Gamma$, is the set of all possible strings of length $l$,

$$\Gamma = \prod_{i=1}^{l} \{0,1\} \tag{3.3}$$

Let *Pop* denote a sample from $\Gamma$ where the sample is selected uniformly and with replacement. The size of a sample *Pop* is

$$N = |Pop| \tag{3.4}$$

The fitness of a string is given by

$$v(S) = fitness \tag{3.5}$$

where $v$ is a "blackbox" function. The average fitness value of the sample *Pop* is

$$\bar{V} = \frac{1}{N} \sum_{S \in Pop} v(S) \tag{3.6}$$

The excess fitness value of a string is denoted by

$$X(S) = v(s) - \bar{V} \tag{3.7}$$

The number of string instances in *Pop* that match $s_i = a$ is denoted by $N_i(a)$. The average allele value is denoted as

$$A_i(a) = \frac{1}{N_i(a)} \sum_{S \in Pop_{s_i=a}} v(S) \tag{3.8}$$

where $Pop_{s_i=a}$ is the set of all strings in *Pop* having the allele $a$ in their $i$th position. The weight of $s_i$ is

$$\Delta_i = |A_i(1) - A_i(0)| \tag{3.9}$$

The excess allele value is defined by

$$X_i(a) = A_i(a) - \bar{V} \tag{3.10}$$

and the excess genic value is

$$X(A_i) = \sum_{i=1}^{l} X_i(a) \tag{3.11}$$

and the genic value of a string $S$ — the predicted string value — is defined as

$$A(S) = X(A_i) + \bar{V} \tag{3.12}$$

Thus, the difference $\varepsilon(S) = v(S) - A(S)$ might reasonably be supposed to be a measure of epistasis of a string $S$.

Consequently, an epistasis measure for the Grand Population, and hence for the representation, is termed the *epistasis variance* and is defined as

$$\sigma_\varepsilon^2 = \frac{1}{N_\Gamma} \sum_{S\in\Gamma} [v(S) - A(S)]^2 \tag{3.13}$$

where the implicit $A_i(a)$ are computed over the Grand Population (note that this definition does not follow the common definition of variance as it involves elements from two different sets). This measure can be estimated from the corresponding expression

$$\sigma_{Pop}^2 = \frac{1}{N} \sum_{S\in Pop} [V(S) - A(S)]^2 \tag{3.14}$$

However, since the computation of $A_i(a)$ is determined by the sample population, this statistic is subject to sampling error (parasitic epistasis), but as yet, confidence measures for the estimate are unavailable. This would require an investigation of the distribution of

$$\sigma_\Gamma^2 - \sigma_{Pop}^2$$

The above definitions (summarized in table 1) provide a method for estimating the epistatic variance for a Grand Population — the base epistasis — from a sample population. The distinction between base epistasis and parasitic epistasis is very important because the effect of the latter is often of equal or higher order of magnitude. This will be demonstrated further in section 4.

The fitness variance is denoted as

$$\sigma_v^2 = \frac{1}{N} \sum_{S\in Pop} (X(S))^2 \tag{3.15}$$

and the genic variance is denoted as

$$\sigma_A^2 = \frac{1}{N} \sum_{S\in Pop} (X(A_i))^2 \tag{3.16}$$

The difference between the fitness variance and the genic variance is important (though not intuitive) for estimating to what extent the sample departures from the Grand Population, and it is denoted as

$$\sigma_{v-A}^2 = \sigma_v^2 - \sigma_A^2 \tag{3.17}$$

| Symbol | Term |
|---|---|
| $S$ | String |
| $v(S)$ | Fitness |
| $X(S)$ | Excess fitness value |
| $a$ | Allele |
| $A_i(a)$ | Allele value of $a$ |
| $X_i(a)$ | Excess allele value |
| $X(A_i)$ | Excess genic value |
| $A(S)$ | Genic value |
| $\varepsilon(S)$ | Epistasis value |
| $\sigma_v^2$ | Fitness variance |
| $\sigma_A^2$ | Genic variance |
| $\sigma_\varepsilon^2$ | Epistasis variance |

Table 1: Summary of the symbols and their definitions in the epistasis discussion.

## 4. Calculating epistasis: A few examples

In the following, the epistasis measuring tools developed in section 3 are applied to two fitness functions of known and characteristic epistasis (the strings and their corresponding fitness values are summarized in table 2). The fitness functions are the algebraic function summation

$$SUM = 2.33 \sum_{i=1}^{3} s_i, \qquad s_i = \{0, 1\}$$

and the logical function $AND$,

$$AND = 28 \sum_{i=1}^{3} s_i, \qquad s_i = \{0, 1\}$$

which represent zero and total epistasis problems respectively. A pseudo semi-epistasis function $SUM\&AND$ is achieved by combining the two functions,

$$SUM\&AND = 1.16 \sum_{i=1}^{3} s_i + 14 \sum_{i=1}^{3} s_i, \qquad s_i = \{0, 1\}$$

The first analysis uses Grand Populations and thus addresses the issue of base epistasis (section 4.1), after which the effect a sampled population has over the statistic is investigated (section 4.2). In section 4.3, a problem of an unknown epistasis — a fully deceptive problem — is analyzed. The functions are arranged to have an equal average fitness value to promote comparability between the epistasis variance in the absence of standard normalizing procedures.

| String | $SUM$ | $AND$ | $SUM\&AND$ | Deceptive |
|---|---|---|---|---|
| 000 | 0.00 | 0 | 0.00 | 7 |
| 001 | 2.33 | 0 | 1.16 | 5 |
| 010 | 2.33 | 0 | 1.16 | 5 |
| 011 | 4.66 | 0 | 2.33 | 0 |
| 100 | 2.33 | 0 | 1.16 | 3 |
| 101 | 4.66 | 0 | 2.33 | 0 |
| 110 | 4.66 | 0 | 2.33 | 0 |
| 111 | 7.00 | 28 | 17.50 | 8 |

Table 2: Strings and their fitness values of four fitness functions: $SUM$, $AND$, $SUM\&AND$, and a deceptive problem of zero, total, semi-, and bounded epistasis respectively.

| $S$ | $v(S)$ | $X(S)$ | $X(A_i)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0.00 | −3.50 | −3.50 | 0.00 | 0 |
| 001 | 2.33 | −1.16 | −1.16 | 2.33 | 0 |
| 010 | 2.33 | −1.16 | −1.16 | 2.33 | 0 |
| 011 | 4.67 | 1.16 | 1.16 | 4.67 | 0 |
| 100 | 2.33 | −1.16 | −1.16 | 2.33 | 0 |
| 101 | 4.67 | 1.16 | 1.16 | 4.67 | 0 |
| 110 | 4.67 | 1.16 | 1.16 | 4.67 | 0 |
| 111 | 7.00 | 3.50 | 3.50 | 7.00 | 0 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 2.33 | −1.16 | 2.33 |
| | 1 | 4.67 | 1.16 | |
| 2 | 0 | 2.33 | −1.16 | 2.33 |
| | 1 | 4.67 | 1.16 | |
| 3 | 0 | 2.33 | −1.16 | 2.33 |
| | 1 | 4.67 | 1.16 | |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 4.08 | 4.08 | 0 | 0 |

Table 3: Calculating the epistasis variance for the $SUM$ function.

### 4.1 Three epistatically different functions

The Grand Populations of three epistatically defined functions are analyzed: zero epistasis (table 3), total epistasis (table 4), and semi-epistasis (table 5).

When analyzing the epistasis variance of the $SUM$ and $AND$ functions, it is possible to observe the strength of the linear assumption. The $SUM$

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0 | −3.5 | −10.5 | −7 | 7 |
| 001 | 0 | −3.5 | −3.5 | 0 | 0 |
| 010 | 0 | −3.5 | −3.5 | 0 | 0 |
| 011 | 0 | −3.5 | 3.5 | 7 | −7 |
| 100 | 0 | −3.5 | −3.5 | 0 | 0 |
| 101 | 0 | −3.5 | 3.5 | 7 | −7 |
| 110 | 0 | −3.5 | 3.5 | 7 | −7 |
| 111 | 28 | 24.5 | 10.5 | 14 | 14 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 0 | −3.5 | 7 |
|  | 1 | 7 | 3.5 |  |
| 2 | 0 | 0 | −3.5 | 7 |
|  | 1 | 7 | 3.5 |  |
| 3 | 0 | 0 | −3.5 | 7 |
|  | 1 | 7 | 3.5 |  |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 85.75 | 36.75 | 49 | 49 |

Table 4: Calculating the epistasis variance for the $AND$ function.

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0.00 | −3.50 | −7.00 | −3.50 | 3.50 |
| 001 | 1.16 | −2.33 | −2.33 | 1.16 | 0.00 |
| 010 | 1.16 | −2.33 | −2.33 | 1.16 | 0.00 |
| 011 | 2.33 | −1.16 | 2.33 | 5.84 | −3.50 |
| 100 | 1.16 | −2.33 | −2.33 | 1.16 | 0.00 |
| 101 | 2.33 | −1.16 | 2.33 | 5.84 | −3.50 |
| 110 | 2.33 | −1.16 | 2.33 | 5.84 | −3.50 |
| 111 | 17.50 | 14.00 | 7.00 | 10.50 | 7.00 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 1.16 | −2.33 | 4.66 |
|  | 1 | 5.84 | 2.33 |  |
| 2 | 0 | 1.16 | −2.33 | 4.66 |
|  | 1 | 5.84 | 2.33 |  |
| 3 | 0 | 1.16 | −2.33 | 4.66 |
|  | 1 | 5.84 | 2.33 |  |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 28.6 | 16.3 | 12.3 | 12.3 |

Table 5: Calculating the epistasis variance for the $SUM\&AND$ function.

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0 | −3.5 | −3.5 | 0 | 0 |
| 001 | 1 | −2.5 | −2.5 | 1 | 0 |
| 010 | 2 | −1.5 | −1.5 | 2 | 0 |
| 011 | 3 | −0.5 | −0.5 | 3 | 0 |
| 100 | 4 | 0.5 | 0.5 | 4 | 0 |
| 101 | 5 | 1.5 | 1.5 | 5 | 0 |
| 110 | 6 | 2.5 | 2.5 | 6 | 0 |
| 111 | 7 | 3.5 | 3.5 | 7 | 0 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 1.5 | −2.0 | 4 |
|  | 1 | 5.5 | 2.0 |  |
| 2 | 0 | 2.5 | 2.0 | 2 |
|  | 1 | 4.5 | −1.0 |  |
| 3 | 0 | 3.0 | −0.5 | 1 |
|  | 1 | 4.0 | 0.5 |  |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 5.25 | 5.25 | 0 | 0 |

Table 6: A three-bit unsigned integer binary representation with zero epistasis.

function can be accurately recomposed from the decomposed $A_i(a)$ values, while the recomposition of the $AND$ function reveals a large epistatic variance. By analyzing the semi-epistatic function $SUM\&AND$, the notion of epistasis variance as a measure for intermediate nonlinearity in a representation is expended.

### 4.2 Samples and sampling noise

Since the population size in all practical GA applications is only a minuscule portion of the genotype pool, it is important to investigate whether calculating the epistasis variance from a sample involves a parasitic bias. This section investigates this sampling bias and suggests that the sampling bias has an overpowering effect on the measurement of base epistasis variance.

It was already shown that calculating epistasis variance with a Grand Population for a representation that contains zero epistasis yields a correct epistasis figure. This section will show that this conclusion is valid only for the Grand Population and erroneous when the calculation is not based on the Grand Population. In tables 7 and 8 such a calculation is shown, and it reveals a substantial parasitic epistasis variance.

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0 | −3 | −2.0 | 1.0 | −1.0 |
| 001 | 1 | −2 | −2.0 | 1.0 | 0.0 |
| 010 | 2 | −1 | −1.0 | 2.0 | 0.0 |
| 011* | | | | | |
| 100 | 4 | 1 | 1.5 | 4.5 | −0.5 |
| 101 | 5 | 2 | 1.5 | 4.5 | −0.5 |
| 110 | 6 | 3 | 3.0 | 6.0 | 0.0 |
| 111* | | | | | |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 1.0 | −2.0 | 4.0 |
| | 1 | 5.0 | 2.0 | |
| 2 | 0 | 2.5 | −0.5 | 1.5 |
| | 1 | 4.0 | 1.0 | |
| 3 | 0 | 3.0 | 0.0 | 0.0 |
| | 1 | 3.0 | 0.0 | |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 4.66 | 3.75 | 0.25 | −0.92 |

Table 7: A 75% Grand Population sample shows parasitic epistasis variance. The starred (*) strings are the ones not included in the statistics.

The analysis of sampled populations suggests the following:

1. The nonlinearity a GA operates with increases as the sample diverts from a Grand Population.

2. Epistasis, as used by biologists, consists of two elements: the base epistasis resulting from the representation and a parasitic epistasis resulting from sampling noise.

### 4.3 A fully deceptive problem

So far, the functions that were analyzed had a known epistasis. To conclude the preliminary discussion on epistasis variance, it is interesting to analyze a function of an unknown epistasis, known to be a "hard" problem for GAs.

The fully deceptive Hamming problem[1] is the archetype of a hard function, and it is that function to which the epistasis tools are applied. A deceptive function may include considerable structure. This structure quality is required by the very definition of the function (in spite of the contradiction

[1] The fully deceptive problem used here is based on the fully deceptive problem as defined by Goldberg [5,9] and involves variations of negligible importance that were adopted for convenience.

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 0 | −3.5 | −6.5 | −3 | 3 |
| 001 | 1 | −2.5 | −5.5 | −2 | 1 |
| 010* | | | | | |
| 011* | | | | | |
| 100* | | | | | |
| 101* | | | | | |
| 110 | 6 | 2.5 | 5.5 | 9 | −3 |
| 111 | 7 | 3.5 | 6.5 | 10 | −3 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 0.5 | −3.0 | 6 |
| | 1 | 6.5 | 3.0 | |
| 2 | 0 | 0.5 | −3.0 | 6 |
| | 1 | 6.5 | 3.0 | |
| 3 | 0 | 3.0 | −0.5 | 1 |
| | 1 | 4.0 | 0.5 | |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 9.25 | 36.25 | 9.0 | −27.00 |

Table 8: A 50% Grand Population sample shows further parasitic epistasis variance. The starred (*) strings are the ones not included in the statistics.

| $S$ | $v(S)$ | $X(S)$ | $X(A)$ | $A(S)$ | $\varepsilon(S)$ |
|---|---|---|---|---|---|
| 000 | 7 | 3.5 | 1.25 | 4.75 | 2.25 |
| 001 | 5 | 1.5 | 0.75 | 4.25 | 0.75 |
| 010 | 5 | 1.5 | 0.75 | 4.25 | 0.75 |
| 011 | 0 | −3.5 | 0.25 | 3.75 | −3.75 |
| 100 | 3 | −1.5 | −0.25 | 3.25 | −0.25 |
| 101 | 0 | −3.5 | −0.75 | 2.75 | −2.75 |
| 110 | 0 | −3.5 | −0.75 | 2.75 | −2.75 |
| 111 | 8 | 4.5 | −1.25 | 2.25 | 5.75 |

| $i$ | $a$ | $A_i(a)$ | $X_i(a)$ | $\Delta_i$ |
|---|---|---|---|---|
| 1 | 0 | 4.25 | 0.75 | −1.25 |
| | 1 | 2.75 | −0.75 | |
| 2 | 0 | 3.75 | −0.25 | −0.5 |
| | 1 | 3.25 | −0.25 | |
| 3 | 0 | 3.75 | 0.25 | −0.5 |
| | 1 | 3.25 | −0.25 | |

| $\sigma_v^2$ | $\sigma_A^2$ | $\sigma_\varepsilon^2$ | $\sigma_v^2 - \sigma_A^2$ |
|---|---|---|---|
| 9.25 | 0.68 | 8.57 | 8.57 |

Table 9: The epistasis analysis for a fully deceptive function.

in the name). For the deception to be effective, most partial fitness evaluations (low-order schemata) must agree with each other (thought not with the global optimum). The more deceptive a problem is, more partial evaluations should be in agreement with each other and with a disagreément with the global optimum. It is reasonable, therefore, to expect that the sum total epistasis embedded in a deceptive function will not be too high, and indeed it ought not to be too high. Calculating the epistasis variance for the fully deceptive problem (table 9) supports the above.

## 5. Conclusions and future work

Some basic problems regarding coding formats and GAs were discussed in this paper. It was suggested that much of the future success in GAs research is dependent on the development of analytical tools with which a given coding of a problem domain can be optimized. In spite of its importance, the precise analysis of representations is uncommon due to the tedious computation involved.

There is a novel approach to the analysis of coding-function relationships originally used by Bethke (and later by others [8,9]). This approach considers the use of Walsh functions in order to evaluate the fitness of schemata. However, this method involves a tedious computation (the computation of $2^l$ Walsh coefficients). As an alternative method for the solution of the coding enigma, and in an attempt to overcome some limitations of the Walsh function analysis (such as the requirement for the representation to be of fixed length), the epistasis variance analysis was suggested.

Epistasis variance is a more flexible method for the measurement of the degree of nonlinearity embedded in a coding format and hence its suitability to a GA. The quantification of nonlinearity is indirect and cannot differentiate between different orders of nonlinearity, but it is very simple and intuitive. The measurement of epistasis is based on the linear assumption that the coding parameters are linearly independent with respect to the fitness function. Assuming linear independence, the fitness is decomposed according to average allele values, with which the original fitnesses are recomposed. The accuracy of this method, or more precisely, the epistasis variance, is an estimate of the total amount of nonlinearity embedded in the coding. The method proposed, though its results are not conclusive, points out new underlying coding-function aspects.

Calculating the epistasis variance in grand populations for two illustrative functions (and for other functions the results of which were not presented here) revealed the following:

1. The epistasis as defined in this work detects all orders of nonlinearity.
2. The epistasis variance cannot be scaled or parameterized with the tools presented here, but can be compared qualitatively.

The epistasis variance is usually determined by a sample. Analyzing the epistasis variance for different samples reveals an extensive sampling error resulting from this approximation. Confidence measures for the extent of this approximation were not presented. As a result of the epistasis analysis for samples, the epistasis variance was redefined by means of two elements: the base epistasis and parasitic epistasis. It is important to distinguish between the two because only the base epistasis is relevant to the issue of suitability of a coding to a GA. It is also intuitively transparent that the parasitic epistasis can help in determining the optimal population size. Notwithstanding the difficulties in applying epistasis measurements, the epistasis variance does provide two important lessons that should be stressed:

1. It is possible to measure the extent of nonlinearity without knowing anything about the fitness function. The analysis holds for all representations where the elements of the representation can be identified.
2. The common use of epistasis in the context of nonlinearity is misleading because it combines two nonlinear properties: a base epistasis and a parasitic epistasis.

To be useful as a tool, the epistasis variance requires two extensions to the work presented in this paper:

1. Means for normalization of the epistasis variance so it can be plotted on a scale.
2. The development of confidence measures for estimating the base epistasis from sample populations.

Further work is also needed to establish the epistasis among groups of representation elements. Such information is desirable because epistasis may not be homogenously distributed.

## Acknowledgments

The work presented in this article was supported in part by the Center for Absorption in Science, The Ministry of Immigrant Absorption, The State of Israel. Judit Bar-Ilan and Antonia J. Jones made valuable comments on early drafts of this article. (Y. Davidor is recipient of a Sir Charles Clores Fellowship.)

## References

[1] F.J. Ayala and J.A. Kiger Jr., *Modern Genetics* (The Benjamin/Cummings Publ., 1980).

[2] C. Berek, G.M. Griffiths, and C. Milstein, "Molecular events during maturation of the immune response to oxazolone," *Nature*, **316** (1985) 412–418.

[3] A.D. Bethke, "Genetic algorithms as function optimizes," *Dissertation Abstracts International*, **41(9)** (1981) 3503B. Univ. Microfilm No. 8106101.

[4] J.F. Crow and M. Kimura, *An Introduction to Population Genetic Theory* (Harper and Row, New York, 1970).

[5] D.E. Goldberg, "Simple genetic algorithms and the minimal, deceptive problem," in *Genetic Algorithms and Simulated Annealing*, L. Davis, ed. (Pitman, London, 1987) 74–88.

[6] D.E. Goldberg, *Genetic Algorithm in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[7] D.E. Goldberg, "Zen and the art of genetic algorithms," 3rd Intl. Conf. on Genetic Algorithms and their Applications, 1988.

[8] D.E. Goldberg, "Genetic algorithms and Walsh functions: Part I, A gentle introduction," *Complex Systems*, **3** (1989) 129–152.

[9] D.E. Goldberg, "Genetic algorithms and Walsh functions: Part II, Deception and its analysis," *Complex Systems*, **3** (1989) 153–171.

[10] R.M. Gorczynski and E.J. Steele, "Simultaneous yet independent inheritance of somatically acquired tolerance to two distinct H-2 antigenic haplotype determinants in mice," *Nature*, **289** (1981) 678–681.

[11] J.J. Grefenslette, *Representation Dependencies in Genetic Algorithms*. Unpublished manuscript, 1979. Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, DC 20375-5000.

[12] J.H. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor, 1975).

[13] H. Jacobson, "Information reproduction and the origin of life," *American Scientist*, **43** (1955) 119–127.

[14] W.S. Klug and M.R. Cummings, *Concepts of Genetics*, 2nd edition (Scott, Foresman and Co., 1986).

[15] J.R. Platt, "Properties of large molecules that go beyond the properties of their chemical sub groups," *J. Theoretical Biol.*, **1** (1961) 342–358.

[16] M. Ptashne, "How gene activators work," *Scientific American*, January (1989) 25–31.

[17] H.A. Simon, "The architecture of complexity," *Proc. of the American Philosophical Soc.*, **106(6)** (1962).

[18] J.K. Tsotsos, "A 'complexity level' analysis of vision," *Proc. of the First Intl. Conf. on Computer Vision*, 1987.