A Deterministic Analysis of Stationary Diploid/Dominance

Buster Greene

6920 Roosevelt NE #126

Seattle, WA 98115

bgreene@accessone.com

Abstract

A deterministic approach is used to study the behavior of diploid vs. haploid efficiency in EP. The deterministic analysis produces an exact result without resorting to multiple trials, at the cost of assuming an infinite population size. A reduction is shown for the required growth time with diploid vs. haploid, using a common deceptive test function. Complete diploid dominance is implemented in a manner which can be applied to any scalar EP, GA or GP problem, stationary or otherwise. In so doing, a dual interpretation of inter- and intra-gene fitness evaluation becomes apparent, and has a natural extension to vector fitness criteria. Results appear to be consistent with previous non-deterministic, multiple trial tests that used stationary fitness criteria.

1. Introduction

Diploid/dominance experiments in EP and GAs have been done with one of two different schemes. The first is similar to that used by Holland (1975) and Hollstien (1971), which utilizes a dominance map and an ad-hoc dominance shift operator. Ng (1995), asserts that such a scheme is no more effective than haploidy, even in a non-stationary fitness environment, and furthermore lacks biological precedent. Ng, et. al. developed an alternative dominance operator that appears to offer some improvement.

A second approach to diploidy was described by Greene (1996) that follows a natural model of diploid fitness known as *complete dominance*. Partial and "complete" dominance are well established mechanisms in biology that provide a simple, but effective explanation for dominance (e.g., see Stansfield (1983)). As a bonus, this implementation of complete dominance requires no genotype modification, and has also shown promise with stationary (i.e., ordinary) fitness criteria. Specifically, implementation extends to real (EP) chromosomes or GP parse trees. For simplicity, the analysis used here will use EP for the experimental test bed.

Analyzing EP behavior typically requires the use of multiple runs to account for different initial (generation 0) conditions. By using an infinite population, and thereby *deterministic* model, multiple trials are avoided. As a result, exact solutions are obtained using a single EP run for a given set of parameters. In support of their usefulness, deterministic approaches were used by Bodmer (1967) and Eshel (1970) to study the effects of recombination on evolutionary efficiency.

Previous *non*-deterministic studies by Greene (1996) indicated that complete diploid dominance can provide a performance improvement in stationary GA's. That approach considers any *scalar* fitness GA to have a single *dominance* locus, and is identical to the approach used here. The term "dominance locus" refers to a distinct chromosome subset, or gene, that functionally maps to a scalar that is the gene's fitness. If the objective criterion maps the entire chromosome to a single scalar result, then the dominance locus also includes the entire chromosome. In this case, *complete* diploid dominance calls for each diploid haplotype to be evaluated using the problem defined fitness criterion The maximum of the two resulting scalars is then the diploid individual's fitness. Greene also discussed a classical deterministic analysis of mutant allele retention, with this scheme, that uses a 1-locus, 2-allele model. 2-alleles, however, are not enough to model a deceptive fitness function. Those results are extended in this paper using a multiple-allele, single dominance-locus model that permits creation of a deceptive fitness function.

A single locus, diploid chromosome was used by Greene (1997) to address a medical signal processing application. Each real encoded, multiple string element haplotype was expanded in a GP-like (depth first) manner. This approach utilized GP function and terminal sets, and one ADF. Multiple trials indicated that diploid efficiency was greater than haploid. This demonstrated the use of single-gene, complete diploid dominance in a somewhat practical scenario.

The proposed multiple string element per gene scheme is unquestionably a complication of the usual EP/GA model. However, accurately modeling the biology (in order to reverse engineer the evolutionary process) may require this complication. Fundamentally, a biological gene doesn't correspond to a single DNA base pair, but to a sequence of 1000 or more base pairs. Such a gene encodes a protein that may range in importance, to the organism, from none to crucial. For example, mutant hemoglobin related genes coming from both parents, can spell very real survival problems. Conversely, having a working gene from just one parent may be all that is required for normal functioning. The latter situation is precisely modeled by complete diploid dominance. The model is admittedly simplified here for scalar fitness, EP/GA usage, in that each individual genome is comprised of a single gene. The single-gene case, however, requires no vector fitness evaluation and still permits an analysis of diploid efficiency.

2. Methods

2.1. Overview

"Complete" dominance, as mentioned, uses the 2 objective values defined by evaluating each diploid homologue. As mentioned, the maximum of these two scalars defines the diploid individual's fitness. In general, each haploid objective value is a real scalar. As such, analysis of the single gene problem calls for a multiple-*allele* model in order to introduce deceptivity[1]. The deterministic model used here, which assumes an infinite population, results in exact difference equations for each individual allele proportion. This non-deterministic model also implements mutation, selection and either haploid or diploid populations for comparison purposes. A real-string mutation operator is implemented in a manner similar to that used by Michalewicz (1992, pg. 88). Selection is non-tournament, fitness proportionate, with non-overlapping generations and replaces 100% of the population each generation.

No recombination will be used here since one desire is to isolate the effects of diploidy. In addition, the complete dominance population is complicated with two kinds of recombination. The first is the usual one, between individual, real string positions. In this case, crossover occurs *within* the gene or dominance locus. The second applies to chromosomes having multiple genes[2]. Such a situation encourages recombination to occur *between* dominance loci. Also, the effects of recombination have been shown elsewhere to not be obvious. For example, Felsenstein (1965) has shown that recombination can actually slow evolution, depending on the 2nd derivative of the fitness landscape. The use of multiple-locus models, to permit both inter and intra gene recombination, might alter or even improve diploid vs. haploid efficiency. Such complexity, however, is beyond the scope of this paper.

2.2. Fitness Function

Mutation and selection are defined for a finite number of M alleles at a single locus. For the experiments here, M=16. The fitness function used is:

$$f(i) = \begin{cases} 1+\alpha; & i = 0,\ \alpha > 0 \\ (i-1)/(M-2); & i \in [1..M-1] \end{cases} \qquad (1)$$

and is shown in figure 1. A value of $\alpha = 0.01$ was found to give adequate difficulty with short run times, and was used for the experiments described in section 3. Equation 1 is intended to simulate a difficult portion of a hypothetical fitness landscape. It is similar to the unitary deceptive trap function used by Deb (1993, pg. 95) and others, except that the operator space is real instead of binary. In order to require that the mutation operator climb against the entirety of the linear trap, the initial population consists entirely of allele type i = 15. The rate of growth of 0's can then be controlled by, among other things, the size of α (larger α makes the problem easier). Equation 1 was chosen because of its simplicity and previous use. In short, this deception problem is a simple test where EPs should show an advantage over hill-climbing.

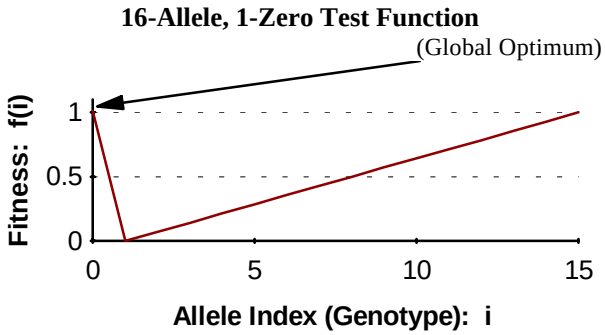

Figure 1. Single zero, deceptive fitness function. There are 16 allele indices or values. The global optimum is at allele 0, and has a value of f(0)= 1.01.

As mentioned, deterministic difference equations can be written and iterated to investigate certain aspects of evolutionary behavior. A fundamental criterion to be used here, in comparing haploid vs. diploid efficiency, is the rate

[1] In principal, a 2-locus, 2-allele model *could* be used to implement deceptivity and crossover, but unnecessarily complicates the diploid vs. haploid issue.

[2] This would apply if the objective function is, itself, a vector.

of increase for the globally optimal genotype[3] when it is initially (at generation 0) non-existent. The initial population consists *entirely* of the allele that is maximally distant from the global optimum in *operator* space. Mutation is the "operator" in question, and is designed so that small allele changes are encouraged, as described below.

2.3. Mutation

Mutation is designed to act like a random walk from the current allele value, according to:

$$x_i^* = (1-p_m)\,x_i + p_m \frac{\sum_{k=0,\,k\neq i}^{M-1} x_k (1-|i-k|/M)^R}{\sum_{k=0,\,k\neq i}^{M-1} (1-|i-k|/M)^R} \qquad (2),$$

where x_i^* is the proportion of the i^{th} allele value after mutation (but prior to selection). x_i is the corresponding proportion before mutation, p_m is the mutation rate and M is the number of alleles. R controls the mutation *dispersal*, or extent of likely change from the current value. This parameter localizes the effect of equation (2) in the space defined by the range of ordered allele values. Larger R makes the mutation operator more localized and generally results in slower, but more detailed hill climbing. Equation 2 is similar to the mutation operator utilized by Michalewicz (1992, pg. 88) in that a double exponential is used to encourage mutation to allele values that are close to the current generation's value. Here, for simplicity, R is held constant throughout a given run.

2.4. Selection

The proportion of the i^{th} allele after selection is given by:

$$x_i' = \overline{w}_i / \overline{w}, \quad i = 1..M \qquad (3)$$

Where: For diploid,

$$\overline{w}_i = \sum_{j=1}^{M} x_j^* w_{ij}$$

$$\overline{w} = \sum_{j=1}^{M}\sum_{k=1}^{M} x_j^* x_k^* w_{jk}$$

and for haploid,

$$\overline{w}_i = x_i^* w_i$$

$$\overline{w} = \sum_{j=1}^{M} x_j^* w_j$$

also:

x_i^* = Proportion of ith allele *after* mutation

w_i = Fitness of the i^{th} *haploid* genotype

w_{ij} = Fitness of the ij^{th} *diploid* genotype

M = Number of alleles.

Equation 3 describes the selection process for haploid and diploid populations, and occurs immediately after mutation. As may be seen, all genotype combinations are enumerated. Also, every possible haploid or diploid genotype, has a fitness that is defined by equation 1 and the use of complete dominance. x_i^* then gives the exact proportion (excluding roundoff error) of the i^{th} allele type in the next generation.

3. Results and Discussion

The deterministic model was iterated for both haploid and diploid populations, using the fitness function of equation 1. The proportion of type 0 alleles (corresponding to the global optimum) were recorded every 2 generations. Runs were then made for various mutation rates, p_m, and mutation dispersals, R. In order to facilitate comparison with normal, finite population EP/GA implementation, the proportion of optimum alleles was used as the criterion for comparing haploid vs. diploid GA efficiency. It should be noted that, in the absence of an elitist selection strategy, the mere appearance of the optimum does not guarantee that it will stay in a finite population, even to the next generation. Such a criterion is nonetheless useful for comparison purposes.

The first set of experiments (section 3.1) show the proportion of optimal alleles vs. generation for various mutation rates and with a fixed value of R=6. This value of R, and range of p_m are similar to that used by Michalewicz (1992) and Greene (1997). In section 3.2, the effect of varying R shows an effect on diploid/haploid efficiency due to variation of the mutation dispersal.

[3]Since the model used here interprets the chromosome as single gene, the terms allele and genotype are used interchangeably.

3.1. Optimal Allele Growth vs. Mutation Rate

The growth of the optimum allele starts at 0 and asymptotes to an equilibrium level in all cases. As these results show, the rates of growth for haploid and diploid can differ.

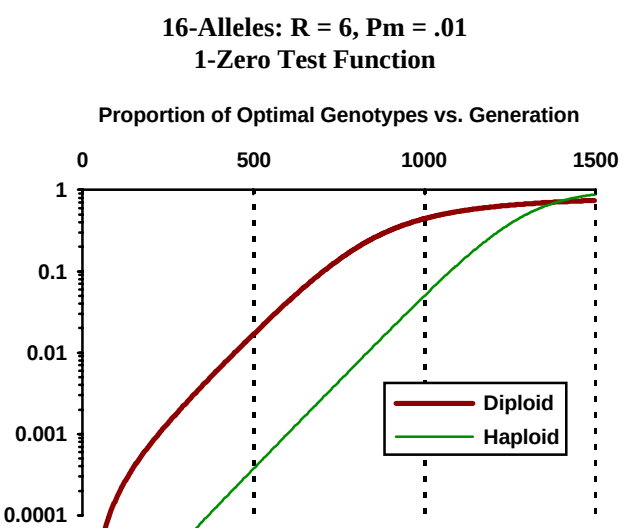

Figure 2. Proportion of optimal alleles vs. generation at p_m= 0.01.

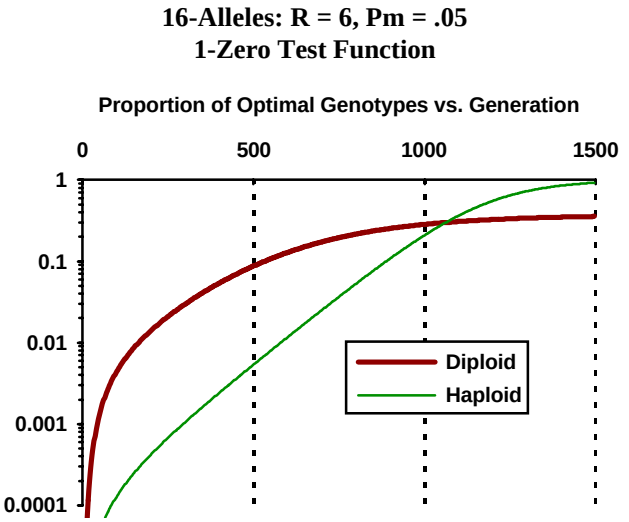

Figure 3. Proportion of optimal alleles vs. generation at p_m= 0.05. As in figure 2, the proportion of diploid optimals increases faster in early generations than with haploid. The final diploid "equilibrium" is lower than haploid because low-fitness mutants are being preserved at a higher level in the population.

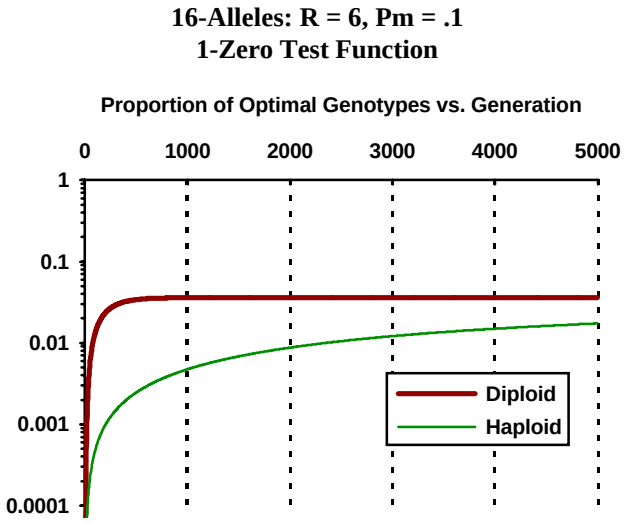

Figure 4. Proportion of optimal alleles vs. generation at p_m= 0.1. Diploid optimals increase more rapidly in early generations. At sufficiently large mutation rates, haploid never catches up to diploidy.

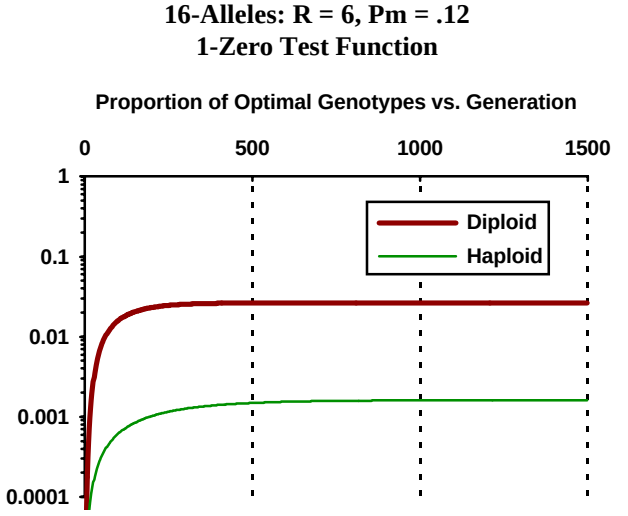

Figure 5. Proportion of optimal alleles vs. generation at p_m= 0.12. Haploid optimals asymptote to a proportion lower than diploid, and between 0.001 and 0.005 (see figure 6).

Figures 2, 3, 4 and 5 plot runs for mutation rates of 0.01, 0.05, 0.1 and 0.12 respectively. In all four cases, the increase in the optimal allele type occurs at a lower generation count with diploid than haploid. The proportion of optimal alleles vs. generation may be used to estimate how soon we may expect to see an optimal genotype appear in a *finite* population (see figure 6 below).

As the mutation rate increases, the final equilibrium values for haploid can be seen to be higher than with diploid, except in figures 4 and 5 where optimal proportions remain

below about 0.1. In these high p_m cases, diploidy maintains higher optimal allele proportions throughout the run.

3.2. Generations to Optimum

Figure 6 plots ratios of diploid to haploid generations that are needed to reach optimal allele proportions of 0.001 and 0.005. These 2 values were chosen in order to represent 1 individual in finite populations of size 1000 and 200 respectively.

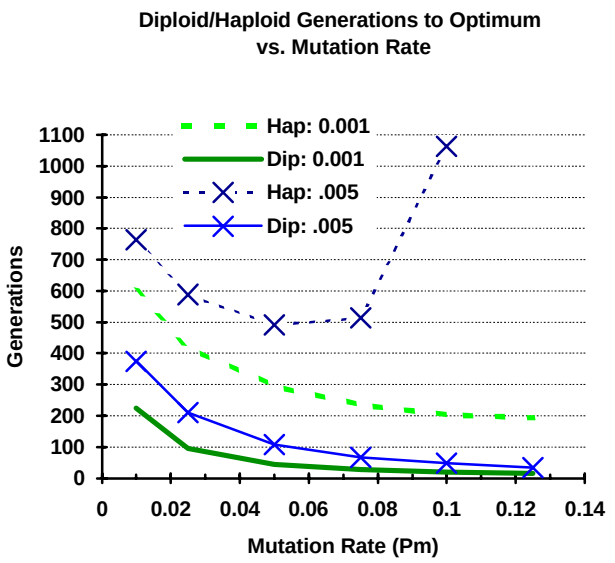

Figure 6. Comparison of haploid and diploid efficiencies for various mutation rates. "Generations" indicate the number of (parallel) diploid or haploid fitness evaluations required to achieve proportions of 0.001 or 0.005 in the optimum allele. Increasing the mutation rate generally increases efficiency. Haploid fails to reach an optimal proportion (p) of .005 for $p_m \geq 0.12$.

The increase in required haploid generations to reach the .005 level in figure 6 ($p_m \geq 0.12$), suggests something akin to premature convergence. As suggested by the graph, and figure 5, the haploid proportion actually asymptotes to approximately 0.0016 for increasing mutation rate (and thus never reaches 0.005). At such an equilibrium point (for haploid), mutation is destroying optimal alleles as fast as they are being created by selection. In a corresponding *finite* EP population of less than 200, the optimum allele would ordinarily remain absent from the population. In the case of diploidy, however, complete dominance maintains a greater supply of low-fitness alleles.

3.3. Summary of Results

Figure 7 shows that the 1-zero problem most quickly reaches solution with a wide mutation dispersal (small R), for both haploid and diploid. As the mutation dispersal becomes completely flat (R=0) the diploid advantage completely disappears. This is as expected, since such a condition causes all allele values to get created (by mutation) at equal rates. In an actual EP problem, consisting of perhaps many millions of real-valued alleles, simply increasing the mutation dispersion without limit is not an option, since it compromises fine detailed hill-climbing.

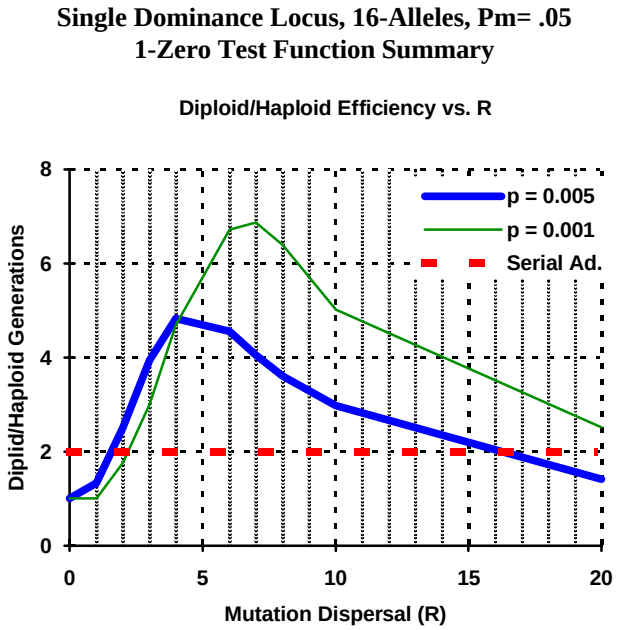

Figure 7. Relative diploid vs. haploid efficiency, measured as the ratio of generations required to reach a proportion p of optimal genotypes. Two values of p are plotted against mutation dispersion, R. Diploid must have an efficiency > 2 on a serial machine to outperform haploid.

In order to determine how sensitive the above results are to fitness behavior, a series of runs were made with various α values for equation 1. Moderate values for R and p_m were used, with results shown in figure 8. As before, the number of generations required to reach optimal allele proportions of 0.001 and 0.005 are compared for diploid vs. haploid. Reducing α increases deceptivity, and makes the problem more difficult. This also appears to make the haploid problem difficulty increase much faster than is seen to occur for diploidy. Sufficiently small α precludes the haploid proportion from ever reaching 0.001.

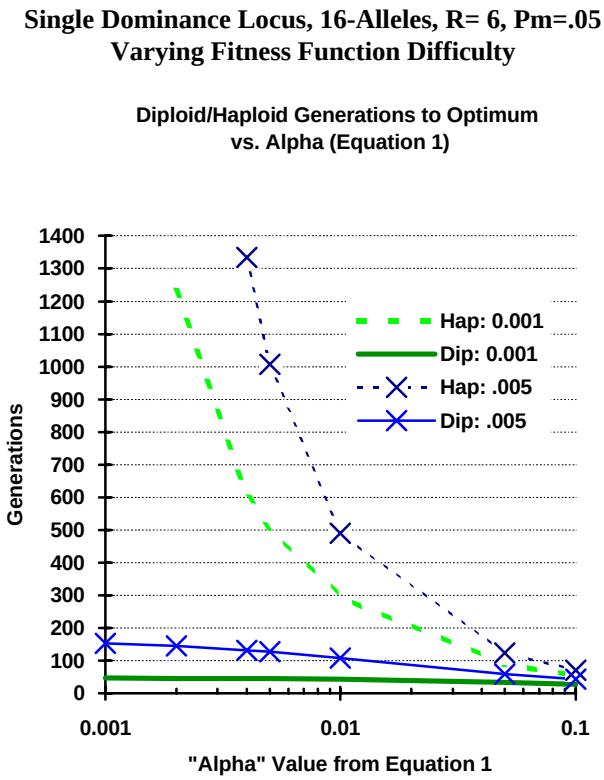

Figure 8. Comparison of haploid and diploid efficiencies. "Generations" indicate the number of (parallel) diploid or haploid fitness evaluations required to achieve proportions of 0.001 or 0.005 in the optimum allele. Decreasing α (equation 1) increases GA difficulty and increases the diploid efficiency advantage over haploid. Haploid fails to reach an optimal allele proportion (p) of 0.001 for α ≤ 0.001.

4. Conclusions

Results with the simple 1-zero, deceptive problem suggest a potential difference in efficiency between diploid and haploid evolutionary systems. Findings are exact due to the use of a deterministic approach. The criterion used for comparison is the rate of increase in the optimal allele, when initially non-existent.

In general, diploidy appears to offer some advantage. For a given α value, the amount of improvement is affected both by the rate and by the *extent* of mutation in operator space. As the mutation rate increases, for a moderate mutation dispersal (R), both the haploid and diploid models increase in efficiency. As expected, a sufficiently large p_m causes efficiency to rapidly decrease. The results of figure 6 show that this value of p_m is reached first with haploidy. Still larger p_m's will cause a similar effect with diploidy (not shown), but this occurs well after diploid and haploid efficiencies have leveled off. This decrease in efficiency with increasing p_m occurs because optimal alleles are being destroyed by mutation faster than they are being created by selection. With the deterministic model, this is seen to be an asymptotic "equilibrium" point in the proportion of optimals for increasing p_m. In this latter case, the level is less than that which represents a single individual in a typical finite population.

Figure 8 indicates that as the problem difficulty *decreases*, the haploid and diploid efficiencies become more nearly identical. From this we may conclude that the diploid speedup seen here will not hold true for all fitness criteria. Conversely, we can imagine scenarios where, with adversely chosen mutation or population size, neither haploid or diploid implementations can reliably locate the global optimum. Such a situation might not show a very clear advantage for either haploid or diploid (partly because the GA/EP is not working well in either case). An example of this is given in the haploid "failure" of figure 6, and the identical fate of diploidy, with extremely high p_m (not shown). It should also be concluded (from figure 8) that a serial implementation of diploidy could result in performance consistently *worse* than haploidy if the problem (deceptivity) is too easy. In such cases the use of EP might be questionable anyway.

A comparison of diploid vs. haploid efficiency at p_m= 0.05 (figure 7) indicates that an optimum mutation *dispersal*, at least for the 1-zero problem, exists. For α=0.01, this optimum is in the range of R=5 to 8. This finding may be applicable to EP strategies that use a localized mutation operator similar to that described by equation 2.

In order to see a diploid speedup, the evidence also suggests that the mutation operator must be sufficiently localized (R ≈ 1 or greater). This observed behavior of the diploid speedup is consistent, in the sense that it occurs over a reasonably wide range of both mutation rates and dispersals. The range of mutation rates and dispersals for which diploid exceeds haploid efficiency, are similar to values typically used with EP. The diploid speedup is also in agreement with previous, *non*-deterministic EP and GA findings described by Greene (1997).

The model used here interprets the genome as a single locus having 16-alleles, to compare haploid vs. diploid EP efficiency. The usual scalar objective function is combined with a diploid mapping function to implement complete dominance. Specifically, the fitness of each diploid individual is taken to be the maximum of its 2 gametic fitnesses. The resulting scalar objective value is thereby associated with a single gene "dominance locus". This dominance locus usually consists of a large number of string elements. Further consideration of this multiple-element gene model suggests that crossover, were it permitted, would occur not only between genes (were there

more than one) but could also occur within them. Such genetic behavior is known to occur in nature.

The usual interpretation of genes suggests that they are the simplest possible elements of recombination. The multiple loci-per-gene model discussed here, however, makes use of a dual (more general) interpretation of genes: As in nature, genes are not exclusively defined by their role in recombination, but in addition have a functional, phenotypic identity (i.e., one protein sub-unit per gene). Each such genotype inherently possess a quantifiable viability, that reflects its phenotype's ability to correctly function in a (potentially) crucial role for the organism. Here, this interpretation has been abstracted to, in one sense, the simplest possible scenario: That of an organism having only one gene. Of course. a single-gene organism is not found in nature. However, the mapping of genes to protein sub-units, at least some of which must function correctly for the organism to be healthy and reproduce, is a very well established occurrence in biology.

The single gene model studied here logically extends to multiple gene viability scalars, which taken together define a fitness *vector*. One such vector is thereby associated with each diploid, multiple gene homologue. In a GA, EP or GP application, the elements of the fitness vector would then map one-to-one with individual genes, whose scalar fitnesses are each problem defined. Multiple "problem defined" genes have in fact been previously used by Langdon (1995) and also Greene (1997). In these examples, the genes corresponded to specific ADF-like functions, each having a predefined fitness behavior, and each mapping to non-overlapping regions of the genome.

The single-gene, "complete" diploid dominance model applies to most if not all EP/GA/GP-like approaches that have a scalar fitness criterion. When modeled as such, diploidy appears to usually outperform haploidy. If put to use in vector fitness problems, as just discussed, diploidy might conceivably provide further improvements by protecting entire, pre-defined genes from premature extinction. In such contexts, it further appears that the use of a deterministic model can provide unique insights into idealized EP behavior, through a more exact and rapid comparison of operator and fitness variations.

Acknowledgments

My thanks to Dr. Joe Felsenstein, who suggested looking at deterministic approaches.

Bibliography

Bodmer, W. F., and Felsenstein, J. (1967). Linkage and selection: Theoretical analysis of the deterministic two locus random mating model. Genetics, 57, 237-265.

Deb, K., Goldberg, D. (1993). Analyzing Deception in Trap Functions. In D. Whitley (Eds.), FOGA-2 (pp. 93-108). San Mateo: Morgan Kaufmann.

Eshel, I. a. F. M. (1970). On the evolutionary effects of recombination. Theoretical Population Biology, 1, 88-100.

Felsenstein, J. (1965). The effect of linkage on directional selection. Genetics, 52(2), 349-363.

Greene, F. (1996). A new approach to diploid/dominance and its effects on stationary genetic search. In D. Fogel (Ed.), Proceedings of the Fifth Annual Conference on Evolutionary Programming, . San Diego: MIT Press.

Greene, F. (1997). Performance of diploid dominance with genetically synthesized networks. Proceedings of the Seventh International Conference on Genetic Algorithms, 615-622.

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

Hollstien, R. B. (1971) Artificial genetic adaptation in computer control systems. Doctoral Dissertation, University of Michigan.

Langdon, W. (1995). Evolving Data Structures with Genetic Programming. In Proceedings of the Sixth International Conference on Genetic Algorithms, (pp. 295-302). Pittsburg: Morgan Kaufmann.

Michalewicz, Z. (1992). Genetic Algorithms + Data Structures = Evolution Programs. Berlin: Springer-Verlag.

Ng, K., Wong, K. (1995). A new diploid scheme and dominance change mechanism for non-stationary function optimization. In L. Eshelman (Ed.), Sixth International Conference on Genetic Algorithms, (pp. 159-166). Pittsburg: Morgan Kaufmann.

Stansfield, W. D. (1983). Schaum's Outline of Theory and Problems in Genetics (2 ed.). McGraw-Hill.