Genetic Repair for Optimization under Constraints Inspired by *Arabidopsis Thaliana*

Amy FitzGerald and Diarmuid P. O'Donoghue

Department of Computer Science, NUI Maynooth, Co. Kildare, Ireland
amyfg@cs.nuim.ie, diarmuid.odonoghue@nuim.ie

Abstract. It has recently been proposed that the model plant, *Arabidopsis thaliana* (thale cress), uses a newly discovered genetic repair system to repair errors at the genetic level. *A. thaliana* uses information from the grandparent's genes as a basis for this correction – so genetic information appears to skip a generation. We apply this gene repair strategy to a combinatory optimization problem, firstly comparing the performance of parent and grandparent based repair. Subsequent experiments expand our understanding of the GeneRepair algorithm, by examining the parameters of fitness and direction involved in the generepair process. Our results point to a tentative explanation as to why *A. thaliana* might have evolved such an apparently complex inheritance process.

Keywords: Evolutionary optimization, genetic repair, constraints, Arabidopsis thaliana.

1 Introduction

Evolutionary Optimization (EO) is an optimization strategy that is inspired by Darwin's idea of survival of the fittest. EO effectively implements a "generate and test" beam search to find near optimal solutions to complex problems, such as NP-Complete problems, in a reasonable amount of computing time. A population of candidate solutions are created and allowed to converge towards a global optimal under the guidance of a suitable fitness function. Evolutionary strategies are effective in exploring complex solution spaces, where each individual explores part of the search space. However, EO is less suited to enforcing validity constraints [1] on these search spaces.

Evolutionary optimization and related approaches, use biology as their inspiration. This paper turns again to the biology domain, looking at some recent advances in the study of the *Arabidopsis thaliana* (thale cress) plant. *A. thaliana* appears uses a genetic repair process to repair errors in its genes. This repair process uses genetic information originating in the genes of the grandparent – information which does not appear to be detectable in the genes of the parent.

This paper presents a comparison of these repair strategies, on a standard combinatoric optimization problem. Results for a biologically inspired penalty points technique [2] act as a benchmark. The results of our initial experiment and presented and

G. Rudolph et al. (Eds.): PPSN X, LNCS 5199, pp. 399–408, 2008.

discussed, followed by two supplementary experiments that clarify some issues raised by the first experiment.

To the best of our knowledge, no authors have previously examined the effectiveness of grand-parent based repair in evolutionary computation, or compared parent based and grandparent based approaches to genetic repair.

2 Gene Repair in Arabidopsis Thaliana

Arabidopsis thaliana (thale cress) is a model plant used for a wide variety of detailed studies and was the first plant genome to be sequenced. The Arabidopsis plant has one of the smallest genomes with about 157 million base pairs and five chromosomes. The Arabidopsis genome encodes 27,000 genes and 35,000 proteins.

Lolle *et al* [3] investigated *A. thaliana* plants with an organ fusion mutation on the Hothead gene (HTH), resulting in an abnormal formation of the plant's flower. Their studies revealed that two plants with the HTH can produce offspring without this abnormality, forming perfectly normal plants. The resultant offspring have the normal form of the hothead gene (hth), even though this information was present in neither of the parent's genomes. That is, approximately 10% of the offspring were found to revert to the normal form of the hothead gene, which is a far higher rate than can be explained by random mutation of these specific alleles (which would be of the order of 1 per billions per allele per generation). It was found that that these revertant genomes all appeared to inherit genetic information from their grand-parents genomes, which had the normal (hth) form. Thus, genetic information appeared to skip a generation, reappearing in a subsequent generation. In an interview with the Washington Post (March 23rd, 2005) Robert Pruitt referred to this as a "parallel path of inheritance", which appears to occur in addition to standard Mendelian inheritance. In essence, a corrective template is used to correct broken or damaged sequences of DNA, possibly in response to stress placed on the plant due to the presence of a genetic mutation.

While Lolle's controversial [4] explanation relies on a cache of RNA inherited from previous generations, we focus on the explanation offered by Ray [5] that is compatible with Lolle's findings. Ray's explanation relies on an archival form of DNA, that serves to store the ancestral DNA but which is not detected by the processes used to sequence the regular encoding of DNA.

Thus, in our implementation each individual maintains its own archive of 2 generations of ancestral genetic information. This yields a custom made repair template for each individual in the population (see Figure 2).

3 The TSP Evolutionary Optimization Problem

To examine the performance of various GeneRepair strategies, we used the standard problem called the Traveling Salesman's Problem (TSP) (or the Hamiltonian Circuit problem). This NP complete problem involves finding the shortest path that visits each of a number of vertices (cities), visiting each just once and returning back to the original vertex (city). The TSP problem is thus a minimization problem, where the

best results correspond to a lower tour length. In this paper we use the results generated for the 51 city traveling salesman problem (eil51) from the standard TSPLib problem set. We point out that our focus was on comparing the effectiveness of GeneRepair strategies and not on producing short tours for this problem set *per se*.

One specific requirement for the problem domain was that it has identifiable validity constraints. That is, invalid solutions to this problem can be generated and can be identified. A TSP solution is invalid if the tour does not visit all cities, if a city is visited twice or if the tour does not return to the starting city.

The mutations of *A. thaliana* studied by Lolle *et al* [3] were from living plants, which were thus viable plants. There are a relatively small number of known viable (living) mutations of *A. thaliana*, corresponding to a tiny fraction of combinations of its 157 million base pairs. In contrast, the TSP does not have such viable mutants as all mutants form invalid (non-viable) solutions to the problem. These non-viable solutions are repaired immediately, whereas *A. thaliana* does not appear to involve genetic repair until the next generation. While our experiments appear to involve a slightly more pro-active gene-repair process, it was considered that is was not a very significant difference. These differences may perhaps lie more in the environmental stress factors that trigger gene repair in *A. thaliana.*

All experiments were run with the same experimental set-up, where only the described parameters were changed between experimental conditions. Initial experiments were conducted with a population size of 500 for 500,000 generations. This yields an overall search space that examines 250,000,000 different possible tours. We point of that this is a tiny fraction of the total search space of approximately $1.5.*10^{64}$ possible tours. Several independent runs were conducted for each experimental condition, to counteract against the randomized nature of EO. (A computer cluster was used to support simple independent simulations).The best results produced at each stage were recorded, as well as the generation at which those results were generated. The best and average results are presented in the next section.

4 Evolutionary Optimization with GeneRepair

This paper applies the genetic repair process described by Lolle [3] and Ray [5] to an otherwise standard EO algorithm (with unmodified crossover and mutation operators). The GeneRepair process is largely independent of the application domain itself. The only influence the problem domain has is through the genetic strings of the ancestor population. Thus, we conclude that this repair process is (largely) domain independent and may work as well or even better on a variety of other problem domains. This may be related in some way to the findings of Lolle *et al* [3] who found that gene repair in *Arabidopsis thaliana* appeared to operate throughout the DNA sequence and thus appear to be a general mechanism for extra-genomic inheritance.

However, before examining the GeneRepair process itself, we must first look at the underlying EO algorithm.

4.1 Representation

Each allele in our EO algorithm encodes a single city and each city is uniquely encoded. Therefore there is a 1-to-1 association between cities of the TSP problem and

the city's representation within the EO algorithm. Solutions to the TSP are formed as an ordered list of cities and the entire population is composed of a fixed number of individual tours (see Figure 1). Tours are stored as a fixed length and ordered list of cities (the number of cities in TSP determining the length of representation). So, the relative order of cities determines their position within a tour.

This representation allows two types of genomic error to occur to individuals within a population. Firstly, duplicate errors may occur when a city is repeated within a candidate tour in the population. Secondly, omission errors occur when a city is absent from a candidate solution in the population. We highlight that error is a violation of the solution constraints as required by the TSP. Because of our fixed-length encoding, omission and duplicate errors are always found in pairs. Thus, an omission error is always has a corresponding duplicate error. As can be seen in Figure 1, duplication of the "2" causes omission of "6" from the genetic sequence. Repairing such errors shall be discussed in Section 4.2 below.

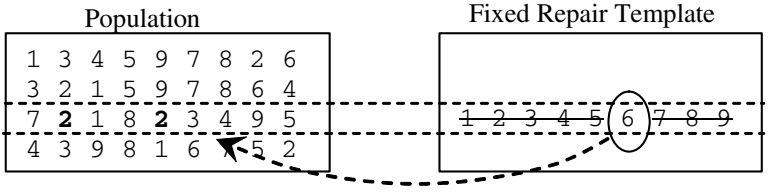

Fig. 1. The third Individual has duplicate and missing information, which must be repaired

4.2 Fitness Function, Crossover, Mutation

Before we present the GeneRepair operator, we first clarify the structure of the EO that is working in conjunction with GeneRepair. We now briefly describe the operators of fitness evaluation, selection, crossover and mutation rates. We point out that these are all generic operators, none of which are tailored to the given problem domain (see [6] for a discussion of specialized operators). The fitness function operates on individual tours, calculating the Euclidean distance between each city pair in turn, returning the sum of the individual inter-city distances.

Previous work on the GeneRepair operator has investigated the performance of parent based repair [7, 8]. This was compared to the performance of a variety of alternative strategies for implementing constraints on EO. In particular, this work explored how GeneRepair interacted with the standard evolutionary parameters, especially mutation rate and crossover mechanism.

While the results of Mitchell [7] indicate that best results are produced using Tournament selection, the results in this paper used Truncation selection with a truncation factor of 2. Thus, at the end of every generation the fittest half of the population we replicated and replaced the less-fit half of the population. The decision to use Truncation selection was made because of its simplicity and because it made detailed analysis of results (not discussed here) easier to conduct. Similarly, single point crossover was used to create new individuals. Thus, a random point on the genetic sequence of both parents is chosen, the first portion of the first parent and the last portion of the second parent are combined to form the new individual (solution).

Mitchell [7] indicates that GeneRepair requires a relatively low rate of (point) mutation - 2% of the alleles in the population are mutated on every generation. This low rate of mutation may be explained because the GeneRepair operator has a mutagenic effect, meaning that the background level of mutation can be somewhat lower than may otherwise be expected.

4.3 The GeneRepair Adjunct Operator

The GeneRepair operator is used in this paper to ensure that all solutions in the population are valid – that there are no omission or duplication errors in any of the solutions stored in the population. Such error can be generated from two different sources. Firstly, the crossover operator combines genetic information from two individual to create a new individual. We use single point crossover that chooses a single point along the allele sequence of both parents, combining the first half of one parent with the second half of the other parent. Thus a new individual is formed.

Genetic errors are identified when the genetic information of newly generate offspring violate the (mathematical) constraints of the TSP. We identify on two categories of error: *omission errors* and *duplication errors* (as discussed in Section 4.1 above).

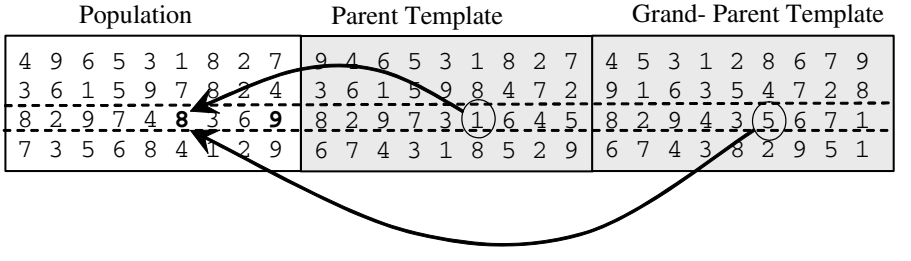

Fig. 2. Does Parent or Grandparent based Correction yields better results?

Mitchell [7] and Mitchell *et al* [8] and others [9] examined several biologically and non-biologically inspired templates, but did not explore the use of a grandparent based repair template.

4.3.1 Template Origin

Our first objective was to compare the performance of parent based GeneRepair with that of grand-parent based GeneRepair. Template driven GeneRepair operates in two phases as follows. The first phase (called error detection) identifies all occurrences of duplicate errors in the current population. In our first experiment, these duplicate errors were identified in a fixed left-to-right manner. So the second and subsequent occurrences of cities within a tour are detected as errors and are sent to the second phase, called correction. (This left-to-right decision shall be addressed further in Section 4.3.2 below.) These duplication errors can be seen as the bold figures in the Current Population of Figure 2 above.

The second phase (called error correction) of GeneRepair repairs the identified errors. While each individual was being examined, the cities of the current population are tagged in the parent and grandparent populations. Thus, un-tagged information in both populations form an ordered list of missing cities. These missing cities are used to replace the duplicate cities in a left-to-right manner.

The first experiment compared the effectiveness of parent and grandparent based GeneRepair, on the TSP problem described above. Table 1 summarizes these results, showing the shortest tour identified across these experiments and the mean results produced by each strategy.

Table 1. GrandParent based GeneRepair outperforms parent based repair

	Min	Mean
Parent Strategies	505.43	549.43
GrandParent Strategies	491.18	548.24

As shown above, the grandparent strategy far outperformed the parent strategy on these experiments. In fact, all grandparent based results outperformed all of the parent based results. Additionally, the relatively high mean of the Grandparent based repair was due to one particularly poor result of this strategy.

Not only did grandparent based repair generate better results, it did so in significantly fewer generations that the parent strategy. The grandparent strategy reached a result within 15% of the optimal in 5,500 generations while the parent strategy reached a result within 19% of the optimal in 8,350 generations. Also, we point out that our focus was on comparing strategies inspired by *Arabidopsis thaliana* and little effort went into tailoring our EO to generate good results for this problem set.

An explanation for the superior performance of grand-parent based repair, we turn to the differences between the offspring and its parent and grandparent. We point out that the grandparent has a higher probability of being *different* to the individual being repaired than its immediate parent. Thus, grandparent based repair generally has a larger disruptive effect on the individual than parent based repair. As our EO converges, the diversity in the population tends to reduce so that there is little difference between parent and offspring. (Mutation or even adaptive mutation is often used to counteract this tendency, allowing convergence to a global rather than a local optimum). Thus, we theorize that the grandparent proves to be a better template for repair than the parent, because of its potential for greater dissimilarity with the individual. This conclusion suggests that great-grandparent based repair should further outperform grandparent based repair – this being the subject of our current work. However, we do expect a decreasing pay-off as additional generations are archived in the repair process. As with Arabidopsis thaliana, it may well be that the additional expense of adding generations may not produce a commensurate payback in performance.

Another interesting observation arose from our analysis of these experiments. When a single occurrence of a duplication error is identified, both parent and grandparent strategies will generate the same *new* repaired individual. So when converging towards a global optimum for the given problem, we might expect fewer errors and

thus less of a difference between parent and grandparent strategies. Therefore, much of the difference between these two strategies will occur earlier in the evolutionary process.

4.3.2 Direction of Error Detection

The next experiment attempted to assess the impact that the direction of error detection has upon solution quality. In addition to the left-to-right error identification strategy, two other strategies were investigated: right-to-left and random direction. The left-to-right and right-to-left were fixed throughout whereas the random direction changed for every individual in each generation.

The next experiment compared these three repair directions: (i) operating repair from right to left, (ii) operating repair from left to right and (iii) operating repair in a random direction**.**

Table 2. The Random Direction GeneRepair Produced Best Results

	Min	Mean
Left-to-Right	471.44	519.67
Right-to-Left	483.36	529.27
Random	459.74	514.25

The results for this experiment are summarized in Table 2 above. Firstly, GeneRepair produces the best results when it proceeds in random and changing directions. The random strategy outperformed the two fixed direction strategies, on the best result generated and as an average across all runs of this experiment.

4.3.3 Fitness of Template

The final factor that we investigated was whether the fitness of the recorded ancestors had any impact on the goodness of the solutions generated. In the earlier experiments, at the end of each generation the genetic material of the fittest parent was recorded for each individual. This then formed part of the repair template for that individual.

In the next experiment, we explored the impact of recording a randomly selected parent for each individual. Thus for the randomly selected parent condition, the parent chosen to be moved into the repair genome was selected randomly, without any reference to the fitness of the two parents. It was expected that the superior fitness of the fittest condition would outperform the random parent conditions.

Table 3. Comparison of Fittest Ancestral Template with Random Ancestral Template

	Min	Mean
Random Parent	493.84	538
Fittest Parent	483.36	529.29
Random Grandparent	459.74	514.25
Fittest Grandparent	475.68	517.99

As shown above a random choice of ancestor proved to be superior to using the fittest of the two. This experiment also concretes the findings of the first experiment in that once again the grandparent was superior to the parent as a GeneRepair template. One explanation for the random ancestor from either generation outperforming the fittest ancestor may be deduced by examining the crossover technique used in this evolutionary strategy. The crossover is single point crossover where the point is chosen randomly for each individual. This means that the fittest ancestor does not necessarily have more impact on the individual than the other ancestor. We can theorize from this that because the fittest ancestor does not always have a larger impact on the individual it is not necessarily the best template to use for repair and a random template is more appropriate. The results above also confirm the results shown in Table 2 as the tour length of 459.74 which was achieved using a Random Grandparent repair template was found by conducting repair in a random direction

4.3.4 Penalty Points

We also examined the performance of the penalty points approach to enforce constraints, using the "death penalty" whereby invalid individuals are prevented from being used in crossover. The result of this experiment is shown in Table 4. As can be seen in Table 4 this approach produces significantly less-fit individuals that was produced by GeneRepair (Tables 1, 2 & 3).

Table 4. Death Penalty Approach

	Min	Mean
Death Penalty	1486.4	1584.94

4.3.5 Summary of Results

Each one of the repair directions described in Section 4.3.2 was tested for each of the inheritance template shown in Table 2 so there are in essence twelve different results to the experiment described in Section 4.3.3 rather than three.

The results of all twelve experiments are summarized in Fig. 3 below. The lines indicate the best solutions produced by each strategy across all runs of that strategy. The depicted results for each strategy were selected by choosing the best results at the end of the 500,000 generations. Each line indicates the best solution found thus far and as can be seen, this gradually converges towards the global optimal. (For this problem the known global optimal was 426. We were very pleased with the results of grandparent based repair – given that truncation selection was used. We expect even better results with Tournament selection and a much larger search space).

Best results are shown by the bottom line on Figure 3. This depicts the result for GeneRepair operating in a random direction using a randomly chosen grandparent as its repair template. This result is followed closely by the other GeneRepair techniques which use the grandparent as a repair template.

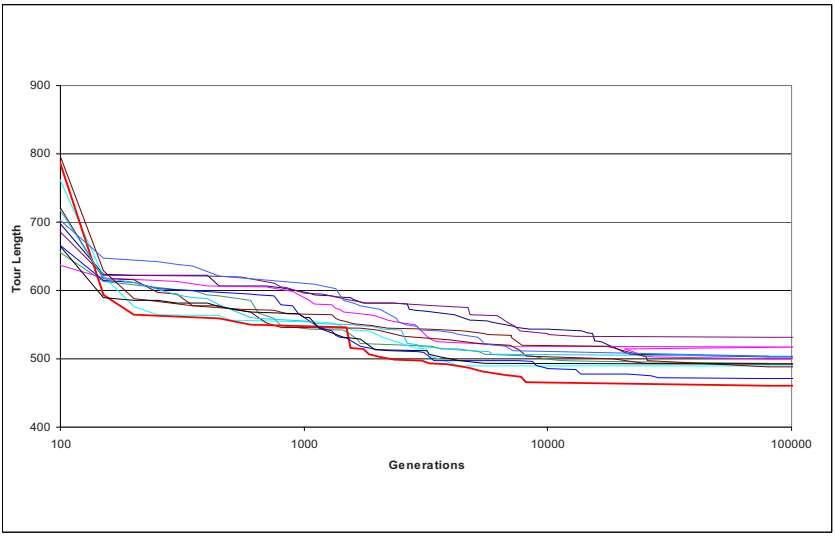

Fig. 3. Comparison of 12 different GeneRepair techniques

5 Conclusion

Evolutionary Optimization (EO) is an approach to optimization inspired by Darwin's idea of survival of the fittest. EO is a very are effective in exploring complex solution spaces, but are less suited to supporting validity constraints between the search parameters. This paper presents an approach to genetic repair that is inspired by the *Arabidopsis thaliana* plant. This plant is capable of making repairs to its own genes by making use of genetic information originating in the individuals grandparent. Controversially, this genetic information appears to skip the parent's generation. Our GeneRepair mechanism is inspired by an "archival DNA" explanation, though an alternative RNA based explanation exists. Errors in an individual plant's genes are repaired by comparison to the grandparents "template" DNA, which also serves to correct these errors.

We adapt this approach to genetic repair by applying it to a standard constrained optimization problem – the Traveling Salesman's Problem (TSP). This applied evolutionary optimization techniques, using unmodified crossover and mutation operators. An adjunction GeneRepair process ensured the validity of all solutions generated. This comparison found that GeneRepair based on a grandparent template produced better results than that of the parent based template. These results echoes recent advances in genetics, identifying non-Mendelian inheritance on the *Arabidopsis thaliana* plant [3]. A subsequent experiment indicates that archiving a randomly chosen parent produced better results than biasing the genetic archival process in favor of the fittest parent – and thus fittest grandparent. Our final experiment showed that the GeneRepair process produces best results when operating in random (and changing) directions. This approach outperformed both of the fixed direction strategies tested.

Not only do our results echo the controversial theory of Lolle *et al* [3], they also shed new light on this theory of non-Mendelian inheritance. First, that the repair seems to work best when it uses a randomly chosen grandparent and, secondly that repair should repair violations in a random order. Building upon Lolle's [3] results, these findings suggest a general approach to enforcing constraints on combinatorial optimization problems, opening up new possibilities for exploration.

Acknowledgement. We thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for their facilities and support.

References

[1] Coello Coello, C.: Theoretical and Numerical Constraint Handling Techniques in Evolutionary Algorithms: A Survey. Computer Methods in Applied Mechanics and Engineering 191(11-12), 1245–1287 (2002)
[2] Kalyanmoy, D.: An Efficient Constraint Handling Method for Genetic Algorithms. Comp. Methods in Applied Mechanics & Engineering 186(2-4), 311–338 (2000)
[3] Lolle, S.J., Victor, J.L., Young, J.M., Pruitt, R.E.: Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis. Nature 434, 505–509 (2005)
[4] Chaudhury, A.: Hothead healer and extragenomic information. Nature 437, E1-E2 (2005)
[5] Ray, A.: Plant genetics: RNA cache or genome trash? Nature 437, E1–E2 (2005)
[6] Michalewicz, Z., Schoenauer, M.: Evolutionary Algorithms for Constrained Parameter Optimization Problems. Evolutionary Computation 4(1), 1–32 (1996)
[7] Mitchell, G.G.: Evolutionary computation applied to Combinatorial Optimization Problems. PhD Thesis, Dublin City University, Dublin, Ireland (2007)
[8] Mitchell, G.G., O'Donoghue, D.P., Trenaman, A.: A New Operator for Efficient Evolutionary Solutions to the Traveling Salesman Problem. Applied Informatics, 0-88986-280-X, 771-774 (2000)
[9] Orvosh, D., Davis, L.D.: Shall We Repair? In: Proc. 5th Intl. Conf. on Genetic Algorithms (1993)