Empirically Evaluated Improvements To Genotypic Spatial Distance Measurement Approaches For The Genetic Algorithm

by

Robert Collier

A Thesis
presented to

The University of Guelph

In partial fulfilment of requirements
for the degree of

Doctor of Philosophy

in

Computer Science

Guelph, Ontario, Canada

© Robert Collier, March, 2012
ABSTRACT

EMPIRICALLY EVALUATED IMPROVEMENTS TO
GENOTYPIC SPATIAL DISTANCE MEASUREMENT APPROACHES
FOR THE GENETIC ALGORITHM

Robert Collier
University of Guelph, 2012

Advisor:
Professor M. Wineberg

The ability to visualize a solution space can be very beneficial, and it is generally accepted that the objective of visualization is to aid researchers in gathering insight. However, insight cannot be gathered effectively if the source data is misrepresented. This dissertation begins by demonstrating that the adaptive landscape visualization in widespread usage frequently misrepresents the neighborhood structure of genotypic space and, consequently, will mislead users about the manner in which solution space is traversed by the genetic algorithm. Bernhard Riemann, the father of topology, explicitly noted that a measurement of the distance between entities should represent the manner in which one can be brought towards the other. Thus, the commonly used Hamming distance, for example, is not representative of traversals of genotypic space by the genetic algorithm – a representative measure must include consideration for both mutation and recombination. This dissertation separately explores the properties that mutational and recombinational distances should have, and ultimately establishes a measure that is representative of the traversals made by both operators simultaneously.
It follows that these measures can be used to enhance the adaptive landscape, by minimizing the discrepancy between the interpoint distances in genotypic space and the interpoint distances in the two-dimensional representation from which the landscape is extruded. This research also establishes a methodology for evaluating measures defining neighbourhood structures that are purportedly representative of traversals of genotypic space, by comparing them against an empirically generated norm. Through this approach it is conclusively demonstrated that the Hamming distance between genotypes is less representative than the proposed measures, and should not be used to define the neighbourhood structure from which visualizations would be constructed.

While the proposed measures do not distort the data or otherwise mislead the user, they do require a significant computational expense. Fortunately, the choice to use these measures is always made at the discretion of the user, with additional costs incurred when accuracy and representativity are of paramount importance. These measures will ultimately find further application in population diversity measurement, cluster analysis, and any other task where the representativity of the neighborhood structure of the genotypic space is vital.
# Contents

1 Introduction  
1.1 Thesis Research: Overview ................................................................. 8

2 Background Literature  
2.1 Genetic Algorithm Overview ............................................................... 14  
2.2 Application of Measurement to Visualization ....................................... 17  
2.2.1 Data Visualization Objectives and Evaluation .................................... 18  
2.2.2 Visualization of Evolutionary Algorithms ......................................... 22  
2.2.3 Spatial Dimensionality Considerations .............................................. 31  
2.3 Typical Approaches to Distance Measurement ..................................... 42  
2.3.1 Distance Measurement Between Strings .......................................... 43  
2.3.2 Population Diversity Measurement .................................................. 44  
2.3.3 Measures Used in Evolutionary Computation .................................... 46  
2.4 Genetic Algorithm Operator Considerations ....................................... 47  
2.5 Distance Measurement Approach Evaluation ...................................... 52  
2.5.1 Extradisciplinary Inspiration ......................................................... 53

3 Operator Traversal Representativity  
3.1 Mutational Distance Measurement ..................................................... 60
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Considerations Specific to Recombination</td>
<td>65</td>
</tr>
<tr>
<td>3.3</td>
<td>Unary Redefinition of Recombination</td>
<td>67</td>
</tr>
<tr>
<td>3.4</td>
<td>Classification of Recombinational Distances</td>
<td>73</td>
</tr>
<tr>
<td>3.5</td>
<td>Recombination Operator Induced Topology</td>
<td>78</td>
</tr>
<tr>
<td>3.6</td>
<td>Abstractions of the Recombination Operator</td>
<td>82</td>
</tr>
<tr>
<td>3.7</td>
<td>Digraphic Representation of Recombination</td>
<td>89</td>
</tr>
<tr>
<td>3.8</td>
<td>Time Complexity Considerations</td>
<td>98</td>
</tr>
<tr>
<td>3.9</td>
<td>Novel Approaches to Measurement</td>
<td>101</td>
</tr>
<tr>
<td>3.10</td>
<td>Additional Time Complexity Considerations</td>
<td>123</td>
</tr>
<tr>
<td>3.11</td>
<td>Operator Representative Distances Summary</td>
<td>124</td>
</tr>
<tr>
<td>4</td>
<td>Mechanism Traversal Representativity</td>
<td>127</td>
</tr>
<tr>
<td>4.1</td>
<td>Simplistic Distance Measure Combinations</td>
<td>130</td>
</tr>
<tr>
<td>4.2</td>
<td>Simplistic Population Centroid Approach</td>
<td>135</td>
</tr>
<tr>
<td>4.3</td>
<td>Integration of Recombinational Distance</td>
<td>138</td>
</tr>
<tr>
<td>4.4</td>
<td>Integrated Variational Distance Definitions</td>
<td>145</td>
</tr>
<tr>
<td>4.5</td>
<td>Time Complexity Analysis</td>
<td>158</td>
</tr>
<tr>
<td>4.6</td>
<td>Mechanism Representative Distance Summary</td>
<td>162</td>
</tr>
<tr>
<td>5</td>
<td>Empirical Distance Measurement</td>
<td>165</td>
</tr>
<tr>
<td>5.1</td>
<td>Analogous Geometric Construction</td>
<td>167</td>
</tr>
<tr>
<td>5.2</td>
<td>Important Considerations</td>
<td>172</td>
</tr>
<tr>
<td>5.3</td>
<td>Adaptation to the Genetic Algorithm</td>
<td>176</td>
</tr>
<tr>
<td>5.4</td>
<td>Empirical Approach Validation</td>
<td>181</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Genotypic and Phenotypic Spaces .......................................................... 5
Figure 2: Research Summary Graphic .................................................................. 10
Figure 3: Manhattan and Chessboard Distances .................................................. 12
Figure 4: Wijk's Visualization Model ................................................................. 20
Figure 5: Example Line Graph of Convergence .................................................. 23
Figure 6: Example Scatter Plot of Convergence ................................................... 24
Figure 7: Evolutionary Activity Wave Diagram Example ..................................... 25
Figure 8: Modified Hinton Diagram Example ..................................................... 26
Figure 9: Topographic Depiction of an Adaptive Landscape .............................. 28
Figure 10: Three-Dimensional Surface Depiction of an Adaptive Landscape...... 28
Figure 11: The Magenta Square Optical Illusion ................................................. 33
Figure 12: Star-Type Glyph Visualization Example .......................................... 36
Figure 13: Chernoff Face Visualization Example ............................................. 37
Figure 14: Parallel Co-ordinate Visualization Example ..................................... 38
Figure 15: Andrew's Curve Visualization Example .......................................... 39
Figure 16: Hypercubic Topology ......................................................................... 61
Figure 17: Possible Mutational Transitions ......................................................... 63
Figure 18: Complementary Index Sets ............................................................... 69
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Possible Operands for Recombination</td>
<td>70</td>
</tr>
<tr>
<td>20</td>
<td>Unary Recombination Paradigm</td>
<td>71</td>
</tr>
<tr>
<td>21</td>
<td>Unsatisfaction of Subadditivity</td>
<td>76</td>
</tr>
<tr>
<td>22</td>
<td>Unsatisfaction of the Requirements for a Topological Space</td>
<td>81</td>
</tr>
<tr>
<td>23</td>
<td>Depicting Recombination With Genotype Pairing Vertices</td>
<td>84</td>
</tr>
<tr>
<td>24</td>
<td>Depicting Recombination With a Bipartite Graph</td>
<td>85</td>
</tr>
<tr>
<td>25</td>
<td>Depicting Recombination With a Hypergraph</td>
<td>87</td>
</tr>
<tr>
<td>26</td>
<td>Overview of the Digraphic Representation of Recombination</td>
<td>93</td>
</tr>
<tr>
<td>27</td>
<td>Digraphic Representation of Recombination</td>
<td>103</td>
</tr>
<tr>
<td>28</td>
<td>Enhanced Digraphic Representation of Recombination</td>
<td>106</td>
</tr>
<tr>
<td>29</td>
<td>Offspring Possible With Multiple Applications of Recombination</td>
<td>109</td>
</tr>
<tr>
<td>30</td>
<td>Depiction of the Proposed Greedy Enhancement</td>
<td>118</td>
</tr>
<tr>
<td>31</td>
<td>Upper Portion of the Research Summary Graphic</td>
<td>126</td>
</tr>
<tr>
<td>32</td>
<td>Mutational and Recombinational Transition Depictions</td>
<td>134</td>
</tr>
<tr>
<td>33</td>
<td>Ranking According to Hamming Distance</td>
<td>139</td>
</tr>
<tr>
<td>34</td>
<td>Ranking According to Recombinational Distances</td>
<td>140</td>
</tr>
<tr>
<td>35</td>
<td>Representing Recombination With the Distance from the Centroid</td>
<td>142</td>
</tr>
<tr>
<td>36</td>
<td>Digraph Representation With Adjusted Centroid</td>
<td>143</td>
</tr>
<tr>
<td>37</td>
<td>Recombinational and Mutational Traversals</td>
<td>144</td>
</tr>
<tr>
<td>38</td>
<td>Offspring Schemata Following Recombination</td>
<td>149</td>
</tr>
<tr>
<td>39</td>
<td>Proposed Integration of Mutational and Recombinational Distances</td>
<td>155</td>
</tr>
<tr>
<td>40</td>
<td>Lower Portion of the Research Summary Graphic</td>
<td>164</td>
</tr>
<tr>
<td>41</td>
<td>Geometric Construction Analogy</td>
<td>169</td>
</tr>
</tbody>
</table>
Figure 42: Genotypic Space of Hypercubic Topology.................................170
Figure 43: Ranking Genotypes by Empirical Distance .........................180
Figure 44: Recombination With an Empty Population..............................190
Figure 45: Connected Digraph Created With a Transition Matrix..............191
Figure 46: Demonstrative Example..........................................................192
Figure 47: Correlation of Transition Matrix and Empirical Approaches .......194
Figure 48: Correlation of Hamming and Empirical Approaches for Mutation.....195
Figure 49: Research Summary Graphic Revisited ....................................197
Figure 50: Fixed Population Anchoring in Genotypic Space....................204
Figure 51: Representativity for Populations of Uniform Distribution .........213
Figure 52: Normality Tests.......................................................................214
Figure 53: Representativity for Populations of Linear Distribution............216
Figure 54: Representativity for Populations of Exponential Distribution.....217
Figure 55: Rearrangement of Genotypes by Hamming Distance...............226
Figure 56: Adaptive Landscapes Extruded from Rearranged Genotypes ......227
Figure 57: First Multidimensionally Scaled Adaptive Landscape Example......229
Figure 58: Second Multidimensionally Scaled Adaptive Landscape Example.....230
Figure 59: Third Multidimensionally Scaled Adaptive Landscape Example ......231
List of Tables

Table 1:  Truth Table........................................................................................................ 97
Table 2:  Transition Matrix Fragment for Mutation.....................................................205
Table 3:  Transition Matrix Fragment for Recombination.............................................205
Table 4:  Transition Matrix Fragment for Complete Exploratory Mechanism....206
Chapter 1

Introduction

At the time this body of research was conducted, the field referred to as evolutionary computation is not more than 50 years old (Kallel, Naudts, and Rogers 2001), with one of the cornerstones, the simple genetic algorithm, having been popularized by John Holland only about 35 years ago (Holland 1975). Holland and his contemporaries (Fogel, Owens, and Walsh 1966; Rechenberg 1964, 1973; Schwefel 1965, 1981) were seeking to exploit the evolutionary mechanism and solve complex optimization problems by emulating processes that had been observed in the natural world. Although numerous variations upon the original genetic algorithm have been introduced in the last four decades, the simple genetic algorithm remains, fundamentally, a population simulation of candidate solution representations. It is also widely accepted (Mitchell 1998; Spears 2000) that a simple genetic algorithm will invariably subject whatever population of candidate solution configurations corresponds to the current iteration of the algorithm to both selection and variation operations, in order to construct a new population suitable for use in the next iteration.
In spite of the fact that the original algorithm (Holland 1975) specified that the population of candidate solutions would be represented as a set of binary strings (often referred to as chromosomes) and replaced entirely after each iteration of the algorithm (i.e. generationally), since the inception of the genetic algorithm, researchers have thoroughly explored adaptations employing both integer and real-valued representations (Rothlauf 2002) and steady state populations (Whitley and Kauth 1988) wherein only one population member is replaced after each iteration of the algorithm. In spite of these developments, the binary and generational genetic algorithm remains a topic of interest to the researcher community (Ortegon-Cano, Hartasanchez, and Stephens 2010; Ter-Sarkisov, Marsland, and Holland 2010; Vafae, Turán, and Nelson 2010), and is frequently employed in novel applications (Downing 2010; Chen and Szeto 2010; Crofford, Eskridge, and Hougen 2010; Lien, Yu, and You 2010; Yehoshua, Avigal, and Unger 2010).

Although a population is formally defined in the context of the genetic algorithm as a multiset of individual candidate solutions (Holland 1975, Mitchell 1998), researchers recognize that some candidate solution configurations are more similar than others. Holland's foundational work further proposed that the set of all possible structures defining the search space for the genetic algorithm should be treated as a set of subsets, where structures in the same subset would have attributes in common (Holland 1975). This became known as the schema theorem for genetic algorithms, and for binary strings of a specified length, a template string referred to as a schema (constructed from the union of the binary alphabet with a unique wildcard character to create a ternary
alphabet) would describe such a subset. To possess membership in a specified schema it is necessary for the binary representation contain the same character at the same index in the schema for every index not assigned a wildcard character, and it was acknowledged (Holland 1975) that membership in a schema indicates some degree of similarity with other members.

This similarity can be extended to define a conceptualization of neighbourhood and implies that the population of chromosomes is actually sampled from a space of chromosomes, as opposed to a set of chromosomes. A space can be formally defined as a set upon which some additional structure has been imposed (Joshi 1983), and any ordered sequence of values can be treated as a point in the space of the same dimensionality as the number of values in the sequence (Nicholson 1995). Thus, the complete set of candidate solution configurations is actually a space, and the population being simulated by the genetic algorithm is sampled from this space. Furthermore, although the binary chromosomes are typically representative of a set of real parameter values that specify a possible solution to the optimization problem in question, the evolutionary mechanism of the simple genetic algorithm introduced by Holland did not act upon ordered sequences of real values, but rather upon the space of binary string representations (Goldberg 1989, Rothlauf 2002).

The heuristic function that would define the evolutionary mechanism of the genetic algorithm is the composition of a selection operation and a variation (or, equivalently, a perturbation or mixing) mechanism. The variation mechanism itself is typically the
composition of a unary mutation operation and a binary recombination (also known as crossover) operation (Spears 2000, Vose 1999). The selection operation is tasked with the exploitative identification of a random subset of the current population that will be used in forming the next generation. The mutation and recombination operations, on the other hand, are expected to generate novel elements of the candidate solution space using members of the current population, to ensure the genetic algorithm is both explorative and exploitative. The selection mechanism defines a probability distribution that each member of the current population will be used in constructing the next population, but while the selection probability associated with a chromosome is typically determined using the fitness function value, the most common implementations of mutation and recombination are randomly applied to the selected subset with a uniform distribution (Mitchell 1998). This entails, naturally, that while the selection component of the evolutionary mechanism is typically dependent on the fitness function, the variation component is typically not.

Nevertheless, as both the selection and variation components are applied to the binary candidate solution representations, the evolutionary mechanism must then operate upon a different space than that which would be intuitive to the researcher that formulated the problem. This distinction is very well understood (Rothlauf 2002) and, as with much of the terminology used in describing the genetic algorithm, the representation of a candidate solution and the solution itself are referred to as the genotype and phenotype, respectively. The distinction was first proposed by geneticist Wilhelm Johannsen (Johannsen 1911), with the genotype being defined as the genetic composition of an
organism (i.e. the chromosome) and the phenotype being the organism itself, the traits of which having been expressed from the genetic material (Hartl and Clark 1997). With this distinction in mind, it is easy to recognize that the mapping of chromosomes (binary candidate solution representations) to fitness function values is itself comprised of a mapping of the elements of genotypic space to the elements of phenotypic space, following by a mapping of the elements of phenotypic space to the space of the possible values of the fitness function (Rothlauf 2002, Lewontin 1974). This relationship is depicted in Figure 1.

Figure 1. For a population of four candidate solutions, the chromosomal representations from genotypic space can be mapped to phenotypic space. The phenotypes can then be mapped to real values using the fitness function.
Thus, researchers who wish to exploit an understanding of the properties of genotypic space (upon which the evolutionary mechanism of the simple genetic algorithm is known to act) must then explore the structure of genotypic space using measures of the distances that exist between genotypes. To accomplish this, researchers frequently employ a data visualization technique known as the adaptive landscape (Wright 1932), which will be introduced and discussed at length in the survey of relevant background literature found in Chapter 2.

The adaptive landscape visualization technique attempts to accurately convey the position and proximity (in the underlying genotypic space) of the candidate solution configurations that are present in the current population, using a two-dimensional depiction of genotypic space from which a three-dimensional surface can then be extruded. The approach by which adaptive landscapes can be constructed is very useful for a number of important tasks, including, but not limited to, data visualization, population diversity measurement, and cluster analyses (McGinley, Maher, et al. 2011, Zhang 2007). However, for the adaptive landscape to convey useful information, the neighbourhood structure of the genotypic space (from which the underlying plane is derived) must be carefully constructed – as the genotypic space is typically traversed by the variational operations of the genetic algorithm (i.e. the unary mutation and binary recombination operators), the neighbourhood structure should consequently be derived from these same variational operators, since they constitute the exploratory mechanism by which the genetic algorithm will traverse genotypic space (Jones 1995b).
Although this conclusion is logical and straightforward, since the effect of a recombination operation cannot be considered independent of the genotypes contained in the current population (because it is a binary operator), the underlying plane from which the adaptive landscape is extruded is typically only constructed with respect to the perceived traversal of genotypic space by a mutation operator (that has a unary arity), and not any of the binary recombination operators. This dissertation will demonstrate that this approach uses an unrepresentative measure of the genotypic spatial distances traversed by the variational component of the evolutionary mechanism of the simple genetic algorithm, and that the use of this unrepresentative measure will invariably lead to severe data distortion and a misrepresentation of the actual neighbourhood structure of genotypic space.

This thesis will introduce and validate a representative measure of the genotypic spatial distance traversed by the genetic algorithm, to be used in defining neighbourhood structures for genotypic spaces that can be used to construct quantifiably superior adaptive landscapes. While demonstrating that a truly representative measure of the genotypic spatial distances traversed is definitely achievable, an empirical technique by which the actual neighbourhood structure traversed by any evolutionary system (and the genetic algorithm specifically) will be introduced to show that the use of the novel measure of genotypic spatial distance (that will be proposed by this thesis for use in the construction of superior adaptive landscape visualizations) will yield a measurable improvement over the typical approach.
1.1 Thesis Research: Overview

The next chapter investigates the relevant background literature, beginning with foundational material on the measurement of distance and the processes of data visualization (and, more specifically, evolutionary algorithm visualization and adaptive landscapes), and exploring the different approaches to handling high dimensionality data, measuring the distance between strings (and other entities employed in evolutionary algorithms), and analyzing the effect of the binary recombination operator used by the genetic algorithm. The chapters that follow this literature review describe the development, validation, and ultimate application of distance measures that are truly representative of the manner in which genotypic space is traversed by the explorative component of the evolutionary mechanism of the genetic algorithm. An overview of the complete development (preceding application) is provided in Figure 2, accurately depicting the course of the investigation that forms this dissertation.

Beginning first with an analysis of the Hamming distance measurement approach, commonly applied to genotypic spatial distances, Chapter 3 will focus on the development of measures for which the assigned values are representative of the length of actual traversals of genotypic space, separately, by each of the two standard genetic operators. Chapter 4 will explore the unification of these two measures (mutation and recombination) in the development of measures that are more representative of the manner in which genotypic space is traversed by the complete explorative component of the evolutionary mechanism. The resulting measures are expected to address many of the
shortcomings the Hamming distance approach to measuring genotypic spatial distances, but to test this expectation a specialized methodology must be introduced for the comparative evaluation of distance measures.

Chapter 5 describes the development of that methodology, referred to as the empirical distance measure, where distances are determined using a carefully designed Monte Carlo simulation. The validity of this approach is then assessed using the infinite population model (Vose 1999) of the genetic algorithm. With the creation of a validated approach to empirical distance measurement, the effectiveness of the measures introduced in Chapters 3 and 4 can be thoroughly assessed, and this evaluation is presented in Chapter 6. At this point, the divide between the customary approach to genotypic spatial distance measurement and the properties of the genetic algorithm itself has been effectively bridged and the relative performance associated with a specific measure can be explored. The visualization application of genotypic spatial distance measures, established previously as a principle motivation, is thoroughly explored in Chapter 7, after which final conclusions about the construction of representative measures and the empirical approach to measurement are discussed.
Figure 2. A summary of the content of this dissertation, revisited at the end of each chapter. The chart illustrates the process by which mutational and recombinational distance measures were developed (Chapter 3) and integrated (Chapter 4), while a methodology was developed (in parallel) for empirically measuring genotypic spatial distances (Chapter 5). The empirical measure is then used in the comparative evaluation of the proposed measures (Chapter 6). Chapter 7 (not depicted above) will explore an application of the novel measures and present a concluding summary.
Chapter 2

Background Literature

Historically, sailors that had become lost in coastal waters would release a crow from captivity to determine the direction of the nearest land mass, originating the idiom of "as the crow flies" as an expression for the shortest distance between points in space (Detweiler 2007). Although this is associated with the common Euclidean measure of distance and the L-2 vector norm (Krause 1975; Nicholson 1995), in practice, knowledge of the distance between points "as the crow flies" cannot always be used to estimate a traversable distance. For the traversal of the distances in a city that is structurally arranged into blocks, the Euclidean distance between intersections is not representative of the number blocks that must be traversed by an automobile. Similarly, the traversal of a chessboard by a queen also cannot be accurately measured by the Euclidean distance.

Mathematicians have introduced very different measures, known as the Manhattan and Chessboard measures of distance (Krause 1975, Agarawal and Sahoo 2008), that reflect the traversal of two-dimensional space described in the former and latter scenarios,
respectively. These measures, depicted in Figure 3, were introduced because the mechanisms by which two-dimensional spaces are being traversed (the automobile and the king, respectively) are not reflected by Euclidean distance measurement.

Figure 3. The Manhattan distance (left) and Chessboard distance (right) over a space of 16 discrete points. The mechanisms by which the spaces are traversed (the automobile and the king, respectively) determine the number of unit transformations necessary for the target point to be reached from the origin.
Similarly, as the evolutionary mechanism of the genetic algorithm explores genotypic space using specialized mutation and recombination operators, useful distance measures must reflect the manner in which that space is traversed by the evolutionary mechanism of the genetic algorithm. As Bernhard Riemann, attributed to be the father of topology (Bourbaki 1994), once stated, "...measurement consists of a superposition of quantities to compare; to measure, there must therefore be a means of bringing one quantity onto another," (Riemann 1854). With the knowledge that there are, then, as many measures of distance as there are approaches to traversal, measuring distances in genotypic space as they would be traversed by the variational operators necessitates that that complete set of requirements for a distance measure classification be formally defined.

Given a set $S$ of points, any function $\delta$ that maps an element of $S^2$ to either an infinite or real value can be considered a measure of distance. This measure ($\delta$) could only be classified more formally as a distance metric if the following conditions are met (Burago 2001), for all points $x, y, z \in S$:

1. $\delta(x, y) \geq 0$ (the non-negativity condition),
2. $\delta(x, y) = 0 \leftrightarrow x = y$ (the identity of indiscernibles condition),
3. $\delta(x, y) = \delta(y, x)$ (the symmetry condition), and
4. $\delta(x, y) + \delta(y, z) \geq \delta(x, z)$ (the subadditivity condition).

The first two conditions are often grouped together (as a single condition of positive definiteness), and the fourth condition is frequently referred to as the triangle inequality.
Although each of these conditions must be upheld for a measure of distance to be classified as a metric (and, consequently, for a set of points to be considered a metric space), more generalized forms (relaxing one or more conditions) have also been defined. The prefixes pseudo-, quasi-, and semi- are affixed to the term metric to indicate a measure that fails to uphold the positive definiteness, symmetry, and subadditivity conditions, respectively (Hart, Nagata, and Vaughan 2004). It should be noted that although the measurements of distance are sometimes treated as simple quantifications of dissimilarity, the measures of genotypic spatial distance discussed in this body of research are taken to represent the ease with which one genotype could be transformed into another by the explorative component of the evolutionary mechanism. It follows then that not all measures of distance need also be defined as metrics - the classification assigned to any novel measure of distance by the preceding criteria is, thus, invaluable for determining the possible application of such a measure. Furthermore, for the development of measures that are representative of the genetic algorithm, it is necessary to define the properties of the generalized genetic algorithm that is the subject of this dissertation.

2.1 Genetic Algorithm Overview

Since the genetic algorithm is essentially a population simulation configured to emulate the evolutionary mechanism observed in the natural world, the operators by which it searches the space of candidate solution are inspired by the real-world processes of mutation and recombination. For the binary, generational genetic algorithm that forms the
basis for this research, every candidate solution (or phenotype) to the optimization problem being solved is mapped to a binary string (or genotype), and a randomly selected initial multiset of genotypes (herein referred to as the current population) is subjected to the four processes – evaluation, selection, variation, and replication – repeatedly and in sequence, until some termination conditions are met.

The evaluation phase of the genetic algorithm computes the performance of each candidate solution in the population (with respect to the optimization problem being solved), and this information is used in the selection phase of the algorithm to randomly identify a submultiset of the current population for further investigation. In most cases, the relative performance of a candidate solution is evaluated using a fitness function for which the function inputs correspond to the variables of a candidate solution genotype and the function output is a value that can be normalized to the range [0, 1], with the values zero and one corresponding to the worst and best possible solutions, respectively.

After each candidate solution in the population has been evaluated, the exploitative and explorative mechanisms of the genetic algorithm are applied in sequence to direct the search, with the former mechanism using only the evaluation function information present in the current population and the latter mechanism submitting new genotypes for addition to the population. These mechanisms correspond to the processes of selection and variation, respectively, and for the genetic algorithm to function properly, candidate solutions that exhibit a relatively superior performance (to the other candidate solution in the population) should be selected, on average, more frequently than candidate solutions
that exhibit relatively inferior performances – the selection strategy then reflects the evolutionary principle of the survival of the fittest.

During the variation phase of the algorithm, the candidate solutions of this subpopulation are perturbed, using both a unary mutation operation (wherein some of the binary digits of the genotype are flipped to their complementary values) and a binary recombination operation (wherein binary digits are exchanged between a pair of genotypes to produce alternative, offspring genotypes). The subpopulation (having been perturbed such that it represents a different set of candidate solutions) then replaces the current population entirely – a practice that can be referred to as a generational replication strategy.

Following the popularization of the genetic algorithm by Holland in 1975, numerous selection, variation, and replication techniques have been introduced, explored, and comparatively evaluated. Although a survey of these techniques would exceed the scope of this dissertation (and is not necessary for comprehending either the methodology or the conclusions of this research), it is important to note that the recombination operation used by Holland (i.e., one-point crossover) has been replaced by the uniform crossover operator for this investigation. The one-point form of recombination takes (as input) two parent solutions, each represented by a binary string, and randomly select an index such that an offspring produced would be the concatenation of a substring selected from each parent, from the first index to the randomly selected index, and from the randomly selected index to the final index, respectively. For reasons that are described in greater detail in Chapter 3, this dissertation considers uniform recombination instead, where each
binary digit of the offspring is selected probabilistically and independently from either parent. Researchers that are entirely unfamiliar with the use of the genetic algorithm and require a more comprehensive introduction to the topic are directed to the texts authored by Holland (1975), Goldberg (1989), and Mitchell (1998).

This body of research has endeavoured to remain entirely fitness function independent (i.e., no assumptions about the evaluation phase of the algorithm are made) and does not assume the application of any particular selection strategy. Having specified that the replication phase of the algorithm will be performed generationally, it remains to investigate the manner in which genotypic space is traversed by the explorative component of the evolutionary mechanism (i.e., the mutation and recombination operators) and, from these operations derive representative measures of distance.

2.2 Application of Measurement to Visualization

The ability to accurately measure the distance between the elements of a data set is fundamental to any representative data transformation, and as such transformations are elementary to the task of constructing exploratory graphics (Hansen and Johnson 2005). Consequently, the field of data visualization offers one of the broadest fields to which accurate and representative measures of the distances can be applied. Exploratory graphics are contrasted from presentational graphics in that, where the latter are concerned with summarizing information and supporting conclusions, the former are constructed to aid analysts and researchers in the search of information (Chen, Härdle,
and Unwin 2008). Notwithstanding the many approaches to data visualization that also rank aesthetic considerations as a foremost objective (Wijk 2005), the widely accepted principles of visualization design presented by Tufte were supplemented with the strong emphasis that "it is wrong to distort the data measures... in order to make an editorial comment or fit a decorative scheme," (Tufte 1983).

Although it has been claimed (perhaps facetiously) that the genuine motivation for developing data visualization techniques is to secure funding (Wijk 2005), earnest definitions have uniformly emphasized that the primary objective of data visualization is to facilitate the acquisition of insight (Wijk 2005) and the conveyance of information that, although concealed, is inherent to the data (Hansen and Johnson 2005). Having already emphasized the distinction between exploratory and presentation graphics (Chen, Härdle, and Unwin 2008), in the context of advancing exploratory tools for application to the genetic algorithm, accepting insight as the chief motivation for data visualization necessitates an operational definition of insight.

2.2.1 Data Visualization Objectives and Evaluation

Developers are sometimes hesitant to cite insight as the primary objective for the field of data visualization because of the intangibility of the concept. This engenders an impression that it is difficult to explicitly evaluate specific data visualization approaches. Although the word insight, in general usage, refers to the outcome of apprehending the inner or hidden nature of a situation (Merriam-Webster Online), a single insight has been
defined in the context of data visualization evaluation as a "unit of discovery" (Saraiya, North, Duka 2005), often corresponding to a particular task that could only be achieved by a person who possess the relevant insight (North 2006). Such insights are characterized, to an extent, by their complexity (in that they are not simply data values themselves), relevance (in that they are pertinent to the domain of the data), and unexpectedness (in that they must be novel and are typically unpredictable) (North 2006).

The correspondence between known insights into a data set and the specific user tasks with which these insights could be associated does suggest that empirical approaches, such as controlled experiments, usability testing, and cognitive walkthroughs (Tory and Möller 2004) can be used to evaluate data visualizations. However, many researchers agree that attempts should be made to measure insight more directly (North 2006, Wijk 2005). Wijk (2005) provided a formal model for the process of data visualization that, although deceptively simple, does include an economic component to elaborate upon the notion of visualization evaluation. In this model, depicted in Figure 4, the image 'I' of a visualization is a function of the data 'D' and the parameter specification 'S'. The model also specifies that the perception 'P' of the user can only be altered by the image of the visualization, and will impart changes in the knowledge base 'K' of the observer.

Each of the changes in knowledge ΔK could be identified as an individual insight, and, although they remain difficult to quantify, Wijk (2005) noted that the veracity of the change in the knowledge base of the user is not guaranteed because the information is generated by the perception of the image of the visualization and not by the raw data.
directly. Consequently, it is possible for a visualization image to distort the actual data such that the user perceives a nonexistent feature and weakens the knowledge base accordingly. Furthermore, this explicit observation (that not every insight is accurate) entails that a given visualization would be improved by the removal of any misleading features in the image that were perceived to the detriment of the knowledge base of the user. Tufte stated that visualizations from which misleading features had been removed "are necessarily better within the principles of the theory, for more information per unit of space and per unit of ink is displayed," (Tufte 1983).

Figure 4. Wijk's visualization model, depicting visualization as the process V that generates an image I from the raw data D. This image is interpreted by the process P of perception by the user and used to contribute to that user's knowledge K. This knowledge can then be used to conduct an exploration E of the data by altering the parameter specification S governing the visualization process.
In addition to evaluating the effectiveness of a visualization approach in facilitating the processes by which users acquire insight, an approach should also be evaluated according to the efficiency by which results can be achieved, with respect to both temporal and monetary expenses (Wijk 2005). Consequently, the economic component of the data visualization model introduced by Wijk (2005) for evaluation purposes set forth a methodology by which the value of a visualization would be computed as the weighted difference between the sum of the changes to the knowledge base of the user $W(\Delta K)$ and the various costs associated with the use of the visualization approach. Wijk constituted the distinction between the cost $C_i$ associated with the initial development of the approach, the cost $C_u$ associated with the training of a user, the cost $C_s$ associated with each sessional construction of an actual image, and the cost $C_e$ associated with each exploratory time step required by a user before an insight is acquired. For a presumably homogenous collection of $n$ users, each using the visualization approach to construct $m$ images, with $k$ exploratory steps per session, the net benefit associated with the use of a visualization approach can be written:

$$W(\Delta K) - C_i + nC_u + nmC_s + nmkC_e.$$  

In spite of the obvious simplification of the model by the assumption of a homogenous user community, it is accepted that although the individual differences between users should be acknowledged, simpler interfaces are correlated with improved performance (Chen and Yu 2000). Furthermore, an emergent pattern of incremental development has been observed (Wijk 2005) to circumvent the typically high initial development costs $C_i$.  

21
As such, modern data visualization approaches employed in the construction of exploratory images are often derived from established techniques.

### 2.2.2 Visualization of Evolutionary Algorithms

Considerable effort has already been invested in the development of approaches suitable for evolutionary algorithm visualization – individual techniques can be classified according to the number of individual candidate solution depictions contained in an image (i.e. the depiction of an entire population versus the depiction of an individual of interest), and by the number of generations depicted in the image (i.e. the state of the population or the course of the current algorithmic process) (Pohlheim 1999).

Perhaps the simplest application of a data visualization approach to an evolutionary algorithm is the convergence diagram (Pohlheim 1999), depicted in Figure 5. In every optimization problem that can be solved there exists at least one candidate solution for which the corresponding fitness function value is never inferior to any other possible solution. Determining the chromosome of this globally optimal solution is, in most situations, the sole objective of an evolutionary algorithm. As such, the actual progress of a genetic algorithm could arguably be depicted as the fitness function value of the best individual chromosome plotted against the generation in which that chromosome appeared. If the algorithm is effectively performing an optimization, the graph will exhibit a trend towards the optimal objective value. Although the simplicity of the convergence diagram certainly decreases initial development and user training costs, this
same simplicity does result in a considerable loss of information. Naturally, researchers are often interested in detailing the state of the population beyond a mere representation of the best or average candidate solution configuration or fitness (Pohlheim 1999), and must employ more complex visualizations.

Figure 5. A simple line graph demonstrating the convergence of a population onto the candidate solution of maximum fitness.

The scatter plot offers a means by which the distribution of the entire population at each generation can be depicted. Although noisy, such graphs do provide users with a sense of the changing distribution of the population (Pohlheim 1999), as is evidenced in Figure 6. These graphs do, naturally, depict the same trend towards optimality as the convergence
diagram, but they also facilitate the conjecture, for each generation, that either the population has converged entirely on a single candidate solution or it still possesses a nontrivial degree of diversity. Unfortunately, the depiction of the entire population subspace at each generation in a two-dimensional image entails that the dimensionality of each population is reduced to a single dimension – a considerably deceptive process that will be discussed in greater detail shortly. Furthermore, a scatter plot of the entire population approaches illegibility as the cardinality of the population increases.

![Scatter Plot of Convergence](image)

Figure 6. A scatter plot depiction of population convergence. In contrast with the line graph depiction of convergence from Figure 5, the scatter plot does convey some information about the distribution of the population at each generation.
Although line and scatter plots are sufficient (and useful) for depicting the convergence of phenotypic variables or objective values, researchers have developed techniques specific to evolutionary algorithms that convey more relevant information with less data distortion. A complete survey of the other data visualization approaches employed by evolutionary algorithm researchers would range considerably, from highly specialized novel techniques, such as the evolutionary activity wave and distribution diagrams (Bedau and Brown 1999; Bedau, Joshi, and Lillie 1999) for the depiction (as in Figure 7) of evolutionarily significant events, to adapted cross-disciplinary techniques, such as the Hinton diagrammatic depiction of a weight matrix (Hinton and Sejnowski 1986). The latter was modified by Routen and Collins (1993) to simultaneously represent (as in Figure 7) both the distribution of a population and the fitness of its constituents. An analysis of these approaches exceeds the scope of this review, but they are included to demonstrate the diverse set of visualization approaches used for evolutionary algorithms.

Figure 7. An evolutionary activity wave diagram is used to depict events of perceived evolutionary significance.
Figure 8. A modified Hinton diagram for depicting the distribution of the values observed at each index of the genotypic representations of each candidate solution in the population.

The merits and weaknesses of each technique could be discussed individually, but it is important to note that many of these approaches were designed to aid practitioners in the application of evolutionary algorithms, and were not intended for use by researchers seeking understand the evolutionary mechanism. However, the technique known as the adaptive landscape (Wright 1932) distinguishes itself by its dual role as both a useful conceptualization, furthering an understanding of evolutionary dynamics (Ruse 1995), and as a surface plot depiction of a candidate solution space that lends itself well to both animate and inanimate depictions of a population of genotypes.
The adaptive landscape visualization technique (Wright 1932) conceived of a space wherein the structure of an organism could be plotted against that same organism's capacity for adaptation to the environment. This approach provides powerful intuition about the underlying evolutionary mechanism, and it has been suggested that the ease with which adaptive landscapes can be interpreted facilitates comprehension of the commonly used optimization metaphors (such as peak or local optima), inspires the design and enhancement of algorithms, and helps motivate more rigorous mathematical analyses (Provine 1989; Jones 1995a).

Stated formally in the context of the simple genetic algorithm, an adaptive landscape is comprised of the set of all possible candidate solutions to the optimization problem in question, a notion of the distance that exists between two candidate solutions (or, equivalently, a notion of the neighbourhood of each candidate solution), and the evaluation function that assigns a fitness value to each candidate solution (Stadler 2002). The landscape technique can then be explained as the space formed by the set of possible solutions, arranged such that the distance between solutions is assigned according to the ease with which one solution can reach the other. An orthogonal axis (graded according to fitness value) is then added to the space, and since each candidate solution can be evaluated and assigned a real fitness value, the point on the landscape for an individual solution will be the point in the space defined by its genotype and elevated according to its fitness. If the space of chromosomes is expressed in two dimensions, this adaptive landscape is easy to represent diagrammatically as either a topographic map or as a three-dimensional surface (Wright 1932), as in Figures 9 and 10, respectively.
Figure 9. A two-dimensional, topographic depiction of an adaptive landscape associated with a two-dimensional Rastrigin's function.

Figure 10. A projection of the three-dimensional surface representation of the adaptive landscape depicted in Figure 9.
The population of solutions maintained by the simple genetic algorithm can then be visualized as a set of points occupying different locations on this surface and, as future generations appear and the population is replaced, the population will appear to be traveling across the fitness surface, searching for optimal values.

The depiction of the relationship between the evolutionary mechanism and the space of chromosomes as a topographic map or three-dimensional surface is closely connected with the practice of classifying fitness functions according to ruggedness (Stadler 2002). It is often considered the case that an adaptive landscape that contains many separate peaks is rugged, in contrast with flat landscapes where no genotype exhibits any significant relative superiority over any other. In fact, the entire mechanism of evolution has been described as the traversal of the landscape in search of the highest peaks (Van Belle and Ackley 2003). Furthermore, significant effort has been invested in developing fitness functions that are tunably rugged (Kauffman and Levin 1987; Hordijk 1996; Altenberg 1996; Whitley, Mathias, et al. 1996) by stochastically generating peaks in the adaptive landscape through parameterized, but random, epistatic interactions.

It should also be observed explicitly that it is not uncommon for researchers to confuse proximity along phenotypic dimensions with proximity along genotypic dimensions. As the distance between points in genotypic space is different from the distance between the same points after a transformation onto phenotypic space, this distance distortion will modify the perceived complexity of the problem. The widely known Hamming cliff problem (Krzanowski and Raper 2001; Whitley 1993; Goldberg and Deb 1991) is useful
for demonstrating this false insight, for although it might be intuitive to treat the unsigned
integer values of 127 and 128 as neighbours (and it would not be surprising to see them
depicted as adjacent on an adaptive landscape), eight bit binary representations of these
values would be the strings ■□□□□□□□ and □■■■■■■■, respectively. Since the binary
digit differs between these chromosomes at every locus, it is implied that they are, with
respect to the common operators employed by the genetic algorithm, maximally distant
(and certainly not adjacent). Although the Gray reflected binary code (Gray 1953) can be
used to address Hamming cliffs specifically, the infidelity of phenotypic distance from
genotypic distance (and vice versa) is not alleviated by the use of Gray encodings. This is
trivially demonstrated by the discrepancy between the phenotypic distance of 128 units
between unsigned integer values 0 and 128, and the genotypic distance of 1 unit between
the Gray encoded representations ■■■■■■■■ and □■■■■■■■, respectively. Although a
unary encoding scheme does not suffer from the preceding restrictions (as the binary
strings ■□□□□□□□ and □■■■■■■■ would actually correspond to unsigned integers 7
and 1, respectively), the unary encoding is generally avoided in genetic algorithm
implementations because of its natural bias and redundancy issues (Rothlauf 2002).

Another challenge associated with the use of this visualization is the projection of the
space of possible candidate solutions onto two dimensions. Although there are many
researchers who openly acknowledge the inherent difficulty in attempting such an
operation (Wright 1932; Provine 1989; Culberson 1995; Jones 1995a, 1995b; Gitchoff
and Wagner 1996; Stadler 2002), there are many others who still acquire false intuitions
from visualizations for which the projection has been done poorly.
Furthermore, although it has been proposed that configuration of genotypic space is most appropriately represented using a binary hypercube (Goldberg 1989; Tanaka 2004), since neither of the variational operators typically employed by the simple genetic algorithm flips exactly one random bit per genotype per generation, this space is most definitely not a binary hypercube (Jones 1995a, 1995b). The operators typically employed result in a search space that is dramatically different from the binary hypercube (Gitchoff and Wagner 1996) and thus the immediate neighbourhood of a population member cannot be deduced from the binary hypercube.

2.2.3 Spatial Dimensionality Considerations

The difficulty associated with the accurate depiction of the underlying space from which the adaptive landscape surface is extruded is a consequence of the high dimensionality typical of genotypic space. For example, an optimization problem with a solution that takes the form of an ordered pair might require the genetic algorithm implementation to employ binary genotypes that are 64 bits in length. This entails that the genotypic space (to which the evolutionary mechanism of the genetic algorithm is applied) is actually 64 dimensional. Fortunately, a number of generic visualization approaches exist to address high dimensionality data considerations.

Although the image produced by a data visualization technique is typically two-dimensional (suitable for display on either screen or page), Tufte observed (1983) that graphics "visually display measured quantities by means of the combined use of points,
lines, a coordinate system, numbers, symbols, words, shading, and color." Although the conveyance of information by numbers, symbols, or words is not in question, North (2006) explicitly specified that the complexity requirements of an insight entails that it be comprised of a large, synergistic fraction of the data set. It can be concluded, then, that a worded, numeric, or symbolic indicator of a maximum value (for example) does not constitute an insight into the data. Consequently, an exploratory image, free of all textual data not used in co-ordinate representation, is comprised of a two-dimensional arrangement of points and lines depicted according to a colouring or shading scheme. Although colour is often used only to differentiate multiple graphs that have been plotted on the same axis, it can, naturally, be used to represent an additional data dimension.

The visible spectrum ranges from 400 to 700nm (Sharma 2004), the colours violet and red respectively. It is estimated that the human eye is adaptable enough to see more than ten million separate colours, but, for any one instant, the visual system can see only about ten thousand (Larson 1998). Although a dimensional resolution of ten thousand might seem to imply that the use of colour would be an effective way to increase the dimensionality of an image, there are a number of shortcomings that must be considered. Firstly, it is estimated (Cummings 2006) that about 8% of the male population suffers from the most common form of colour blindness. Thus, it could be inferred that a visualization relying heavily on colour will present decreased information to almost one in ten users. Additionally, although the human visual system is able to see ten thousand colours at a given moment, there is no definitive answer for how many colours an average person could differentiate between. Furthermore, the ability to identify a colour
correctly is affected by the other colours by which it is surrounded - a fact that can be demonstrated by the optical illusion in Figure 11. Although this image appears to employ four distinct colours, only the colours white, magenta, and green are present.

![Image](image.png)

**Figure 11.** The magenta square optical illusion, typically (and incorrectly) perceived as using two shades of magenta. In fact, only a single shade of magenta is used.

Finally, it is also noteworthy that despite the total ordering implied by the visual spectrum (i.e. red < orange < yellow < green < blue < violet), human perception does not readily impose an ordering to the colours perceived (Tufte 1983). Fortunately, as the various degrees of shading associated with a single colour (typically grey) do have an implicit perceived ordering, grayscale imagery can address many of the shortcomings associated with colour imagery while effectively functioning as an additional data dimension that can be used by a visualization approach.
Naturally, as data set dimensionality increases, the maximum effectiveness of a two-dimensional depiction (whether chromatic or achromatic) decreases, for both presentation and exploratory images. Motivated by the overabundance of data with which researchers are currently confronted (Wijk 2005), researchers have introduced a multitude of approaches to increase the dimensionality of an image.

It has been empirically demonstrated (Lawrence, Badre, and Stasko 1994) that interaction with animated imagery confers an advantage to student comprehension. It follows, logically, that researchers seeking to improve their own comprehension of a data set (and, ultimately, garner insight about the data) would also benefit from animate visualizations. H.G. Wells is the author credited with characterizing temporality as a fourth dimension, by having his protagonist of "The Time Machine" (1895) state that "... any real body must have extension in four directions: it must have Length, Breadth, Thickness, and Duration ... four dimensions, three which we call the three planes of Space, and a fourth, Time." It should be recognized that, excluding the presence of a holographic or volumetric display (Grossman and Wigdor 2010), only two spatial dimensions can be depicted in any single frame of an animated image. This establishes the maximum dimensionality of an animated image to be three dimensions.

It should also be explicitly noted that (during the depiction of an animated image) only those data points with temporal co-ordinates preceding the temporal co-ordinate of the current frame of the animation will have been observed by the user. Thus, to avoid distorting the data and misleading the user, the temporal co-ordinate of each data point
should be independent of any value that is depicted solely in a later frame. Consequently, a temporal dimension is seldom employed for depicting data sets that do not have a temporal component themselves. Since the genotypes that are used by the genetic algorithm to represent the difference candidate solution configurations are often considered with respect to the generation or generations in which they appear in the population, animated images are suitable for use in genetic algorithm visualization (Jackson and Fovargue 1997).

Further exploiting the ability of the human visual system to differentiate and group images by similarity (as noted previously when the use of colour to increase visualization dimensionality was discussed), some researchers replace individual data points with more complex images, referred to as glyphs. In this context, a glyph is defined as a small image wherein the shape or appearance of the glyph is entirely determined by one or more dimensions of the data point. Figure 12 includes depictions of the most commonly used glyphs. Although both triangular and weathervane-like glyphs (Pickett and White 1966, Cleveland and Kleiner 1974, Fienberg 1979) have been utilized to increase dimensionality, the most frequently employed glyph is a k-sided polygon (Siegel, Goldwyn, and Friedman 1971) or star (Mazza 2009), with a collection of these glyphs often referred to as a star plot. For a data set of a certain dimensionality n, depicted achromatically at a nonarbitrary position in a two-dimensional image, each data point is represented as a star with n-2 points. The value for each of the n-2 dimensions that are not depicted by the position of the glyph will directly determine the length of each point of the star.
Figure 12. Ten arbitrarily positioned glyphs (of the star type), associated with data points from a space of dimensionality five. If the position of each glyph was not assigned arbitrarily, the depicted data dimensionality would be increased to seven.

The preceding technique may surpass the use of colour or shading for maximizing dimensionality, but it is also understood that an increase in the number of points on the star decreases the probability that a user will recognize different values. Although glyphs are easy to differentiate, it becomes very difficult to extract data values from the dimensions depicted using the points of the star. Consequently, the area of application for glyphs can be somewhat limited.

It should also be mentioned, for the sake of inclusion, that the glyph type known as a Chernoff face (Chernoff 1973) is a possible (albeit bizarre) alternative to the star glyph. As humans are particularly adept (Mazza 2009) at recognizing other humans, a case can be made for the used of abstracted facial feature representations over stars, especially for situations that require a user to differentiate data points. Unfortunately, though this technique results in an increase in differentiability, it becomes virtually impossible for a user to extract any approximate data values for any of the higher dimension variables.
The technique, depicted in Figure 13, is presented here only for the sake of adequately exhausting the set of existing techniques that are used to increase the dimensionality of a data visualization.

Figure 13. A set of Chernoff faces generated using the same data points as depicted in Figure 12. Although it is exceedingly difficult to determine what values are actually being depicted, it is somewhat easier to detect differences when Chernoff faces are used instead of star plots.

Aside from altering the manner in which each data point is depicted, the dimensionality of a visualization can be increased by adjusting the locations of each data point. For example, the parallel co-ordinate visualization technique (Wegman 1990) uses a sequence of parallel, one-dimensional axes such that the $i^{th}$ component of each data point is plotted on the $i^{th}$ axis, and would be connected to the $i-1^{th}$ and $i+1^{th}$ component depictions on the two adjacent axes. Such an approach (as depicted in Figure 14) is closely related to the k-sided polygonal glyphs referenced earlier (Siegel, Goldwyn, Friedman 1971), where the magnitude of each point of the star is similarly determined by the $n^{th}$ component of each data point.
Figure 14. The parallel co-ordinate visualization technique was used here to depict data points from a set of dimensionality four.

A similar alternative, known as the Andrews curve (Andrews 1972), maps data points to the equation:

\[ f(x(t)) = x_1 / \sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(t) + x_5 \cos(2t) + ..., \]

where \( x_i \) corresponds to the \( i^{th} \) component of the data point. Unfortunately, as it has been recommended (Andrews 1972) that no more than 10 data points be plotted thusly on the same image, this approach has limited application for evolutionary algorithm visualization. Nevertheless, this projection of higher dimensionality data onto a two-dimensional curve (as depicted in Figure 15) does reflect the clustering of data points in the higher dimensional space, by using the proximity of the depicted curves.
Figure 15. The same data points from Figure 14 were depicted here using the Andrew's curve approach.

As an alternative to increasing the number of visualizable dimensions associated with an exploratory image, many researchers have proposed more efficient and representative use of the two available dimensions. These approaches differ from those described previously in that, rather than attempting to increase the number of dimensions, the fidelity of the lower dimensionality representation to the higher dimensionality data set is increased instead (by minimizing the discrepancy between the representation and the original data). The projection of data set of high dimensionality to a representation of lower dimensionality is, in fact, already performed implicitly during the construction of an adaptive landscape (albeit often incorrectly). The mapping (Rothlauf 2002) of genotypic space onto phenotypic space (and, if necessary, the subsequent projection down to two dimensions) does constitute an approach to addressing data sets of high dimensionality.
Although simplistic, the approach can be enhanced through the application of multidimensional scaling approaches.

It is the objective of a multidimensional scaling technique to construct a lower dimensional representation (often a two-dimensional image) of a higher dimensional space such that the distance between every pair of points depicted in the image is as representative as possible of the distance between the corresponding pair of data points in the higher dimensional space (Cox and Cox 2001). Although there are many effective multidimensional scaling approaches (Pohlheim 2006; Cox and Cox 2001, Ashlock 2005), the nonlinear mapping technique introduced by Sammon (1969) was selected for the body of research described in this document because of its iterative nature and its previous (albeit limited) application to evolutionary and genetic algorithm data visualization (Pohlheim 2006; Dybowski, Collins, and Weller 1996). It should be noted that an alternative approach to generating low-dimensional fitness landscape representations of high-dimensional genotypic spaces was introduced very recently (McCandlish 2011), using the eigenvectors of a transition matrix associated with the evolutionary process to construct the low dimensionality representation. This constitutes a dramatically different approach to representative depiction of genotypic space, but it must also be noted that this approach does not consider the effect that recombination has on the traversal of genotypic space by a population. Instead, this approach makes the assumption that the perceived difficulty by which one genotype can be transformed into another is dependent upon the fitnesses of the intermediate genotypes on the traversal of genotypic space. Naturally this entails prior exhaustive knowledge of the fitness values
that would be assigned to each genotype, and although the approach proposed by McCandlish is encouraging it is emphasized that the approach to genotypic spatial distance measurement that will be described in this dissertation is fitness function independent and does include consideration for the effect of both the mutation and recombination operators.

The distinction between a nonlinear mapping of the search space and a straightforward projection (by an arbitrarily chosen vector) is the series of iterative adjustments made by the nonlinear mapping technique. Although there is virtually always data loss whenever points from a high dimensionality space are mapped onto a lower dimensionality space, every iteration of the nonlinear mapping technique introduced by Sammon repositions each projected data point such that data distortion is decreased. This repositioning is performed using a gradient descent approach to the minimization of the stress between the pairwise distance matrices measured between points in the high dimensional space and in the low dimensional representation, respectively. In so doing, this technique attempts to ensure that the pairwise distance between all data points in the representation is maintained (as closely as possible) with the pairwise distance between the original data points.

With no upper bound on the number of conveyable dimensions, nonlinear mapping can be used to enhance the adaptive landscape visualization technique for the genetic algorithm even if the most effective metric cannot be defined (Pohlheim 1999). Dybowski, Collins, and Weller (1996) successfully applied Sammon's nonlinear mapping
to the development of a simple genetic algorithm visualization technique using the Euclidean distance metric, but noted that the computational complexity was prohibitive and the nonuniform errors present in the final visualization could be misleading to the user.

2.3 Typical Approaches to Distance Measurement

Although it is, of course, possible to limit analyses of the evolutionary mechanism of the genetic algorithm to those instances wherein the chromosomes are short enough to be depicted, without ambiguity, using the aforementioned techniques (Wright 1932; Culberson 1995; Jones 1995a, 1995b), this restriction is impractical if the researcher is seeking insight on the evolutionary mechanism in general. Fortunately, as it was noted previously, Sammon’s nonlinear mapping can effectively scale a space of arbitrarily high dimensionality down to a two-dimensional space, provided that there exists a measure of distance (that need not satisfy the subadditivity condition) (Pohlheim 1999) that is both accurate and representative. This entails that the surface of a superior adaptive landscape visualization could only be extruded from a two-dimensional arrangement generated by a nonlinear mapping of genotypic space if the measure of interchromosomal distance in the higher dimensional space is both accurate and representative. Accordingly, this motivates an investigation of the measures that have previously (and successfully) been applied to the distances between strings and, more specifically, to the distances between genotypes in other evolutionary algorithms.
2.3.1 Distance Measurement Between Strings

As the candidate solution representations employed by the genetic algorithm are typically binary strings, the effectiveness of the measures employed during error correction and detection (for computing the distance between binary signals) must be investigated. The Hamming distance (Hamming 1950), already encountered frequently in the field of evolutionary computation (Jacob 2001), is computed between two binary strings of the same length as the number of indices for which the value of one string differs from the other. Equivalently, the Hamming distance between strings is the number of substitution operations (which are, for binary strings, bit flip operations) necessary to transform either string into the other. This measure was introduced to ascertain the amount of damage inflicted on a signal (Hamming 1950), by comparing the minimum distance between code points before and after transmission. Since transmission errors are unbiased across strings and can be modeled with a uniform probability distribution, this measure is appropriate for tasks in the signal processing domain. That being said, it can only be considered representative of the distance traversed within a hypercubic space. As it was noted previously that the genotypic space traversed by the genetic algorithm is not hypercubic, a more complex measure should be employed to ensure that the measured distances are representative of the actual genotypic spatial distances traversed by the genetic operators.

Record linkage applications, attempting to assess the degree to which two entities could be considered counterparts, also employ measures of the distance between strings that could be modified to measure genotypic spatial distances. The Levenshtein and
Damerau–Levenshtein distances (Damerau 1964; Levenshtein 1966) could both be considered extensions of the Hamming distance. These measures compute the distance between strings as the number of edit operations necessary to transform either string into the other, where the sets of edit operations are \{insertion, deletion, substitution\} and \{insertion, deletion, substitution, transposition\}, respectively.

Alternatively, the Jaro-Winkler and Jaro distances (Jaro 1989, 1995; Winkler 1999) between strings are computed as a function of the number of matched or transposed characters between strings, with and without preferential weighting assigned to characters near the front of the string, respectively. It must be emphasized that these techniques were not intended specifically for application to binary alphabets, and although these distances include additional consideration for the alternative operations by which a space might be traversed, they remain unrepresentative of the variational operations employed by the genetic algorithm.

### 2.3.2 Population Diversity Measurement

It should also be noted that there is a significant overlap between approaches to the measurement of the distance between strings and many population diversity assessment techniques developed for use by biologists and ecologists. Furthermore, as many approaches to the measurement of population diversity in genotypic space can be easily defined as some aggregate of the distance measured between each pair of genotypes (Wineberg and Oppacher 2003), the various approaches to population diversity
estimation are both a source for, and application of, distance measurement techniques. Although it is widely recognized that the use of a proportional selection strategy results in a trend of decreasing population diversity over time (Goldberg 1989), it is generally accepted that a nontrivial degree of population diversity must be maintained if the algorithm is to function properly (Oppacher and Wineberg 1999).

One of the most commonly employed (Wineberg and Oppacher 2003) measures of population diversity employed by ecologists is known as Shannon entropy. Although not based on a pairwise measurement approach, this measure was developed to quantify the volume or rate at which information is produced by an information source (Shannon 1948). Equivalently, the calculation estimates the number of binary digits necessary for the encoding of a particular piece of information – an estimate of the information content of what is, essentially, a random variable. With a single bit both necessary and sufficient to distinguish between two elements, two bits both necessary and sufficient to distinguish between four elements, and, in general, n bits both necessary and sufficient to distinguish between $2^n$ elements, it is easy to recognize that the amount of information contained by a set $S$ can be quantified as $\log_2(S)$. It follows that if the probabilities associated with characters at specific indices are defined by frequency values, the expected entropy associated with an index is the sum of the products of each probability $p$ and the number of binary digits necessary to encode it, $\log(p)$, for a total entropy $H = - \sum p \log(p)$.

The application of this measure to population diversity assessment can be demonstrated easily using the population of strings {'aaa', 'bba', 'caa', 'dba'} from the quaternary alphabet {'a', 'b', 'c', 'd'}. At the first index, each of the four strings uses a unique
character, and the entropy calculation at this index would yield the (maximum) value of two. At the second index, two of the four strings use the character 'a' and the other two use the character 'b', resulting in an entropy value of 1. Finally, at the third index, each of the strings uses the same character and the entropy is calculated as 0. Naturally, the maximum amount of information is conveyed by the character in first index, and no information in conveyed by the character in the third index – equivalently, the set of values used at the first index has the maximum possible diversity, and the set of values used in the third index has the minimum possible diversity.

2.3.3 Measures Used in Evolutionary Computation

Contrary to the field of ecology, Morrison and De Jong (2002) observed that, in the evolutionary computation domain, the most common measure of population diversity is the pairwise computation of the Hamming distance for the elements of the population. Although he and other researchers (Rényi 1961, Simpson 1949) have introduced additional measures of population diversity, suitable for application to the populations employed by evolutionary algorithms, Wineberg and Oppacher (2003) demonstrated that many approaches to measuring the diversity of populations of binary strings (including entropy, variance, expected heterozygocity, and nucleotide diversity) are only minor variations of the simple pairwise sum of the individual Hamming distances between entities. Thus, to effectively measure the diversity of a simulated population, it is sufficient to accurately and representatively measure the distance between the individual
candidate solution representations, and then use this measure between all-possible-pairs to assess the diversity of the complete population.

2.4 Genetic Algorithm Operator Considerations

It was noted (O'Reilly 1997) that distance measurement techniques can be applied, not only as approaches to the supervision of population diversity, but as a means by which the effect of a genetic operator can be quantified and subsequently evaluated. For the genetic programming subdomain of evolutionary computation, techniques by which the distances between candidate solution programs are measured often include consideration for the operators that will be applied. This consideration clearly defines a notion of the edit distance that separate candidate programs. Not surprisingly, there is a significant intersection between these edit operations (O'Reilly 1997) and those operations considered by the record linkage techniques mentioned previously (i.e. the substitution, insertion, and deletion operations). It has been explicitly recognized that these measures of edit distance consider the actual operations applied to the population, but it should be noted that, in virtually all cases, the only operations being considered are of unary arity.

It is difficult to conceptualize a measurement of the distance between individual chromosomes with respect to the traversal of genotypic space by a binary operator. This difficulty can be clarified with an example from the space of integers. Although it is easy to conceptualize the distance between the values of +1 and +8 in terms of an increment operator (i.e. \( x = \text{increment}(x) = x + 1 \)), the distance between these values in terms of a
binary operator (such as addition) is undefined. Equivalently, although it can be stated that +1 is separated from +8 by exactly seven increment operations, it is impossible to state how many binary addition operations are necessary without knowing what other operands are available. This restriction will be addressed in extensive detail in Chapter 3.

In spite of this difficulty, Altenberg's (1997) development of an evaluation function for his fitness distance correlation counterexample did introduce a measure of "crossover distance" between chromosomal pairings, defined as the number of single point recombination operations that must be applied to transform one pair of complementary chromosomes into another complementary pair. This technique is not entirely unlike the recombinational distance measure introduced by Manzoni, Vanneschi, and Mauri (2012), which was also defined in terms of one-point recombination. Techniques have also been introduced (Gustafson and Vanneschi 2005, 2008; Vanneschi, Gustafson, and Mauri 2006) to the genetic programming subdomain for the measurement of the distance between programs that reflect the operation of 'crossing-over' program subtrees between candidate solutions. Unfortunately, in each of these cases the distance is being measured from an element of the space of possible populations and not the space of candidate solutions. Similarly, it was noted by both Jones (1995a, 1995b) and Culberson (1995) that considering each point in any genotypic space to be a single chromosome does not permit researchers to explicitly connect points through a recombination operator. In their respective attempts to model the effect of recombination, it was proposed that points in the search space should be representative of possible chromosomal pairings, between
which connections exist when one pair could be recombined to produce the elements of the other pair as offspring.

An alternative model investigated by Gitchoff and Wagner (1996) proposed a hypergraphic depiction of the recombination operator, wherein exactly one vertex exists for each possible candidate solution configuration, with as many hyperedges between any two vertices as there are possible offspring chromosomes that could be the result of a recombination operation between the two hyperconnected vertices. To clarify, two candidate solution configurations that could be recombined (each depicted using a single vertex) are connected by one hyperedge for each of the different offspring that could be produced. Since these hyperedges always have exactly two endpoints (one for each parent chromosome), the offspring represented by the hyperedge is always the result of a binary function (recombination) on two operands (the parents). As a result, it is not possible for this depiction to represent the distance between a single parent chromosome and a possible offspring, and, consequently, traversals of genotypic space by the binary recombination operator cannot be represented using this approach.

Unlike any standard unary mutational operator, the current level of diversity in the population of the genetic algorithm has a dramatic impact on the effect of the binary recombination operator. Ignoring the recognized bias in odd-point crossover operations (Rana 1999), the binary recombination of two strings between which there is a maximal Hamming distance is actually capable of producing an offspring chromosome of any structure, belonging to any schemata. Contrarily, the application of the recombination
operator to a pair of identical strings is incapable of generating offspring that are any different from the parent structure. Stated simply, the magnitude of the effect of recombination varies directly with the diversity of the population to which it is applied. Although it is widely recognized (for instances of the simple genetic algorithm) that the use of a proportional selection strategy results in a trend of decreasing population diversity over time (Goldberg 1989), this does not entail that binary recombination becomes a less useful operator over time. In fact, assuming the selection scheme employed by the algorithm is operating properly, it is this property that allows the repeated application of binary recombination to focus the original population into a population of reduced diversity but relatively superior fitness (Culberson 1995). Similarly, the investigation of a modified genetic algorithm for application to optimal control problems (Michalewicz, Janikow, and Krawczyk 1992) indicated that binary recombination was essential for the early, nonpremature convergence of the population in the search space. These authors also observed that the significance of certain genetic operators does vary over time.

It should be noted that the work of Moraglio and Poli (2004) does establish an operator (referred to as the topological crossover operator) that can be derived using a pre-existing definition of the neighbourhood associated with a genotype. It should be explicitly noted that the definition of this operator specifies that it is still necessary to perform a binary operation that cannot be defined without explicit consideration of the population. It should also be noted that a genetic algorithm that is being used to generate a data visualization has already established its variational operators – it is the neighbourhood
definition that must be defined and extended to develop a measure of genotypic spatial distance.

Further addressing the absence of a suitable measure of the distance between individual binary genotypes (that does include consideration for the recombination operation), it must be recognized that the effectiveness of binary recombination is not universally accepted by the researchers exploring evolutionary computation. There have been many analyses (Fogel and Atmar 1990; Fogel 1993; Fogel and Stayton 1994; Schaffer and Eschelman 1991; Schaffer, Caruana, et al. 1989) that have compared the performance of a genetic algorithm that does employ a recombination operator against a genetic algorithm or evolutionary system that does not employ any recombination operator. In these analyses, when the genetic algorithm that does use recombination is outperformed by the genetic algorithm that does not, it is interpreted as evidence that the usefulness of the binary recombination operators is dubious. However, it can be observed that these results alone cannot be used to conclusively justify declarations that recombination is ineffective, as these investigations were performed on a relatively small set of problems.

In spite of the existence of this transformation, even among those researchers of the opinion that binary recombination is both useful and powerful, there is often a tendency to avoid any description of the impact such an operator has on the movement of population members within a search space. The difference between mutational and recombinational operators has been compared to moving through a search space with either a small or large step size, respectively (Fogel and Stayton, 1994). As Jones (1995a)
observes, describing the effect of a crossover operation as a large jump is both inaccurate and insufficient. Instead, he suggested that each operator in use should be conceptualized as acting upon a different landscape. Nevertheless, to accurately and representatively measure the distance between the binary string candidate solution representations used by the genetic algorithm, consideration must be included for both unary mutational operations and binary recombinational operations. Operating under the assumption that the creation of such a measure is actually achievable, it is also necessary to introduce a methodology whereby the accuracy of the measure could be evaluated, before any responsible application of this measure could be attempted.

2.5 Distance Measurement Approach Evaluation

As an analogous hypothetical scenario, if a surveyor was provided two unique rulers or measuring rods and asked to evaluate which of the two was more accurate, it would be necessary for the surveyor to have some third distance measuring apparatus (or set of known distances) against which the rulers could be calibrated. In the absence of an available measure with which to calibrate, the surveyor could only indicate whether or not there was a detectable difference between the two unique rulers or measuring rods. Similarly, a researcher tasked to contrast two or more distance measures and select the most suitable approach for a given application would require known distances against which the accuracy of the possible measurements could be gauged and, in this manner, evaluated. Wijk's model does include some economic criteria by which an approach could be evaluated (Wijk 2005), but just as the Manhattan distance is more suitable for
distance measurement in a city than the Euclidean distance, so to is the representativity (and, thus, effectiveness) of a measure maximized if the manner by which the genotypic space is to be traversed is reflected in the measurement.

Unfortunately, as the genetic algorithm typically exploits a complex genotypic space and offers only a stochastic approach to optimization, it is exceedingly difficult to compute a precise value for the actual distance between any two chromosomes from genotypic space. However, outside of the discipline of computer science there are a number of approaches to the measurement of proximity that do not simplistically measure Euclidean distance. These approaches are discussed in the following section and did ultimately serve as the inspiration for the representative empirical measure introduced in Chapter 5 for the tasks of calibration and evaluation.

2.5.1 Extradisciplinary Inspiration

Urban analysts often conceptualize the distance between buildings as a function of time or accessibility (Páez and Scott 2005), because the approximate velocity of traffic on a street can be easily estimated or empirically measured. Similarly, as the speed of light in a vacuum is a fixed, absolute value (Dukish 2009), astronomers conceptualize the proximity of distant celestial bodies in terms of units known as light years (Seeds and Backman 2009). Although the light year unit is a measurement of distance, it is the constancy of light speed that allows the proximity to be expressed in temporal units. Although the speed of convergence of a genetic algorithm is objective function dependent
(Fogel 1995) and certainly not constant, it should be emphasized that the expected number of applications of the variational (mutation and recombination) operators per generation is parameterized and, as such, the evolutionary mechanism does operate at an estimable speed.

Although a reasonable estimate of the speed of the evolutionary mechanism can be computed without knowledge of the selection strategy component of the evolutionary mechanism (which, as noted previously, is objective function dependent), it is not possible to generalize about the direction along which the evolutionary mechanism will proceed in the higher dimensional space. However, paralleling a two-dimensional geometric construction using a drafting compass to construct the bounding (circular) subplane associated with the fixed compass radius, the genetic algorithm itself can be used to identify the subspace of genotypes that can be reached from any specified genotype. The act of turning the compass in the preceding geometric construction analogy is actually an exhaustive (and continuous) exploration of all possible directions along which a transformation could proceed; similarly, a representative sample from the set of all possible variational operator applications used by the genetic algorithm must be explored for the approach to be effective, and the next chapter will begin exploring the traversal of genotypic space by these operators.
Chapter 3

Operator Traversal Representativity

Overview

It is the aim of this chapter to investigate the traversals of genotypic space that occur when either the mutation operator or the recombination operator (but not both) is used to transform one genotype into another. It will be demonstrated that the former (i.e. traversal using only the mutation operator) is well represented by conventional measures (such as the Hamming distance), but the latter (i.e. traversal using only the recombination operator) is not. This fact is well known and can be demonstrated trivially - for a population that is comprised entirely of copies of a single parent genotype, it is impossible to produce new genotypes through the application of recombination alone. This would entail that the recombinational distance (i.e. the number of recombination operations necessary) to produce a genotype \( \alpha \) from a genotype \( \beta \) (where \( \beta \neq \alpha \)) is infinite, regardless of the Hamming distance between \( \alpha \) and \( \beta \). Consequently, the
Hamming distance measure is not representative of the genotypic spatial distance that would be traversed by the recombination operator, and since a representative measure of the distance traversed by recombination is necessary to construct a measure that is representative of traversals by the complete explorative component (i.e. both mutation and recombination), a comprehensive investigation of recombinational distance measurement is the focus of this chapter.

This chapter refers to the traversal of genotypic space by the recombination operator as though the uniform recombination operator was the only recombination operation in use by the genetic algorithm. Since the uniform recombination operator is an operator with binary arity (in contrast with the unary arity of the typical mutation operator used by the simple genetic algorithm), the task of combining distinct measures of mutational and recombination distance is complicated. Furthermore, the binary arity of the recombination operator indicates that the transformation (through genotypic space, from a parent to an offspring) defined by the recombination operator is not independent of the current population. Mutation, on the other hand, can be used to define a transformation that is independent of the current population. In addressing both these complications, this chapter will discuss an approach where a single binary recombination operator is treated as though it were a set of unary operators defined by the unique members of the population.

Although it is conceivably possible to restrict this set of unary operators to only those operations associated with a reduced subset of the unique population members, this
would entail that several potential traversals of genotypic space will not be considered. Although such a restriction could be included in a future application (at the discretion of the user), it would be inappropriate at this stage of analysis. It is also important to recognize that the most logical criteria for selecting the subset of the unique population members from which the function of the recombination operation is defined would be either the available evaluation function data or the perceived distance from a fixed point in genotypic space. Since it has been established that genotypic spatial distances cannot be representatively measured without considered the traversals associated with recombination, the only conceivable avenue for restricting the set of unary operators defining the recombination operation would be the use of evaluation function data. In discouraging this practice, the user is reminded that the adaptive landscape visualization technique already employs evaluation function data in the extrusion of the three dimensional surface from the two-dimensional representation of genotypic space – the orthogonality of the evaluation function dimension to the dimensions in which genotypic space is depicted reinforces that adaptive landscape visualization surfaces are extruded from an accurate representation of genotypic space being traversed that is, itself, independent of the evaluation function being optimized. Furthermore, this chapter will demonstrate that the time complexity required for measuring the distance (between genotypes) that is associated with a specific variational operator is substantially less than what might be expected, and that the use of more representative distance measures for tasks such as data visualization is not precluded by the required increase in time complexity.
Under the unary recombination paradigm proposed in this chapter, the recombinational distance between two genotypes can be defined in one of two ways, denoted $\Delta_\chi$ and $\Delta_{\xi\chi}$. Where $\Delta_\chi$ represents the number of times one genotype must be recombined with the population member denoted $\iota$ to produce some other genotype (a quasimetric definition of recombinational distance), $\Delta_{\xi\chi}$ represents the number of times one genotype must be recombined with any member of the current population to produce some other genotype (a semiquasimetric recombinational distance). Both of these proposed measures of genotypic spatial distance are described in detail and the latter ($\Delta_{\xi\chi}$) is used as the initial measure of recombinational distance.

Although there have been several previous attempts to depict recombination graphically, this chapter also demonstrates that none of these visualization methodologies would convey recombinational distance in a way that could be adapted for use in constructing adaptive landscape visualizations. As a result, it was necessary to introduce a new approach to the graphical representation of recombination, based on the unary recombination paradigm. The introduction of this graphical depiction was very useful for exploring the perceived complexity of measuring the recombinational distance that separates candidate solution representations in genotypic space.

By this point, the recombinational distance value that could be assigned to the traversal from one genotype to another (based on quasimetric $\Delta_{\xi\chi}$) could only be assigned a value of zero (if the two genotypes are the same), infinity (if recombination alone cannot produce the second genotype as a descendent of the first), or some other nonzero, finite
value (if recombination alone can be used to produce the second genotype as a
descendent of the first). Although this might have been sufficient for enhancement of the
adaptive landscape visualization technique, this chapter also explored an alternative
measure of recombinational distance measurement that uses different subpopulations.
Since it is also possible to use this approach to justify the assignment of distance values
ranging from one (the smallest possible subpopulation) to the minimum of the number of
unique populations members (the largest possible subpopulation) and the chromosome
length (the number of possible changes to a genotype), the two measures of
recombinational distance defined by this chapter are referred to as the low resolution and
high resolution recombinational distance measures, respectively.

The proposed algorithm for the precise measurement of recombinational distance will be
demonstrated to have a nonpolynomial worst-case time complexity, but moreover it will
be demonstrated that the task of measuring recombinational distance is essentially
equivalent to finding a solution for a set cover optimization problem – a task that is
known to belong to the class of NP-hard problems. This motivates a preliminary
investigation into the heuristic techniques that may be suitable for recombinational
distance measurement, including the details of a polynomial time greedy heuristic that,
while efficient, can be shown to overestimate the recombinational distance between
genotypes.

The development of these representative measures of recombinational distance proved to
be a very challenging task, and, consequently, the research presented in this chapter is
quite dense. This high level of detail was necessary to ensure that every relevant aspect of recombinational distance measurement was thoroughly considered, but many of the complex analyses that comprise this chapter are not be required for the understanding the integrated distance measures ultimately derived in Chapter 4 and comparatively evaluated in Chapter 6. Consequently, readers may wish to bypass much of the content of Chapter 3 in favour of the overview and the summary that immediately precedes Chapter 4. Alternatively, a condensed version of these analyses was presented at the International Conference on Evolutionary Computation in 2010 (Collier and Wineberg 2010) and may be consulted for a somewhat less complex treatment of the topic.

### 3.1 Mutational Distance Measurement

It was noted previously that, although the most typical approach to the measurement of genotypic spatial distances when using a binary representation is the Hamming distance, the Hamming distance between binary string genotypes assumes a hypercubic genotypic space, as depicted in Figure 16. Each genotype of length $\lambda$ is depicted as a vector in the $\lambda$-dimensional hypercube and is connected to every other genotype depiction that differs at only a single vector component (and is, thus, separated from the original genotype by a Hamming distance of 1). It follows, then, that a (hypothetical) variational operator that flips exactly one binary digit would transform any specific genotype into one of the other $\lambda$ genotypes to which it is connected. This would entail that the transformation of one genotype $\alpha$ to another genotype $\beta$ would require as many applications of the operator as the Hamming distance between $\alpha$ and $\beta$. 
Figure 16. The genotypic space associated with genotypes of length four, depicted with an underlying hypercubic structure.

Although this could certainly be conceptualized as a mutation operator, it must be emphasized that the mutation operator is not typically implemented as an operator that flips exactly one digit each generation. The far more common implementation of the mutation operator flips the binary value at each index of a chromosome both probabilistically (with parameterized probability $p_{\text{mutation}}$) and independently. This allows the operator to make as few as zero and as many as $\lambda$ mutations to a chromosome. With the expected number of individual bit mutations per application thus defined by the product $\lambda(p_{\text{mutation}})$, the $p_{\text{mutation}}$ probability parameter is often set a $\lambda^{-1}$ such that the expected number of individual bit mutations is one per genotype. Where the formerly described mutation operator flips exactly one binary value per genotype and, thus, can only mutate from one chromosome into one of its $\lambda$ hypercubic neighbours, the latterly
described mutation operator can technically mutate any genotype into any of the $2^\lambda$ chromosomes of the same length in a single application.

It must be noted, however, that the probability distribution for transformations between genotypes is certainly not uniform. In fact, any genotype subjected to mutation has a:

$$(1-p_{\text{mutation}})^\lambda$$ probability of remaining unchanged by mutation,

$$(p_{\text{mutation}}) \cdot (1-p_{\text{mutation}})^{\lambda-1}$$ probability of experiencing a single bit flip

$$(p_{\text{mutation}})^2 \cdot (1-p_{\text{mutation}})^{\lambda-2}$$ probability of experiencing two bit flips, etc.

For $\lambda = 3$, the space of chromosomes can be depicted as a fully connected graph, where each node represents a genotype and an edge between genotypes indicates that it is possible to mutate either genotype into the other. As this is a complete graph, it can be made more representative by adjusting the weight of each edge according to the relative likelihood of the corresponding transition, as in Figure 17.

Since chromosome $\alpha$ differs from chromosome $\beta$ at exactly as many indices as the Hamming distance $\delta$ between $\alpha$ and $\beta$, then of the $\lambda C_\delta$ ways in which $\delta$ bits can be flipped (and $\lambda-\delta$ bits remain unaffected), exactly one would effect the transformation from $\alpha$ into $\beta$. The transition probability defined distance between $\alpha$ and $\beta$ can then be calculated as:

$$(\lambda^{-1})^\delta \cdot (1-\lambda^{-1})^{\lambda-\delta}.$$
Figure 17. A graph of the transitions between genotypes that are possible using the mutation operator. Edge weights are determined by the relative likelihood of each transition.

This relationship between the Hamming distance $\delta$ between chromosomes and the distance between chromosomes that must be traversed by the typical mutation operator (for an arbitrary $p_{\text{mutation}}$ parameter) can be explored in greater detail. With the transition probability $p_{\text{transition}}$ between two chromosomes as the product of the probability of mutating the value at each index that differs between them and the probability of not mutating the value at each index that is shared, the expression reduces to exponential with respect to the Hamming distance between the chromosomes, as demonstrated on the following page.
\[ p_{\text{transition}} = p_{\text{mutation}}^\delta \cdot (1 - p_{\text{mutation}})^{1-\delta} \]

\[ = p_{\text{mutation}}^\delta \cdot (1 - p_{\text{mutation}})\delta \cdot (1 - p_{\text{mutation}})^{-\delta} \]

\[ = (1 - p_{\text{mutation}})^\lambda \cdot p_{\text{mutation}}^\delta \cdot (1 - p_{\text{mutation}})^{-\delta} \]

\[ = (1 - p_{\text{mutation}})^\lambda \cdot p_{\text{mutation}}^\delta \cdot \frac{1}{1 - p_{\text{mutation}}} \]

\[ = (1 - p_{\text{mutation}})^\lambda \cdot e^{\ln(p_{\text{mutation}}^\delta)} \cdot e^{\delta \ln\left(\frac{1}{1 - p_{\text{mutation}}}\right)} \]

\[ = (1 - p_{\text{mutation}})^\lambda \cdot e^{\delta \ln(p_{\text{mutation}}) + \ln\left(\frac{1}{1 - p_{\text{mutation}}}\right)} \]

\[ = (1 - p_{\text{mutation}})^\lambda \cdot e^{k_1 \cdot e^{k_2 \cdot \delta}} \]

It should be noted that the constant \( k_2 \) will be negative as long as \( p_{\text{mutation}} \) is less than 0.5, resulting in the transition probability being subject to exponential decay with increasing Hamming distance. Furthermore, as this is a monotonic function, the ordering of genotype according to the likelihood with which they will be the result of a mutation of any specific genotype is the same as the ordering of chromosomes according to their Hamming distance from that same genotype.

The precision of this calculation indicates only that it is possible to develop a measure of the distance between genotypes that is representative of the manner by which genotypic space is traversed by a unary mutation operator. Since it is also possible for the genetic algorithm to traverse genotypic space using the binary recombination operator, a recombinational distance measure is needed before a truly representative measure for the genotypic spatial distance traversed by the complete exploratory mechanism can be constructed. This chapter will explore the development of alternative distance measures.
that are more representative of the manner in which genotypic space is traversed by the recombination operator. These recombinational distance measures represent the next logical step in the development and validation of measures representative of traversals by the genetic algorithm.

### 3.2 Considerations Specific to Recombination

Binary recombination operators can be classified as either k-point recombination operators (a set comprised of operators for all values of k between one and the fixed length of the chromosomes of the representation) or uniform recombination operators. Although the recombination operator used in the simple genetic algorithm performed one-point recombination, the argument can be made that since the set of possible offspring of any k-point binary recombination operation is a subset of the set of possible offspring of a uniform recombination operation between the same parent chromosomes, then uniform recombination is more suitable as the general case recombination operator.

This is not to be interpreted as a criticism of the use of k-point recombination, and should only serve to specify an important consideration of the analyses presented herein. The declaration that uniform recombination is a more suitable general case recombination than any one form of k-point recombination is not unreasonable and, as such, all subsequent references to recombination, including the operator symbol $\chi$, refer to the binary uniform recombination operator unless otherwise specified.
The binary arity of the typical recombination operator, in contrast with the unary arity of the typical mutation operator, poses the most significant barrier to the introduction of a measure of the distance traversed by recombination. Since recombination requires more than a single operand, the notion of recombinaional distance in the candidate solution space is undefined if the composition of the population is unspecified. Consequently, when considering traversal of the search space using only recombination, the population must be explicitly considered.

This might seem to suggest, initially, that a measurement of the distance between individual chromosomes is not possible. It might also seem to suggest that a recombinaional distance measure can only measure the distance between two sets of chromosomes, of which at least the first set must have a cardinality greater than or equal to two (to satisfy the binary arity requirements of the recombination operator). This is emphasized by the aforementioned works of Jones (1995a, 1995b), Culberson (1994), and Altenberg (1996), wherein adjacencies defined with respect to traversal by a recombination operator were only defined between sets of chromosomes (typically of cardinality two). Although the development of a measure of recombinaional distance from a definition of adjacency is trivial, since adjacency was defined only from a set with a cardinality of two, a measure of the distance separating two individual chromosomes would not appear to be possible. In spite of this perception, the next section will demonstrate that the mapping induced by the binary operator can be replaced by the collective mappings induced by each member of a set of unary operators.
### 3.3 Unary Redefinition of Recombination

In order to conceptualize the redefinition of binary recombination to a set of operators of unary arity, an analogous scenario is examined, where binary addition operator + defined over the set $\mathbb{Z}$ of integers is replaced with a set of unary operators $\xi$, where, $\forall \alpha, t \in \mathbb{Z}, \exists +_t \in \xi: +_t(\alpha) = +(\alpha, t)$. In demonstrating this, consider the subset of $\mathbb{Z} \{1, 2, 3\}$ to which the binary addition operator can be applied to create the set of possible operations $\{1+1, 1+2, 1+3, 2+1, 2+2, 2+3, 3+1, 3+2, 3+3\}$. It should be obvious that it is possible to replace the binary addition operator with three unary operators – an increment by a value of one (denoted ++), an increment by a value of two (denoted +=2), and an increment by a value of three (denoted +=3), respectively. Upon the same subset of $\mathbb{Z}$ these operators define the set of operations $\{1++, 1+=2, 1+=3, 2++, 2+=2, 2+=3, 3++, 3+=2, 3+=3\}$ and it is clear that these operations are functionally equivalent to those that can be performed with the binary addition operator.

Similarly, for a simple genetic algorithm that is simulating population $P$ to search a solution space $S$, the function of the binary recombination operator $\chi: \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{S}$ can be replaced by the set of unary operators $\xi_\chi$ where:

$$\forall \alpha, t \in \mathbb{S}, \exists \chi_t \in \xi_\chi: \chi_t(\alpha) = \chi(\alpha, t).$$

In clarification, for a population of chromosomes with fixed length $\lambda = 2$, the function of a binary recombination operation $\chi(\alpha, t), \forall \alpha, t \in \{□□, □■, ■□, ■■\}$ can be replaced
with the set of operators $\xi_\chi = \{\chi_{\Box\Box}, \chi_{\Box\Diamond}, \chi_{\Diamond\Box}, \chi_{\Diamond\Diamond}\}$ where operations $\chi_{\Box\Box}(\alpha)$, $\chi_{\Box\Diamond}(\alpha)$, $\chi_{\Diamond\Box}(\alpha)$, and $\chi_{\Diamond\Diamond}(\alpha)$ are functionally equivalent to the operations $\chi(\alpha, \Box\Box)$, $\chi(\alpha, \Box\Diamond)$, $\chi(\alpha, \Diamond\Box)$, $\chi(\alpha, \Diamond\Diamond)$, respectively. Having established the membership to the set of unary recombination operators, it suffices to note that the application of a single binary recombination operation to a population is replaced by the probabilistic application of a member of the set of unary recombination operators, where the likelihood of performing a unary recombination operation corresponding to a chromosome that is not present in the current population is set to 0.

To fully specify the unary recombination operation it is first necessary to recognize that every unary recombination operator denoted $\chi_\iota$ that is contained in $\xi_\chi$ is essentially performing a binary recombination operation with the first operand fixed to chromosome $\iota$. It is thus possible to redefine the solution space $S$ with respect to this fixed chromosome and deduce general properties applicable to every unary recombination operator contained in $\xi_\chi$.

With the recombination operators of the genetic algorithm defined for chromosomes of fixed length $\lambda$, the set of possible chromosomes of the same length with which the fixed chromosome $\iota$ could be recombined is identified as the ordered set $S'$, of cardinality $2^\lambda$. Within the set $S'$ there are $C(\lambda, \delta)$ unique chromosomes at a Hamming distance of $\delta$ from $\iota$, for $0 \leq \delta \leq \lambda$. From the binomial theorem it is established that $\delta = 0^\lambda \sum^\lambda C(\lambda, \delta) = 2^\lambda$ and, consequently, the subsets of $S'$ associated with each possible Hamming distance value of $\delta$ where $0 \leq \delta \leq \lambda$, are mutually exclusive and exhaustive. Any chromosome $S'_i$ belonging
to ordered set $S'$ can be uniquely identified as the chromosome of length $\lambda$ that has values complementary to those of $\iota$ at the set of indices $\phi$, where the cardinality of set $\phi$ can range from 0 (for chromosome $S'_0$ at Hamming distance 0 from $\iota$) to $\lambda$ (for chromosome $S'_{\lambda}$, complementary to chromosome $\iota$, at a Hamming distance of $\lambda$). This is illustrated in Figure 18 for $\lambda = 3$.

**Figure 18.** For $\lambda = 3$, the set $S$ of possible genotypes contains exactly eight members. Each of these chromosomes can be defined by the complementary index set encoding a transformation from a fixed genotype $\iota$ to the chromosome. The set can then be ordered by the cardinality (and contents) of these complementary index sets.

It is stressed that any binary string chromosome $\iota$ of length $\lambda$ is suitable for defining a unary recombination operation, provided that the set of chromosomes $S'$ is the set of all binary strings of length $\lambda$, ordered such that $S'_0$ for $\phi = \{\}$ will be the binary string that is identical to $\iota$, having a Hamming distance of 0, $S'_1$ for $\phi = \{1\}$ will be the binary string that is identical to $\iota$ except at index 1 for which it will be complementary, having a Hamming distance of 1, etc. A further example for $\lambda = 3$ is depicted in Figure 19.
Figure 19. Although the set of possible genotypes \( S \) is not dependent on the fixed chromosome \( \iota \), the ordered set \( S' \) is sorted by complementary index set cardinality. Thus, a different fixed chromosome \( \iota \) results in a different ordering of the set \( S' \).

It is now possible to define a unary recombination operator such that the domain is a single chromosome and the range is a set of those chromosomes that are possible offspring of the recombination operation. The set of possible offspring chromosomes \( \varepsilon \) that is associated with a uniform recombination operation between parent chromosomes \( \alpha \) and \( \iota \) is the set of chromosomes having values complementary to those of \( \alpha \) at any set of indices that is a member of the power set \( P(\phi) \). Equivalently, it could be stated that
every element of the set of possible offspring chromosomes $\varepsilon$ is contained within the highest order schema that contains both parent chromosomes $\alpha$ and $\iota$, as depicted in Figure 20.

![Diagram](image.png)

**Figure 20.** Binary recombination between $\bullet\bullet\bullet$ and $\bullet\square\square$ is equivalent to the unary recombination of $\bullet\square\square$ with the fixed chromosome $\iota = \bullet\bullet\bullet$. The set of possible offspring is equivalent to the highest order schema that contains both $\bullet\bullet\bullet$ and $\bullet\square\square$.

This schema would only contain wildcard characters at indices where chromosomes $\alpha$ and $\iota$ differ and, thus, the set of wildcard character indices would be equivalent to the set $\phi$. For recombination between $\alpha$ and $\iota$ (between which there is a Hamming distance value of $\delta$) the cardinality of set $\phi$ will be $\delta$, and thus the cardinality of the power set $P(\phi)$ will be $2^\delta$. The Hamming distance between the chromosomes $\bullet\bullet\bullet$ and $\bullet\square\square$ (for example) is two, and the complementary index set $\phi$ that encodes a transformation from one to the other is $\{2, 3\}$. Accordingly, the power set $P(\phi) = \{\{\}, \{2\}, \{3\}, \{2, 3\}\}$ specifies
(relative to $t = \text{■■■}$) the set of chromosomes \{\text{■■■}, \text{■■□}, \text{■□■}, \text{■□□}\}, which is equivalent to the set of possible offspring and the set of chromosomes defined by the schema $\text{■**}$ (the highest order schema to contain both $\text{■■■}$ and $\text{■□□}$).

It should be noted that since every chromosome $\alpha$ is described relative to chromosome $t$ using a complementary index set $\phi$, the actual binary values of the chromosome need not have been explicitly noted. As a clarifying example, for the previously depicted chromosome $t = \text{■■■}$, the chromosome defined by the arbitrary complementary index set $\phi = \{2, 3\}$ would be $\text{■□□}$ and the set of possible offspring chromosomes that could be produced by the operation $\chi_{\text{■■■}}(\text{■□□})$ would be \{\text{■■■}, \text{■■□}, \text{■□■}, \text{■□□}\}, defined by complementary index sets \{\}, \{2\}, \{3\}, and \{2, 3\}, respectively. If, on the other hand, chromosome $t = \text{□□■}$, the chromosome defined by the arbitrary complementary index set $\phi = \{2, 3\}$ would be $\text{□■□}$ and the set of possible offspring chromosomes that could be produced by the operation $\chi_{\text{□□■}}(\text{□■□})$ would be \{\text{□□■}, \text{□□□}, \text{□■■}, \text{□■□}\}, also defined by complementary index sets \{\}, \{2\}, \{3\}, and \{2, 3\}, respectively.

Clearly, the possible offspring would remain the chromosomes defined by complementary index sets \{\}, \{2\}, \{3\}, and \{2, 3\}, entailing that the function of the unary recombination remains consistent regardless of the actual values of chromosome $t$. With every chromosome $\alpha$ described relative to $t$, it is sufficient to associate each set of possible offspring chromosomes, denoted $\varepsilon$, with the parent chromosome $\alpha$ which, when recombined with $t$, could produce those chromosomes as offspring.
With the preceding approach for redefining the space of possible chromosomes with respect to a single, fixed chromosome using complementary index sets, the set of unary recombination operators necessary to replace the binary recombination operator can be constructed. For every unique chromosome ι in the population that could act as one operand of the binary recombination operator, there exists a unary operator upon the chromosome space defined in terms of ι. This unary operator takes a single operand chromosome and generates a set of possible offspring chromosomes equal to the set of possible offspring for a binary recombination operation between that operand chromosome and the fixed chromosome ι.

3.4 Classification of Recombinational Distances

Under this paradigm, the distance between two chromosomes α and β in terms of the traversal of the solution space by the recombination operator could be defined in one of two ways. The distance would represent the smallest number of applications necessary to produce chromosome β as the offspring of chromosome α, using (for each application) either any unary recombination operation contained in ξχ and corresponding to a chromosome present in the current population or, alternatively, a single specified unary recombination operator χι contained in ξχ. The distinction between these measures of recombinational distance (denoted Δξχ and Δχι, respectively) is not trivial, as it can be demonstrated that they do not satisfy the same conditions as necessary for classification as a metric. As noted previously in Chapter 2, the conditions for the classification of a
measure as a metric are non-negativity, identity of indiscernibles, symmetry, and subadditivity, respectively.

For the former measure $\Delta_{\xi \chi}$, the distance is defined as the smallest number of applications of any unary recombination operator in $\xi \chi$ necessary to transform one chromosome into another. Since there does not exist a definition for the negation of a recombination operation, the number of recombination operations separating two chromosomes is an integer value from the range $[0, +\infty]$ and, thus, non-negativity is satisfied.

For the identity of indiscernibles condition, if the distance $\Delta_{\xi \chi}$ between chromosomes $\alpha$ and $\beta$ is zero, then, since no recombination operations are necessary for $\alpha$ to produce $\beta$ as an offspring, they must be equivalent. Conversely, if $\alpha$ and $\beta$ both represent the same chromosome, then no recombination operations need be applied for $\alpha$ to produce $\beta$ as an offspring. As such, the identity of indiscernibles condition is also satisfied.

It is, however, trivially demonstrable that the distance $\Delta_{\xi \chi}(\alpha, \beta) \neq \Delta_{\xi \chi}(\beta, \alpha)$ for the singleton population $P = \{\square\square\square\}$, and, consequently, that the symmetry condition is not satisfied. In proof of this, although the unary recombinational operation $\chi_{\square\square\square\square}(\square\square\square\square)$ can produce chromosome $\square\square\square\square$ as an offspring, entailing that $\Delta_{\xi \chi}(\square\square\square\square, \square\square\square\square) \leq \infty$, the unary recombinational $\chi_{\square\square\square\square}$ comprising $\xi \chi$ cannot be applied to chromosome $\square\square\square\square$ to produce $\square\square\square\square$ as an offspring. As this entails that $\Delta_{\xi \chi}(\square\square\square\square, \square\square\square\square) = \infty$, the symmetry condition is not satisfied.
Similarly, it can also be demonstrated that the subadditivity condition (or, equivalently, the triangle inequality), is not satisfied by the definition of the $\Delta_{\xi\chi}$ measure of recombinational distance. It can be easily deduced that the recombinational distance $\Delta_{\xi\chi}(\alpha, \beta)$ is finite if and only if $\beta$ is contained in at least one of the schemata that can be created by replacing the values of each distinct member of the current population with the wildcard character at each index that differs from chromosome $\alpha$. Equivalently, for any unary recombinational operator $\chi_i \in \xi\chi$, $\chi_i(\alpha)$ can produce $\beta$ as an offspring if and only if $\beta$ is a member of the highest order schema that contains both $i$ and $\alpha$. Thus, for exemplar population $P = \{□□□□, □□■■, ■■□□\}$, the distance between chromosomes $\Delta_{\xi\chi}(□□□□, □□■■)$ is infinite since not one of the three unary recombinational operators $\chi_{□□□□}, \chi_{□□■■}, \chi_{■■□□}$ that comprise $\xi\chi$ can be applied to chromosome $□□□□$ to produce $■■■■$ as an offspring.

However, as the operation $\chi_{□□■■}(□□□□)$ can produce $□□■■$ as an offspring (entailing that the distance $\Delta_{\xi\chi}(□□□□, □□■■)$ is finite) and the operation $\chi_{■■□□}(□□■■)$ can produce $■■■■$ as an offspring, (entailing that the distance $\Delta_{\xi\chi}(□□■■, ■■■■)$ is also finite) the substitution of $\alpha = □□□□$, $\beta = □□■■$, and $\kappa = □□□□$ into the subadditivity inequality $\Delta_{\xi\chi}(\alpha, \beta) \leq \Delta_{\xi\chi}(\alpha, \kappa) + \Delta_{\xi\chi}(\kappa, \beta)$ results in an infinite value on the left hand side of the expression and the sum of two finite values on the right hand side. This example, depicted in Figure 21, clearly indicates that the subadditivity condition is not satisfied.

With only non-negativity and the identity of indiscernibles satisfied, this first measure of recombinational distance $\Delta_{\xi\chi}$ must be classified as a semiquasimetric.
Figure 21. Since □□□□ cannot be recombined with one element of the population to produce ■■■■, the distance $\Delta_{\xi\chi}(\Box\Box\Box\Box, \Box\Box\Box\Box) = \infty$. However, □□□□ can produce □□□□, which can itself produce ■■■■. This entails that $\Delta_{\xi\chi}(\Box\Box\Box\Box, \Box□□□□) < \infty$ and $\Delta_{\xi\chi}(\Box□□□□, \Box□□□□) < \infty$. The latter measure of recombinational distance $\Delta_{\chi\iota}$ defines the recombinational distance between two chromosomes $\alpha$ and $\beta$ as the smallest number of applications of any single unary recombination operation $\chi\iota$ contained in $\xi\chi$ necessary to produce chromosome $\beta$ as the offspring of chromosome $\alpha$. The proof that this measure does satisfy the nonnegativity and identity of indiscernibles conditions mirrors the discussion in the previous paragraph and need not be restated. Furthermore, as it was demonstrated that the
symmetry condition was not be satisfied by recombinational distance $\Delta_{\xi\chi}$ with the singleton population example of the previous paragraph, it naturally remains unsatisfied for any single operator $\chi_t \in \xi\chi$. This indicates that the symmetry condition remains unsatisfied by this recombinational distance definition.

However, contrary to the definition of $\Delta_{\xi\chi}$, it can be demonstrated that $\Delta_{\chi_i}$ does satisfy the subadditivity condition that $\Delta_{\chi_i}(\alpha, \beta) \leq \Delta_{\chi_i}(\alpha, \kappa) + \Delta_{\chi_i}(\kappa, \beta)$, if it is accepted that the recombinational distance $\Delta_{\chi_i}(\alpha, \beta) \in \{0, 1, \infty\}$ and, further to the satisfaction of the identity of indiscernibles, if $\alpha \neq \beta$ then $\Delta_{\chi_i}(\alpha, \beta) \in \{1, \infty\}$. As a clarifying example, as the distance between chromosomes $\alpha$ and $\beta$ is now defined as the smallest number of applications of any specific unary recombination operation $\chi_t$, if chromosomes $\alpha$ and $\beta$ are written $\alpha_1--\alpha_2--...--\alpha_\lambda$ and $\beta_1--\beta_2--...--\beta_\lambda$, respectively, then $\beta$ can only be produced by a single recombination (entailing a distance $\Delta_{\chi_i}(\alpha, \beta) = 1$) of $\alpha$ with the population member $t$ if and only if, $\forall 1 \leq j \leq \lambda$, chromosome value $\beta_j = \alpha_j \vee t_j$. If, however, $\exists 1 \leq j \leq \lambda$: $\beta_j \neq \alpha_j \wedge \beta_j \neq t_j$, then irrespective of the number of applications of the unary recombination operation $\chi_t$ to chromosome $\alpha$, chromosome $\beta$ will never be produced as an offspring (entailing a distance $\Delta_{\chi_i}(\alpha, \beta) = \infty$). Having demonstrated that $\Delta_{\chi_i}(\alpha, \beta) \in \{1, \infty\}$, it follows that the left hand side of the inequality $\Delta_{\chi_i}(\alpha, \beta) \in \{1, \infty\}$, and the right hand side of $\Delta_{\chi_i}(\alpha, \kappa) + \Delta_{\chi_i}(\kappa, \beta) \in \{1 + 1, 1 + \infty, \infty + 1, \infty + \infty\} = \{2, \infty\}$. Since the minimum value of the sum on the right hand side of the inequality is 2, all cases where the left hand side is 1 will satisfy the inequality and it suffices to demonstrate that whenever the distance on the left hand side of the inequality is infinite, then the distance
on the right hand side is also infinite. Since cases where $\Delta_{\chi_i}(\alpha, \kappa) = \infty$ ensure that the sum on the right hand side of the inequality will also be infinite, the only case requiring further investigation is when $\Delta_{\chi_i}(\alpha, \kappa) = 1$, which, as noted previously, is equivalent to the operation $\chi_i(\alpha) = \kappa$. Substitution of $\chi_i(\alpha)$ for $\kappa$ in the second addend of the right hand side of the inequality, $\Delta_{\chi_i}(\kappa, \beta)$, results in the expression $\Delta_{\chi_i}(\chi_i(\alpha), \beta)$ and, since the left hand side value of the inequality, $\Delta_{\chi_i}(\alpha, \beta) = \infty$, entails that $\exists 1 \leq j \leq \lambda: \beta_j \neq \alpha_j \land \beta_j \neq \iota_j$, it is recognized that chromosome $\beta$ will never be produced as an offspring of chromosome $\alpha$, regardless of the number of times it is subjected to uniform recombination with chromosome $\iota$. With the subadditivity condition demonstrably satisfied, and only the symmetry condition unsatisfied, this second measure of recombinational distance $\Delta_{\chi_i}$ is most correctly classified as a quasimetric.

3.5 Recombination Operator Induced Topology

Of the semiquasimetric $\Delta_{\xi\chi}$ and the set of relevant quasimetrics of the form $\Delta_{\chi_i}$, for each unique chromosome $\iota$ in the current population, both effectively describe the distance between chromosomes $\alpha$ and $\beta$ in terms of the application of the recombination operator between the operand chromosome $\alpha$ and a single chromosome $\iota$ from the current population. Thus, a topology can be induced upon the set of candidate solution genotypes by using the measure of recombinational distance to generate the recombinational landscape component of the underlying genotypic space. In the formal definition of an adaptive landscape, for which a typically three-dimensional surface is extruded along an
objective function dimension from a planar representation of the underlying genotypic
space, it is recognized that the topology of the genotypic space, defined by a
conceptualization of distance or neighbourhood, is typically specified using either a set of
transition probabilities or a set of move sets. In this context, for any transformation of one
chromosome \( \alpha \) into another chromosome \( \beta \), achieved through the application of a
variational operator, the likelihood with which this transformation can be achieved is
recognized as the transition probability from \( \alpha \) to \( \beta \), and, in only those instances when
this transition probability is greater than 0, chromosome \( \beta \) is assigned membership to the
move set of \( \alpha \). It is not, then, unreasonable to recognize that a move set defined
conceptualization of distance or neighbourhood is merely a relaxation of the precision of
a transition probability defined conceptualization. Indeed, the decision to specify a set of
move sets instead of a set of transition probabilities is equivalent to the simplification of
the task of determining the likelihood with which chromosome \( \alpha \) can be transformed into
chromosome \( \beta \), into the simpler task of determining whether or not chromosome \( \alpha \) can be
transformed into chromosome \( \beta \) (and, in the more general case, what is the set of
chromosomes into which chromosome \( \alpha \) can be transformed). Although a transition
probability conceptualization of distance would provide a more precise view of the
underlying genotypic space topology, and, consequently, a superior conceptualization of
distance, determining the move set of each chromosome in the population is the simpler
task and will be addressed first. Although the word topology is associated with a precise
mathematical definition, the term is commonly (and originally) associated with a study of
the structural features associated with a region. Furthermore, as recombinalional distance
does not meet the criteria for classification as a metric, it would be inaccurate to refer to
genotypic space as a metric space. That being said, Wright's original definition of an adaptive landscape conceived of the evaluation function data extruded from a space that was, intuitively, Euclidean. As it has been demonstrated that the underlying space in not Euclidean in nature (nor is it defined by the Hamming distance measure alone), this dissertation must operationally define a more generalized concept. As such, the word topology is used (on occasion) to describe the neighbourhood structure of genotypic space. It must be stressed that the anatomy of a genotypic space that may be referred to as the topology induced by a operation does not actually satisfy the necessary requirements for the formation of an actual topological space. In demonstration of this, if open-ball type subsets of genotypic space are denoted $T_R(C)$ and defined such that genotype $\alpha$ is an element of subset $T_R(C)$ if and only if the number of population members with which chromosome $C$ must be recombined to produce $\alpha$ is less than $R$, then these would be open subsets of genotypic space. In order for the collection of these subsets to satisfy the requirements for a mathematical definition as a topology, it must be possible to prove that both the union and finite intersection of these open sets are also open. It can, however, be demonstrated in a genotypic space of as few as four chromosomes (i.e. $\lambda = 2$) that the union of two open subsets defined by recombination with a population of only two members can yield a subset which is not, itself, an open subset. Figure 22 depicts such an example for a singleton population of chromosome ■□ – although the open subset centered around chromosome $C = ■□$ and defined by a single recombination operation (i.e. $R = 1$) would yield subset {■■, ■□}, and the open subset centered around chromosome $C = □■$ and defined by a single recombination operation (i.e. $R = 1$) would yield subset {■■, □■}, the union {■■, □□, □■} is not an open subset.
Figure 22. For a singleton population \{■■\}, the open subset centered around ■□ for a single recombination operation yields the subset \{■■, ■□\}, and the open subset centered around □■ for a single recombination operation yields the subset \{■■, □■\}. Since the union of these subsets \{■■, ■□, □■\} is not itself a subset specified by recombination, the necessary conditions to define a topological space have not been satisfied.

Having investigated the actual requirements necessary for mathematical definition as a topology, the word topology will be operationally defined for the entirety of this document as the arrangement of chromosomes that preserves and represents, as closely as possible, the relative lengths of the traversals of genotypic space necessary to transform one chromosome into another. This usage is consistent with the common definition of topography as the physical features of an entity and their structural relationships (Merriam-Webster Online).
3.6 Abstractions of the Recombination Operator

The possible transformations of a chromosome $\alpha$ of the set of possible operand chromosomes and the set $\varepsilon$ of the possible offspring of a recombination operation (between chromosome $\alpha$ and a specified operand chromosome $\iota$) can be stored as an adjacency matrix that would define a directed graph structure representative of the possible recombination operations. Although this would be structurally similar to a component of the matrix employed by Vose (1999) to encode mixing information (the probability that a pair of chromosomes, through both unary mutation and binary recombination, can produce a specific offspring), the adjacency matrix for the proposed digraphic representation of recombination would represent only those transformations that can be achieved through the application of the unary recombination operation alone (as defined by chromosome $\iota$), and would, at least initially, encode only Boolean values for whether or not each chromosome could produce any other in the space solely through the act of recombining with a chromosome $\iota$ of the population. It should be emphasized that Vose provided only the necessary and sufficient properties for a mixing matrix, and was not concerned with the determination of the actual values that would be associated with any specific finite population. Furthermore, as it was Vose's intention to employ the mixing probabilities in tandem with the selection probabilities (which cannot be computed without the evaluation function and a corresponding decrease in generality), for the present task of determining whether or not a given chromosome can be created through the recombination of elements of the current population, the proposed adjacency matrix of Boolean values would incur a lesser computational expense.
Although additional graphical depictions of the functionality of the recombination operator could be generated from each of the previous abstractions discussed in the literature survey, before any comparisons can be made it must be emphasized that these previous approaches were each constructed under different motivations that those of this body of research, and did not employ the unary recombination paradigm introduced here. The conceptualization of recombination as an operator wherein both the domain and the range are defined as pairs of chromosomes (as opposed to those conceptualizations wherein individual chromosomes are the possible states of a recombinationally defined transformation) would not require the establishment of a unary recombination paradigm.

It has been established that the number of applications of uniform binary recombination necessary to transform an origin chromosome into a target chromosome is dependent upon the set of chromosomes with which the origin could be recombined. Since recombination is certainly capable of producing a pair of offspring chromosomes (instead of a single offspring chromosome) with each application, the conceptualization of recombination as a operation (or transformation) that separates pairs of chromosomes is certainly an accurate and reasonable approach. This conceptualization, which would be most naturally represented with something akin to the undirected graph representation depicted in Figure 23 (where the vertices identify with unique pairs of chromosomes) does effectively depict the function of the binary recombination operation. It is certainly possible to create an inter-chromosomal pairing distance matrix using this information, and would, with solution space size $S = 2^λ$, contain $S^4 = (2^λ)^4$ entries. Unfortunately, as will be discussed in greater detail in a later chapter, for the purpose of depicting an
underlying genotypic space suitable for specific data visualization applications, the
definition of $S^4 = (2^λ)^4$ interpoint distances between $S^2 = (2^λ)^2$ distinct points when only $S = 2^λ$ unique genotypes have been defined is counterintuitive and decreases the value
associated with the use of such a visualization.

![Diagram](image)

Figure 23. The set of vertices in this graph is the set of all possible pairs of
cchromosomes to which a recombination operator could be applied. Two vertices are
connected if and only if recombination can be applied to one pair to produce the
other pair as offspring (and vice versa). Although self-loops have been omitted from
this graph, any pair of chromosomes could be recombined to produce the same pair
of chromosomes as offspring.
Furthermore, it is neither necessary nor desirable to treat recombination as an operation with a range defined as a pair of offspring chromosomes. In the absence of a unary recombination paradigm, the cardinality of the operand set is certainly two, (because of the binary arity of recombination), but it is perfectly reasonable to generate only a single offspring chromosome with each application of the operator. Under this conceptualization the set of possible operand sets is different from the set of possible results, and the most logical depiction would take the form of a bipartite graph (such as that which is depicted in Figure 24). The bipartite graph is defined here between a first class that contains $S^2$ vertices and a second class that contains only $S$ vertices. As required for a bipartite graph definition, edges representative of individual recombination operations would connect a pair of chromosomes (depicted with a vertex from the class of $S^2$ vertices) to each of the possible individual chromosomes that could be created as offspring (each of which being depicted as a single vertex from the class of $S$ vertices).

**Figure 24.** A bipartite graph that depicts the application of recombination to a pair of chromosomes (from the set $S^2$) to produce a single chromosome (from the set $S$) as an offspring.
Although this conceptualization addresses concerns regarding the binary cardinality constraint placed upon the result set of a recombination operation, an interpoint distance matrix defined with this conceptualization would have $S^2$ rows and only $S$ columns and would, consequently, be neither a square matrix nor suitable for the desired distance visualization applications.

The final, albeit unusual, abstraction of the recombination operator discussed previously used a hypergraph representation. The set of vertices in a depiction defined thusly would represent each of the $S$ individual chromosomes that could be either an operand or the result of a recombination operation. Contrary to the previously discussed abstractions, where possible offspring chromosomes were represented by vertices in the graph, this proposed representation (for which an example is depicted in Figure 25) would represent each possible offspring chromosome of a recombination operation using a separate hyperedge (depicted as a single edge with multiple labels), connecting the pair of associated operand chromosomes.

As recombination operations applied to two distinct chromosomes could produce (depending upon the cardinality of the complementary index set between the operand chromosomes) between 2 and $\lambda$ offspring chromosomes, two vertices are connected by as many undirected hyperedges as there are possible offspring chromosomes of the recombination of the two chromosomes identified by the two hyperconnected vertices. There would be $2^\lambda$ hyperedges between each pairing of vertices that is representative of a pair of complementary chromosomes, and, since recombination between a pair of
identical chromosomes is well-defined (although unproductive), exactly one hyperedge connecting each vertex with itself. Although such a representation is clearly inadequate for defining any measure of the distance between chromosomes that would be traversed by a variational operator, it must be emphasized that this approach was neither conceptualized nor intended for data visualization applications.

Figure 25. A hypergraphic representation of the functionality of the recombination operator when applied to the genotypic space for $\lambda = 2$. The graph clearly indicates that recombination applied to the complementary chromosomes ■■ and □□ could produce any genotype as an offspring. However, unlike the depictions in Figures 23 and 24, it must be emphasized that no connections are depicted between parents and their offspring.
Despite having revisited the previous abstractions of the recombination operator introduced in the literature survey, the task of defining a conceptualization (and graphical depiction) that would facilitate the substantiation of whether a pair of chromosomes could be recombined to produce another specified chromosome as offspring, remains open. Under the unary recombination paradigm, this task can be rephrased as the substantiation of whether one chromosome could be transformed into another using only the unary recombination operation defined by a single member of the current population. Although the abstractions discussed previously did not explicitly include the complexity analyses for such a task, the complexity implied by each graphical representation can be computed and used as a baseline against which the novel approaches can be contrasted. As noted previously, for a population of size $\rho$ and a solution space size $S$, the undirected graph of chromosome pairs would have $\rho^2$ vertices, of which the number of connected edges is greater than zero. Since each edge actually corresponds to a possible crossover mask, each vertex would have a set of up to $S$ neighbouring vertices, to which a specified offspring may or may not have membership.

The bipartite graph was defined with a class of $S^2$ vertices (for all possible pairs of chromosomes) and a class of $S$ vertices (for each possible offspring chromosome). A recombination operation in such a graph is represented by a set of up to $S$ edges (i.e. connecting up to $\rho^2$ members of the cardinality $S^2$ class with the cardinality $S$ class, for a population of size $\rho$). Any chromosome that could be created as an offspring would be a member of the cardinality $S$ class, connected by one of the edges of the set.
Finally, each recombination operation depicted in the hypergraph abstraction could have, between every pairing of the vertices representing chromosomes of the current population (of which there are $\rho^2$), a set containing as many as $S$ hyperedges (each representing a single possible offspring chromosome), to which an offspring chromosome may or may not have membership.

It can be inferred from the preceding graphical analyses (for each of the previously noted abstractions of the recombination operators) that the substantiation of whether any pair of chromosomes from the current population could be recombined to produce a specific offspring would have a worst-case time complexity on the order of $\rho^2 S = \rho^2 2^\lambda$. However, it will be demonstrated shortly that the time complexity associated with such a task can be reduced substantially.

### 3.7 Digraphic Representation of Recombination

Since the recombination operations discussed herein probabilistically determine whether or not the value at each index of a chromosome will be exchanged independently, the adjacency matrix used to define the directed graph representation for recombination between chromosomes of length $\lambda$ can be constructed recursively from adjacency matrices for chromosomes of length $\lambda-1$. Under the temporary assumption that chromosome $t$ (for which the unary recombination operator is to be defined) is the binary string of length $\lambda$ comprised entirely of zero bits, there exists a $2^\lambda \times 2^\lambda$ matrix $\mu$ of Boolean values where entry $\mu_{ij}$ indicates whether or not recombination between $t$ and the
The $i^{th}$ member of the chromosome space can yield the $j^{th}$ member of the chromosome space as an offspring. The matrix that would function as the basis for a recursive construction would be used for a chromosome length of 1 and, thus, entry $\mu_{00}$ would indicate whether or not chromosome $\iota$ (which is '0') and the zeroth member of the chromosome space (which is also '0') can be recombined to produce the zeroth member of the chromosome space (which is also '0') as an offspring. Entry $\mu_{01}$, on the other hand, would indicate whether or not chromosome $\iota$ (which is '0') and the zeroth member of the chromosome space (which is also '0') can be recombined to produce the first member of the chromosome space (which is '1') as an offspring. For single bit chromosome recombination, the entries $\mu_{00}$, $\mu_{01}$, $\mu_{10}$, and $\mu_{11}$ would be assigned the Boolean values true, false, true, and true, respectively.

$$\text{for } \lambda = 2: \quad \mu = \begin{bmatrix} \text{true} & \text{false} \\ \text{true} & \text{true} \end{bmatrix}$$

For the recursive step in the construction of an adjacency matrix of the digraph representation for a chromosome of length $\lambda$, assume that the adjacency matrix of the digraph representation for a chromosome of length $\lambda - 1$ is complete and accurate. For entry $\mu_{ij}$ of the adjacency matrix for a chromosome of length $\lambda$ to have a value of true, it must be possible to recombine the $i^{th}$ member of the chromosome space of length $\lambda$, denoted "$i_1 i_2 i_3 \ldots i_{\lambda}$", with a chromosome $\iota$ of length $\lambda$ of only zero bits, such as $\square \square \square \ldots \square$, and produce the $j^{th}$ member of the chromosome space of length $\lambda$, denoted "$j_1 j_2 j_3 \ldots j_{\lambda}$" as an offspring. In the case where $i_1 = \square$ this recombination is possible if and only if "$i_2 i_3$
... i\(\lambda\)" and ■■...■ can be recombined to produce "j\(2\) j\(3\) ... j\(\lambda\)", since an i\(1\) of □ can be recombined with a ■ from ι to produce either possible value of j\(1\). Consequently, the \(2^{\lambda-1} \times 2^{\lambda-1}\) entries \(\mu_{ij}\) of the adjacency matrix for length \(\lambda\) for i from \([2^{\lambda-1}+1...2^{\lambda}]\) and j from \([1...2^{\lambda-1}]\) and the \(2^{\lambda-1} \times 2^{\lambda-1}\) entries \(\mu_{ij}\) of the adjacency matrix for length \(\lambda\) for i from \([2^{\lambda-1}+1...2^{\lambda}]\) and j from \([2^{\lambda-1}+1...2^{\lambda}]\) will both be precise copies of the adjacency matrix associated with chromosomes of length \(\lambda-1\). In the alternative case, where i\(1\) = "0", recombination between "i\(1\) i\(2\) i\(3\) ... i\(\lambda\)" and ■■...■ can only produce "j\(1\) j\(2\) j\(3\) ... j\(\lambda\)" as an offspring chromosome if and only if j\(1\) = "0" and "i\(2\) i\(3\) ... i\(\lambda\)" and ■■...■ can be recombined to produce "j\(2\) j\(3\) ... j\(\lambda\)" as an offspring. Consequently, the \(2^{\lambda-1} \times 2^{\lambda-1}\) entries \(\mu_{ij}\) of the adjacency matrix for length \(\lambda\) for i from \([1...2^{\lambda-1}]\) and j from \([1...2^{\lambda-1}]\) will also be a precise copy of the adjacency matrix associated with chromosomes of length \(\lambda-1\) and the \(2^{\lambda-1} \times 2^{\lambda-1}\) entries \(\mu_{ij}\) of the adjacency matrix for length \(\lambda\) for i from \([1...2^{\lambda-1}]\) and j from \([2^{\lambda-1}+1...2^{\lambda}]\) will all have a value of false.

To demonstrate, consider the construction of the 4 \(\times\) 4 adjacency matrix of the digraph for chromosomes of length two. Under the assumption that chromosome ι is comprised entirely of zero bits (in this case, ■■), recombination with the first chromosome, ■■, can produce only ■■ as an offspring. Thus, the first row of the adjacency matrix for the digraph will be [true false false false]. Recombination between ι and the second chromosome, ■□, can produce either ■■ or ■□ as an offspring and, thus, the second row of the matrix will be [true true false false]. Similarly, the third and fourth rows of the matrix will be [true false true false] and [true true true true], respectively.
for $\lambda = 3$:

\[
\begin{bmatrix}
\text{true} & \text{false} & \text{false} & \text{false} \\
\text{true} & \text{true} & \text{false} & \text{false} \\
\text{true} & \text{false} & \text{true} & \text{false} \\
\text{true} & \text{true} & \text{true} & \text{true}
\end{bmatrix}
\]

As expected from the structural induction proof of the preceding paragraph, if the adjacency matrix of the digraph representation for chromosomes of length two is bisected vertically and horizontally into exactly four $2 \times 2$ adjacency matrices, the top-left, bottom-left and bottom-right submatrices are precise copies of the basis matrix (i.e. the adjacency matrix of the digraph representation for chromosomes of length one). The top right, on the other hand, is a $2 \times 2$ matrix containing only Boolean values of false.

\[
\begin{bmatrix}
\text{true} & \text{false} & \text{false} & \text{false} \\
\text{true} & \text{true} & \text{false} & \text{false} \\
\text{true} & \text{false} & \text{true} & \text{false} \\
\text{true} & \text{true} & \text{true} & \text{true}
\end{bmatrix} = \begin{bmatrix} A & B \\ A & A \end{bmatrix}
\]

where: $A = \begin{bmatrix} \text{true} & \text{false} \\ \text{true} & \text{true} \end{bmatrix}$ and $B = \begin{bmatrix} \text{false} & \text{false} \\ \text{false} & \text{false} \end{bmatrix}$

It also follows that if the adjacency matrix for chromosome length three is bisected vertically and horizontally into exactly $4 \times 4$ adjacency matrices, the top-left and bottom matrices are copies of the adjacency matrix for the digraph representation of recombination between chromosomes of length two, and the top right is a $4 \times 4$ matrix comprised entirely of zeros. Digraphic representation derived from these adjacency matrices are depicted in Figure 26.
Figure 26. The digraphic representations specified by the adjacency matrices specified for $\lambda = 2$ (above) and $\lambda = 3$ (below). The presence of the former digraph (for $\lambda = 2$) as a component in the latter digraph (for $\lambda = 2$) is evident, especially in the subgraph containing only vertices ■■■, ■■□, ■□■, and ■□□.
It can be concluded, from the proof and discussion contained in the previous section, that as long as the fixed parent chromosome $\iota$ of a recombination operation is a binary string of zero digits, there is a trivially simple recursive algorithm that will determine whether the chromosome $\beta$ can be produced as an offspring of a recombination operation between the fixed parent chromosome $\iota$ and any chromosome $\alpha$. This algorithm, in order to determine whether the $i^{th}$ member of the chromosome space can produce the $j^{th}$ member of the chromosome space as an offspring through recombination with a chromosome comprised entirely of zero bits, entails determining whether the entry $\mu_{ij}$ of the adjacency matrix lies in the top right quadrant of the adjacency matrix. If so, it can be concluded that a recombination operation between the $i^{th}$ member of the chromosome space and the zero bit chromosome cannot produce the $j^{th}$ member of the chromosome space as an offspring. If, however, the entry $\mu_{ij}$ lies in any other quadrant of the adjacency matrix, the same algorithm is recursively applied to the $2^{nd}$ through the $\lambda^{th}$ bits of chromosomes $i$ and $j$, and repeated until only the adjacency matrix associated with chromosomes of length 1 need be consulted.

As an alternative to the development of a similar proof for every other possible value of the fixed parent chromosome $\iota$, it would suffice to demonstrate that there exists a reversible transformation that, when applied to both the parent and offspring chromosomes, would convert one of the parent chromosomes into the binary string comprised entirely of zeros without otherwise affecting the relationships between them. Under this transformation, denoted $\tau$, the Boolean value describing whether or not recombination between a pair of chromosomes $\kappa_i$ and $\kappa_j$ can yield chromosome $\kappa_k$ as an
offspring would be equivalent to the Boolean value describing whether or not a recombination operation, applied to a chromosome that is comprised entirely of zero bits and chromosome $\tau(\kappa_j)$ can yield the chromosome $\tau(\kappa_k)$ as an offspring.

Vose (1999) noted such a transformation in the second lemma of his technical report on the formalization of the genetic algorithm to be the application of the bitwise exclusive disjunction operator. The use of this operator entails that a single digraph representation of the application of a recombination operation with a chromosome comprised entirely of zero bits is, in fact, effectively representative of the application of a recombination operation to any chromosome.

If the previously mentioned adjacency matrix has already been constructed, wherein the Boolean value of entry $\mu_{ij}$ indicates whether or not recombination between a chromosome $\iota$ comprised entirely of zero bits can be recombined with the $j^{th}$ member of the chromosome space to yield the $k^{th}$ member of the chromosome space as an offspring, then the question of whether uniform recombination between the pair of simulated chromosomes $\kappa_i$ and $\kappa_j$ can yield chromosome $\kappa_k$ as an offspring is equivalent to the question of whether recombination between a chromosome comprised entirely of zero bits and $\tau(\kappa_j)$ can yield chromosome $\tau(\kappa_k)$ as an offspring. This Boolean value, in turn, can be read directly from the adjacency matrix.

If transformation $\tau$ is the application of a bitwise exclusive disjunction operation (represented with the symbol $\oplus$) to the $i^{th}$ member of the chromosome space and each of
the i\textsuperscript{th}, j\textsuperscript{th}, and k\textsuperscript{th} members of the chromosome space, then \( \tau \left( "x_1 x_2 x_3 \ldots x_\lambda" \right) \) would be equivalent to "\( i_1 \oplus x_1 i_2 \oplus x_2 i_3 \oplus x_3 \ldots i_\lambda \oplus x_\lambda \)". Since exclusive disjunction results in a value of false if and only if the two operands are either both true or both false, then \( \tau \left( "i_1 i_2 i_3 \ldots i_\lambda" \right) \) would be equivalent to "\( i_1 \oplus i_1 i_2 \oplus i_2 i_3 \oplus i_3 \ldots i_\lambda \oplus i_\lambda \)", also equivalent to ■■■...■.

To solve for the Boolean value of whether recombination between the i\textsuperscript{th} and j\textsuperscript{th} member of the chromosome space, denoted "\( i_1 i_2 i_3 \ldots i_\lambda \)" and "\( j_1 j_2 j_3 \ldots j_\lambda \)" respectively, can produce the k\textsuperscript{th} member, denoted "\( k_1 k_2 k_3 \ldots k_\lambda \)", as an offspring, the application of a bitwise exclusive disjunction operations with "\( i_1 i_2 i_3 \ldots i_\lambda \)" will transform the i\textsuperscript{th}, j\textsuperscript{th}, and k\textsuperscript{th} members of the chromosome space into chromosomes ■■■...■, \( \tau \left( " j_1 j_2 j_3 \ldots j_\lambda" \right) = "i_1 \oplus j_1 i_2 \oplus j_2 i_3 \oplus j_3 \ldots i_\lambda \oplus j_\lambda \)\), and \( \tau \left( " k_1 k_2 k_3 \ldots k_\lambda" \right) = "i_1 \oplus k_1 i_2 \oplus k_2 i_3 \oplus k_3 \ldots i_\lambda \oplus k_\lambda \)\), respectively. It then suffices to prove that the Boolean value describing whether uniform recombination between chromosomes ■■■...■ and "\( i_1 \oplus j_1 i_2 \oplus j_2 i_3 \oplus j_3 \ldots i_\lambda \oplus j_\lambda \)" can produce chromosome "\( i_1 \oplus k_1 i_2 \oplus k_2 i_3 \oplus k_3 \ldots i_\lambda \oplus k_\lambda \)" as an offspring is equivalent to the Boolean value describing whether uniform recombination between the i\textsuperscript{th} and j\textsuperscript{th} member of the chromosome space can produce the k\textsuperscript{th} member of the chromosome space as an offspring.

For this to be true it must be shown that, for all values of \( x \), \( (i_x \oplus k_x) = 0 \lor (i_x \oplus j_x) \) will be true if and only if \( k_x = (i_x \lor j_x) \) is also true. This particular fact can be most easily demonstrated through the use of a simple truth table, which has been included as Table 1.
Table 1. A truth table demonstrating that \((i_x \oplus k_x) = 0 \lor (i_x \oplus j_x)\) will be true if and only if \(k_x = (i_x \lor j_x)\) is also true.

Since the set of possible offspring chromosomes that can be produced by the application of uniform recombination operations to chromosomes of length \(\lambda\) is equivalent to the set of possible chromosomes with which chromosome \(\tau\) could be recombined to create offspring chromosomes, and since both sets are present in the digraph representation of recombination, the number of possible resultant offspring chromosomes is \(2^{\lambda}\).

Furthermore, since the \(C(\lambda, \delta)\) unique chromosomes at Hamming distance \(\delta\) (\(0 \leq \delta \leq \lambda\)), represent every possible chromosome with which a chromosome could be recombined, and the cardinality of the set of possible offspring that could be produced from a recombination operation applied to chromosomes between which there is a Hamming distance of \(\delta\) is \(2^\delta\), the exact number of arcs present in the offspring digraph is \(\delta=0^{\lambda} C(\lambda, \delta) \cdot 2^\delta = (1 + 2)^\lambda = 2^\lambda\).
3.8 Time Complexity Considerations

As noted previously, the abstractions of the recombination operation specified previously implied a worst-case time complexity on the order of $\rho^2 S = \rho^2 \lambda^2$ for the task of substantiating whether or not any pair of chromosomes taken from the current population could be recombined to produce another specified chromosome as offspring. Under the proposed methodology, substantiating whether a given chromosome can be produced by a population through a single application of a binary recombination operator remains equivalent to determining whether a given chromosome can be produced from any pair of chromosomes in the population, necessitating the same $S^2$ component of the worst-case time complexity associated with the examination of every possible chromosome pairing. However, although it remains true that recombination between a pair of complementary chromosomes could result in any chromosome in the search space $S$ as an offspring, performing the determination of whether or not a matrix entry is located in the top right quadrant, at most $\lambda$ times, has a worst-case time complexity of $O(\lambda)$.

Overall, the time complexity of the proposed recursive algorithm is the sum of the complexity of locating the appropriate matrix entries for all possible chromosome pairings, $O(\rho^2 \lambda)$, and the complexity of the application of the bitwise exclusive or operations necessary to redefine the chromosomes of the current population in terms of each possible fixed parent, also $O(\rho^2 \lambda)$, for a total worst-case time complexity of $O(\rho^2 \lambda)$. Thus, the time complexity has been reduced from $O(\rho^2 \lambda^2)$ to $O(\rho^2 \lambda)$, which constitutes a logarithmic speedup.
It was noted previously that every possible offspring chromosome that could be produced by recombining parent chromosomes $\alpha$ and $\iota$ would belong to the set of chromosomes that is defined by the highest order schema that contains both parent chromosomes $\alpha$ and $\iota$. This schema definition would contain wildcard characters at indices where chromosomes $\alpha$ and $\iota$ differ and non-wildcard characters equivalent to the corresponding values of chromosomes $\alpha$ and $\iota$ at all other indices. Consequently, the construction of such a schema for every unique pair of chromosomes present in the population (of which there are $\rho^2$), is straightforward (and will be discussed in greater detail shortly). A chromosome $\kappa$ is assigned membership to a specific schema if and only if either the chromosome $\kappa$ and the schema both contain the same value at index $i$, or the schema contains a wildcard character value at index $i$, for each index $i$ where $0 \leq i \leq \lambda$. As such, the substantiation of the membership of a specific chromosome in a given schema is associated with a worst-case time complexity of $O(\lambda)$, and, as there are $\rho^2$ unique pairs of chromosomes, the worst-case time complexity of the complete determination is $O(\rho^2\lambda)$, which is the same as what would be achieved using the self-similar matrix described in the previous section.

Consulting the self-similar matrix described in the previous section for substantiating whether or not any pair of chromosomes taken from the current population could be recombined to produce another specified chromosome as offspring has the same time complexity as determining whether or not a chromosome has membership in any of the highest order schemata generated for each unique pair of chromosomes. Although the latter approach can, perhaps, be described with greater ease than the former, it should be
observed that both approaches are, in fact, equivalent. By defining the highest order schema that contains both members $\alpha$ and $\iota$ of a unique pair of chromosomes, the complete set of possible offspring that could be produced by recombining chromosomes $\alpha$ and $\iota$ is being generated. However, instead of performing a search of this set randomly, an arbitrary (but straightforward) ordering of the possible offspring chromosomes is inducing a structure on the set of possible offspring. By testing a specific chromosome $\kappa$ against each non-wildcard value of the ternary label that defines a schema, the set of the possible offspring of a chromosome pairing that might match chromosome $\kappa$ is being iteratively bisected, entailing that the set of possible offspring is actually being subjected to a binary, rather than a linear, search. The logarithmic speedup of the highest order schemata membership approach to determining whether or not a population can produce a specific chromosome as an offspring is, thus, not an entirely unexpected result.

The preceding section has precisely defined the structure and properties of the adjacency matrix of a directed graph representation of recombination, and represents a considerable investment of effort. This effort was justified because, in addition to its functionality in defining an approach to the substantiation of whether any pair of chromosomes from the current population could be recombined to produce a specific offspring, it has also provided a useful visual representation that can be extended to the adaptive landscape visualization.

It was previously noted that (with the sufficient condition that no operators other than point mutation and uniform recombination are employed) the mixing matrix component
of the mechanism of Vose’s infinite population genetic algorithm model is completely independent of the fitness function. Similarly, the structure and properties of the directed graph representation of the recombination operator discussed previously are also fitness function independent. Consequently, the adjacency matrix remains constant for any problem for which a binary representation of length \( \lambda \) was selected for the genotypic representation of candidate solutions. This independence ensures that any time-complexity constraints upon the substantiation of whether or not any pair of chromosomes taken from the current population could be recombined to produce another specified chromosome as offspring can be alleviated by storing the \( S^2 = 2^{2\lambda} \) values (for the \( 2^\lambda \times 2^\lambda \) Boolean matrix associated with a chromosome \( \eta \) of entirely zeroes) for access in constant time, albeit with the \( O(\lambda) \) application of the transformation \( \tau \).

### 3.9 Novel Approaches to Measurement

This chapter has, to this point, established an approach to the substantiation of whether or not a specific chromosome can be created through the application of recombination with a given population. The ability to make this determination makes it possible to construct move set conceptualizations of neighbourhood (as introduced in Section 2.5) for the genotype of every candidate solution configuration. Thus, the recombinational distance measurements denoted \( \Delta_{\xi\chi} \) and \( \Delta_{\chi\eta} \) (representing the application of either any number of unary recombination operations defined by the population or a specific unary recombination operation defined by the population, respectively) can now be defined using a move set conceptualization of distance.
For a given population, the $\Delta_{\chi t}$ distance between chromosomes $\alpha$ and $\beta$ (determined with respect to recombination with chromosome $t$) will be $\infty$ for all cases where chromosome $\beta$ does not belong to the move set of $\alpha$ that is defined by unary recombination with fixed chromosome $t$. Equivalently, if chromosome $\beta$ neither belongs to the highest order schema containing both $\alpha$ and $t$, nor is the Boolean entry $\mu_{ij}$ for $i = \tau(\alpha) = \alpha \oplus t$ and $j = \tau(\beta) = \beta \oplus t$ of the adjacency matrix assigned a value of true, $\beta$ is not a member of the move set of $\alpha$. Since it was noted previously that the $\Delta_{\chi t}$ distance between chromosomes $\alpha$ and $\beta$ would only be assigned the value 0 if and only if chromosome $\alpha$ was identical to chromosome $\beta$, it remains only to determine what other finite values could be assigned to the $\Delta_{\chi t}$ distance. Unfortunately, if the $\Delta_{\chi t}$ distance between chromosomes $\alpha$ and $\beta$ is not assigned a value of 0, and $\beta$ does belong to the highest order schema containing both $\alpha$ and $t$ (with entry $\mu_{ij}$ for $i = \tau(\alpha) = \alpha \oplus t$ and $j = \tau(\beta) = \beta \oplus t$ in the adjacency matrix having a value of true), it can only be concluded that the move set defined recombinational distance $\Delta_{\chi t}$ between $\alpha$ and $\beta$ lies in the range $[1, +\infty)$.

To evaluate possible approaches by which a less impractical range for nonzero, finite recombinational distances $\Delta_{\chi t}$ can be calculated, it is necessary to introduce an operational definition for the resolution of a distance measure. This distance resolution is defined as the number of possible intervals or values that a recombinational distance could be assigned if it has already been determined that the distance is neither zero nor infinite. Currently, as the set of resultant values or intervals for a recombinational distance $\Delta_{\chi t}$ contains only the two values 0 and $\infty$ and the single interval $[1, +\infty)$ for a
total cardinality of 3, the distance resolution would also be 3. This represents the minimum possible resolution if support for both zero and infinite distances is included.

For digraphic representations (such as Figure 27) depicting each chromosome with a single node, every self-loop (i.e. edge from a node to itself) would represent the assignment of a distance of 0 to the traversal of a chromosome to itself. Since one zero length traversal exists for each chromosome, self-loops can be omitted with no loss of information. As the only other values that could be assigned are 1 and $+\infty$, the presence or absence of an edge between chromosomes would indicate finite or infinite distance, respectively. Figure 27 illustrates this for $\lambda = 3$ and the population $\{\text{■■□}, \text{■□■}\}$.

Figure 27. The digraphic representation for $\lambda = 3$ and the population $\{\text{■■□}, \text{■□■}\}$.
The distance resolution could certainly be improved upon if a transition probability defined conceptualization of distance $\Delta \chi_i$ were to replace the move set defined conceptualization. Under the assumption that the selection mechanism has already identified the subset $P'$ of the current population $P$ to which the recombination operator will be applied, if chromosome $t$ appears exactly $n_t$ times in the subpopulation $P'$ it is possible (albeit difficult) to compute the transition probability governing the likelihood that another member $\alpha$ of the population will recombine with $t$ to produce $\beta$ as an offspring. The probability of such an occurrence would be the joint probability that $\alpha$ is selected to undergo recombination in the first place (since only a fraction of $P'$ is recombined, typically), combined with the probability that any copy of $t$ would be selected as the second parent in the recombination operation, further combined with the probability that the recombination operation actually does produce $\beta$ as an offspring. If $\beta$ has membership in the move set of $\alpha$ defined by recombination with $t$, then at every index where $\beta$ differs from $\alpha$, the offspring chromosome must be assigned the value from $t$ (instead of the value from $\alpha$). Furthermore, for every value from $t$ to be assigned to the offspring chromosome, the value assigned must be identical to the value from $\beta$.

Uniform recombination is applied to each index independently, and the probability with which the value at any specific index from $t$ is given to the offspring instead of the value from $\alpha$ is denoted $p_{\chi}$. It follows that if the number of wildcard characters in the highest order schema that contains both $\alpha$ and $t$ is denoted $w$, and the cardinality of the complementary index set between $\beta$ and $\alpha$ is denoted $q$, then the likelihood that a recombination operation between $\alpha$ and $t$ would produce $\beta$ as an offspring (assuming $\beta$
has membership in the move set of $\alpha$ defined by recombination with chromosome $\iota$ would be $q \cdot p_\chi \cdot (\lambda-w) \cdot (1- p_\chi)$, for chromosomes of length $\lambda$. Since values of $p_\chi \neq 0.5$ imply that the ordering of the parents is significant, the transition probability determined previously for $\alpha$ producing $\beta$ as an offspring by recombining with $\iota$ must be combined with the transition probability determined for $\iota$ producing $\beta$ as an offspring by recombining with $\alpha$. Furthermore, it is emphasized that this represents a transition probability defined conceptualization of measure $\Delta_\chi$, and must still be computed for each unique chromosome in the population to construct a transition probability defined conceptualization of measure $\Delta_\chi$.

For a transition probability defined conceptualization of distance measure $\Delta_\chi$, calculated with the aforementioned technique, since the set of possible values contains the two values 0 and $\infty$ and as many transition probability values from the interval $(0, 1)$ as are supported by the implementation, the distance resolution of such an approach is vastly increased. Figure 28 revisits the example from Figure 27 (where $\lambda = 3$ and population $\{\text{■□□}, \text{■□■}\}$) but the depiction has been made more representative by adjusting the edge weights according to the number of recombination operations required.

It is important to observe that this technique requires the assumption that both $\alpha$ and $\iota$ were present in the subpopulation to which the variational operators will be applied – although it might appear that a move set defined conceptualization does allow the distance between two chromosomes not present in the population to be computed, both techniques entail similar assumptions.
Figure 28. An enhancement to the digraphic representation from Figure 27, for $\lambda = 3$ and the population \{■■□, ■□■\}. Contrary to the previous depiction, the weight applied to each edge is representative of the number of unary recombination operations (from the set \{recombination with ■■□, recombination with ■□■\}) that must be applied to achieve the transformation. The fact that the edge between chromosomes ■■■ and ■■□ is depicted with a heavier weight than the edge between chromosomes ■■■ and ■□□ indicates that it requires fewer unary recombination operations to transform ■■■ into ■■□ than to transform ■■■ into ■□□.

The possible resolutions defined for the recombinational distance measurement denoted $\Delta_\lambda$ were bounded from below at a value of three and from above by two plus the sum of the cardinality of the set of possible transition probabilities supported by the
implementation. Unfortunately, this substantial increase is achieved at a considerable computational expense. In an effort to balance complexity and computational expense against effectiveness, a third approach was proposed that would yield a recombinational distance measure with an intermediate resolution at a reduced computational expense. The possible recombinational distance value extrema $\infty$ and 0 between chromosomes $\alpha$ and $\beta$ in the presence of current population $P$ are retained, for cases where there exists at least one index for which the value in $\beta$ is neither the corresponding value in $\alpha$ nor any member of $P$ and cases where $\alpha \equiv \beta$, respectively.

It suffices then to contrast the nonzero, finite distances between $\alpha$ and $\beta$ when it is known that some subset of the current population can be recombined with chromosome $\alpha$ to ultimately produce $\beta$ in a future generation as an offspring. It is the cardinality of this population subset that can be used to establish recombinational distances between 0 and $\infty$ and, thus, can be used to derive an improvement to the resolution of recombinational distance measure $\Delta_{\xi\chi}$ from the lower bound of three. It is reasonable to establish that any chromosome $\alpha$ that can be recombined with a single chromosome from the current population to produce chromosome $\beta$ as an offspring is at a lesser recombinational distance from chromosome $\beta$ than a chromosome $\alpha'$ that must be recombined with a two or more chromosomes from the current population to produce $\beta$ as an offspring.

To clarify using the example from the previous figure (Figure 28), ■■■ need only be recombined with one member of the current population {■■□, ■□■} to produce ■■□ as an offspring, but ■■■ must be recombined with both members of the current population
{■■□, ■□■} to produce ■□□ as a descendent. Consequently, the recombinational distance assigned to the traversal of genotypic space from ■■■ to ■□□ is less than the recombinational distance assigned to the traversal of genotypic space from ■■■ to ■□□. This was indicated in Figure 28 by the difference in the weights with which the respective edges were depicted.

Equivalently, if the complementary index set encoding the transformation between α and β is a subset of one of the complementary index sets that encodes the transformation between α and each member of the population, then α is at a lesser recombination distance from β than if the complementary index set between α and β is only a subset of the union of two or more complementary index sets defined between α and each element of the population. Since the cardinality of the complementary index set between α and β can be no greater than the chromosome length λ, the resolution of the recombinational distance measure $\Delta_{\xi\chi}$ for nonzero and finite distances will have a value of λ. This result is not surprising given that the minimum and maximum number of values that a single parent chromosome can contribute to its offspring (assuming a contribution is made at all) are 1 and λ, respectively.

Thus, overall resolution for recombinational distance measure $\Delta_{\xi\chi}$ that can be achieved by this approach is $\lambda+2$, representing a set of possible values $[0, \lambda] \cup \{\infty\}$. Figure 29 illustrates that the range of possible finite recombinational distances for $\lambda = 3$ and a population {■■□, ■□■, □■■} is bounded by the interval $[0, \lambda]$. 

108
Figure 29. Although the chromosome ■■■ can produce any binary genotype of length three through the application of the recombination operators defined by the population {■□■, □■■, □■■}, the production of specific offspring will never require more than three recombinational operations.

While investigating the move set defined approach to recombination distance measurement, techniques were introduced for substantiating whether or not a given chromosome could be transformed into another using only a unary recombination operation defined by a specified member of the current population. Naturally, these can be extended to techniques for substantiating whether a given chromosome could be
transformed into another using only the recombination operator and a subset of the current population. Two techniques to the substantiation have been introduced – the former employing a self-similar matrix representative of a proposed graphical depiction and the latter using highest order schemata membership. It was previously demonstrated that these approaches are equivalent, and since the latter is less complex it will form the basis by which a measure of the recombinalional distance $\Delta_{\xi\chi}$ from $\alpha$ to $\beta$ will be assigned a value from the set of possible values $[0, \lambda] \cup \{\infty\}$.

It was noted previously that, in the absence of a mutation operator, chromosome $\alpha$ cannot produce an offspring chromosome $\beta$ with a specific value at a given index unless at least one chromosome in the current population has the same value as chromosome $\beta$ at the same index. For each index in the chromosome of the offspring of the uniform recombination of two binary strings, if the value at that same index in the first parent is equal to the value at the same index in the second parent, the value at that index in every possible offspring must also be the same. If, on the other hand, the values at that same index in the parent chromosomes are complementary, the possible offspring might have either value at the specified index. By assigning the wildcard character from Holland's schema theorem to indices for which the parent values are complementary and the binary typed values of $\alpha$ to every other index, the set of possible offspring for a given pair of chromosomes can be identified by the resulting schema.

Although it remains true that each chromosome of length $\lambda$ is a member of $2^\lambda$ schema, by employing wildcard characters only at indices where the parent chromosomes differ, the
The resulting schema is the highest order schema that contains every possible offspring. Similarly, the diameter of the set of chromosomes identified by a schema can be increased to include additional chromosomes simply by using this approach to determine the highest order schema that contains both the smaller diameter schema and the additional chromosome. The process for expanding a schema to include an additional chromosome is included, in pseudocode, below.

*The variable `<chromosome>` is the binary string for the chromosome to be included.*

*The variable `<schema>` is the ternary string that specifies the schema that will be made to include `<chromosome>` by replacing some of the fixed characters with wildcards.*

```plaintext
expand-schema

(<schema, chromosome>)

for i from 0 to length(<chromosome>)

    if <schema>[i] != '*' \&<schema>[i] != <chromosome>[i]

        <schema>[i] = '*'

return <schema>
```

To demonstrate the functionality of the preceding algorithm, consider, as an example, the schema $Z = \text{■■■*}$ that defines the possible offspring of a population of chromosomes ■■■■ and ■■□. If the chromosome $\kappa = \text{■■□■}$ is added to the population, this algorithm
would return schema ■■** to which all offspring that can be generated by any number of recombination operations with the new population are assigned membership.

This approach can be applied to any population of chromosomes P to construct the highest order schema that contains every possible offspring that can be created by the application of any number of recombination operations to the current population. Furthermore, for each chromosome $\alpha$ of the chromosome space S, the move set of chromosomes that can be reached by the application of a uniform recombination operation between the chromosome $\alpha$ and a member of the current population P remains equivalent to the set identified by the highest order schema that contains all the chromosomes that are members of the current population and chromosome $\alpha$.

With the previous approach to expanding a given highest order schemata (containing a set of chromosomes) to include an additional chromosome, the complete technique for computing recombinational distance $\Delta_{\xi\chi}$ between $\alpha$ and $\beta$ at a resolution of $\lambda+2$ can be constructed easily. Testing first to establish whether or not $\alpha \equiv \beta$, to determine whether the assignment of a recombinational distance of 0 between them is necessary, schemata are constructed for each unique ordered pair of chromosomes of the form $(\alpha, \iota)$, for any chromosome $\iota$ of the current population. Membership of $\beta$ to any of these schemata would indicate that a single recombination operation between $\alpha$ and a member of the current population could produce $\beta$ as an offspring, and, thus, that a recombinational distance of 1 should be assigned to the transformation from $\alpha$ to $\beta$. Otherwise, if $\beta$ cannot claim membership to any of these schemata, new schemata are constructed for each
unique ordered chromosome triples of the form \((\alpha, t_1, t_2)\), for chromosomes \(t_1\) and \(t_2\) of the current population, where the complementary index between \(t_1\) and \(t_2\) has a cardinality greater than zero.

Membership of \(\beta\) to any of these new schemata would indicate that although a single recombination operation between \(\alpha\) and a member of the current population cannot produce \(\beta\) as an offspring, the recombination of chromosome \(\alpha\) and a member of the current population \(t_1\) can result in an offspring that could be recombined with chromosome \(t_2\) to produce chromosome \(\beta\) as an offspring. This would, naturally, indicate that a recombinational distance of 2 should be assigned to the transformation from \(\alpha\) to \(\beta\).

This process of finding each possible schemata using ordered \(n\)-tuples of the form \((\alpha, t_1, t_2, ..., t_{n-1})\) is similarly repeated for \(4 \leq n \leq \lambda+1\) to establish appropriate recombinational distance measurement values from the range \((3, \lambda)\). If, after having constructed every \(\lambda+1\)-tuple defined schema for the current population and established that \(\beta\) cannot claim membership in any such schema, it can be concluded that the recombination operator alone, applied between \(\alpha\) and any number of members the current population, cannot produce chromosome \(\beta\) from chromosome \(\alpha\) and the recombinational distance between them is, thus, infinite.

The complete algorithm for this computation of recombinational distance \(\Delta_{\xi\chi}\) can then be summarized with the pseudocode on the following page. As stated previously, this
approach establishes the maximum finite distance that might separate two chromosomes at \( \lambda \) recombination operations, and although it is now feasible to compute the recombinational distance between two chromosomes to be either an infinite value or an integer value in the interval \([0, \lambda]\), it should be noted that as there are \( \rho \text{C}_2 \) possible chromosome pairings, \( \rho \text{C}_3 \) possible chromosome triples, and, in general, \( \rho \text{C}_n \) possible chromosome \( n \)-tuples for \( n < \rho \), in spite of the fact that not all \( n \)-tuple collections of chromosomes would meet the complementary index set constraints specified by the algorithm, it would be necessary to construct on the order of \( \rho \text{C}_0 + \rho \text{C}_1 + \rho \text{C}_2 \ldots + \rho \text{C}_\rho \) schemas to achieve this level of recombination distance calculation precision. Nevertheless, the approach does ensure that if the recombinational distance between \( \alpha \) and \( \beta \) is assigned a value \( \Delta_{\xi\chi} \), then \( \beta \) can be produced from \( \alpha \) in as few as \( \Delta_{\xi\chi} \) recombinations with the population, for \( \Delta_{\xi\chi} \in [0, \lambda] \).

*The variable* \(<\text{population}>\) *is the unique set of chromosomes for recombination.*  
*The variable* \(<\text{origin}>\) *is the chromosome from which the distance is measured.*  
*The variable* \(<\text{target}>\) *is the chromosome to which the distance is measured.*  
*The variable* \(<\text{schema}>\) *is the array of highest order schemata used to compute distance.*

\[
\text{measure-recombinational-distance}
\]
\[
(\text{<population>}, \text{<origin>}, \text{<target>})
\]

\[
\text{if } \text{<origin>} = \text{<target>}
\]

\[
\text{return } 0
\]
\[
\langle \text{schema} \rangle(0) = \langle \text{origin} \rangle \\
\text{for } \iota \in \langle \text{population} \rangle \\
\quad \langle \text{schema} \rangle(0) = \text{expand-schema}(\langle \text{schema} \rangle(0), \iota) \\
\quad \text{if } \text{<target>} \in \langle \text{schema} \rangle(0) \\
\quad \quad \text{return } 1 \\
\quad \text{else} \\
\quad \quad \langle \text{schema} \rangle(0) = \langle \text{origin} \rangle \\
\langle \text{schema} \rangle(1) = \langle \text{origin} \rangle \\
\text{for } (\iota_1, \iota_2) \in \langle \text{population} \rangle \\
\quad \text{if } \neg (\iota_1 \equiv \iota_2) \\
\quad \quad \langle \text{schema} \rangle(1) = \text{expand-schema}(\langle \text{schema} \rangle(1), \iota_1) \\
\quad \quad \langle \text{schema} \rangle(1) = \text{expand-schema}(\langle \text{schema} \rangle(1), \iota_2) \\
\quad \quad \text{if } \text{<target>} \in \langle \text{schema} \rangle(1) \\
\quad \quad \quad \text{return } 2 \\
\quad \quad \text{else} \\
\quad \quad \quad \langle \text{schema} \rangle(1) = \langle \text{origin} \rangle \\
\ldots \\
\text{return } \infty
\]
It is evident that the preceding algorithm must, in the worst-case, test whether or not the target genotype is a member of each of the $2^\rho$ possible schemata, with each test requiring an operation with an $O(\lambda)$ time complexity, entailing that the overall time complexity would be $O(\lambda 2^\rho)$ and, thus, nonpolynomial. Although this approach closely parallels the intuition surrounding how a measure of recombinational distance must be calculated, it should be noted that it is possible to find an upper bound on the number of population members necessary to expand the origin genotype into a schema that contains the target genotype (and, in so doing, estimate the recombinational distance) with a polynomial time algorithm. A greedy approach is summarized in the pseudocode below and the process by which this approach estimates (and, potentially, overestimates) the recombinational distance from origin genotype ■■■■■■■ to target genotype □□□□□□□ is depicted in Figure 30, for the population {■□□■■■□, □■■■■■■, ■■■□□□□■, ■■■■■■■□}

*The variable `<population>` is the unique set of chromosomes for recombination.*

*The variable `<origin>` is the chromosome from which the distance is measured.*

*The variable `<target>` is initially the chromosome to which the distance is measured and is also used to specifying the remaining genotypic data to be exchanged.*

*It should also be noted that although specify-remaining-values is structurally similar to expand-schema (introduced previously), the ternary string `<target>` does not constitute a schema – it represents the values of the target that have not yet been matched.*
improved-measure-recombinational-distance

(<population>, <origin>, <target>)

if <target> == <origin> then return 0

for ι = 1 to λ
    if <target>[ι] == <origin>[ι] then <target>[ι] = '_'

for δ = 1 to |<population>|
    <best> = find-best-match(<population>, <target>)
    <target> = specify-remaining-values(<target>, <best>)
    <population>.remove(<best>)
    if <target>[x] = '_' for 1 ≤ x ≤ λ then return δ

return ∞

find-best-match (<population>, <target>)

for <member> ∈ <population>
    if Δ_Hamming(<member>, <target>) < Δ_Hamming(<best>, <target>)
        <best> = <member>

return <member>

specify-remaining-values (<target>, <chromosome>)

for ι from 0 to λ
    if <target>[ι] == <chromosome>[ι] then <target>[ι] = '_'

return <target>
Figure 30. A sequence of diagrams corresponding to the proposed greedy enhancement to the recombinational distance measurement approach introduced previously.
It should be noted that the task of measuring the recombinational distance from an origin genotype to a target genotype is essentially equivalent to the task of identifying the smallest set of population members (and the cardinality thereof) that can be used to expand the origin into a schema that contains the target. This would indicate that the task of computing the recombinational distance could be restated as an instance of the well-known set cover optimization problem [Chvatal 1979]: for a set of elements designated to be the universe and n sets for which the union of these sets is equivalent the universe, the set cover decision problem is to determine if there is a group of k or fewer of these sets for which the union still contains the universe, and the set cover optimization problem is to determine the smallest number of sets for which the union is still equivalent to the universe. Consider, as an example, the universe of entities \{a, b, c, d, e\} and the sets \{a, b, c\}, \{b, d\}, \{c, d\}, and \{d, e\}, whose union is equivalent to the universe – the correct solution to this set cover optimization problem would be the sets \{a, b, c\}, \{d, e\}\} and the solution to the set cover decision problems for k = 2 and k = 1 would be true and false, respectively. Where the latter (i.e., the decision problem) is known to be NP-complete, the former (i.e., the optimization problem) is known to be NP-hard.

The mapping of the measurement of recombinational distance to an instance of the set cover optimization problem is most easily demonstrated by example – for a pair of genotypes designated as the origin and the target, the set of indices between which these genotypes differ is designated as the universe, and each of the sets composing the universe corresponds to a population member such that the set contains only those indices at which the values of the population member are necessary to transform a genotype into
the target. If $A$ is a set of population members considered (according to some algorithm) for the recombination-defined traversal from origin to target, then we can define the $i^{th}$ element $a_i$ of the set $A$ as a non-empty set of indices at which the value of the corresponding population member is the same as the value of the target genotype. For sample origin and target genotypes ■■■□□□ and □□□□□□, respectively, the set $D$ of indices at which they differ (where $D = \{1, 2, 3, 4, 5, 6\}$), and population {□□■■■■, ■■□□□■, ■■■□□■, □□■■■□}, the elements of $A$ (i.e., $a_1$, $a_2$, $a_3$, and $a_4$) would be $\{1, 2, 3\}$, $\{3, 4, 5\}$, $\{5, 6\}$, and $\{1, 2, 6\}$, respectively. It is thus apparent that the task of determining the smallest set of population members necessary to effect the transformation (using recombination) from the origin genotype to the target genotype is equivalent to the task of solving the set cover optimization problem for the universe $D$ using the sets of $A$. Since (as stated previously) the time complexity of the set cover optimization problem is known to be NP-hard, it must (unless $P = NP$) be necessary to use an algorithm with exponential time complexity to measure recombinational distance precisely in all situations.

For the example in the preceding paragraph, the greedy algorithm would have selected the population members in the sequence □□■■■■, ■■□□□■, and ■■■□□■, evaluating the recombinational distance between the origin and target genotypes to have a value of three. Naturally, this is an overestimate of the recombinational distance, as the two members ■■□□□■ and □□■■■□ would also have be sufficient to transform the origin into the target using recombination alone. However, it should be noted that if a subset of the population that can effect the transformation does exist, then the greedy algorithm
(which doesn't terminate until a sequence containing the entire population is considered) is guaranteed to provide a covering set. This entails that the greedy approach described previously will never underestimate the recombinational distance that separates two genotypes and thus constitutes an upper bound that can be computed in polynomial time. It should also be noted that the nonpolynomial time recombinational distance measurement algorithm could always be terminated after considering every tuple up to an arbitrary size (at the discretion of the user) to provide a lower bound on the recombinational distance between the origin and target genotypes. Consequently, although the proposed recombinational distance measure had a nonpolynomial time complexity, it is possible to generate both an upper and a lower bound on the recombinational distance in polynomial time.

Having established that the measurement of recombinational distance is largely equivalent to the set cover problem, it becomes apparent that the greedy algorithm introduced previously would be directly related to the known greedy heuristic that is used to solve the set-cover problem. It is, thus, reasonable to speculate that some of the many heuristics designed for more general purpose application to the set cover problem [Chvatal 1979, Caprara, Fischetti, and Toth 1995, Haddadi 1997, Feige 1998] could be adapted to measure recombinational distance. It should also be noted that apriori knowledge that the fitness function being solved can produce regularities in the genotypes of an evolving population might be exploited such that some heuristic techniques will exhibit superior performance to others. To conclusively demonstrate this would require further investigation.
It should be noted, at this point, that an approach to recombinational distance measurement defined with respect to the simpler, one-point recombination operation (often associated with the most rudimentary instances of the genetic algorithm) could not be expected to yield any substantial improvement over the time complexity of the approach introduced previously (using uniform recombination). The approach introduced in this chapter must, in the worst-case, substantiate whether the target genotype is a member of set of possible offspring defined by the application of uniform recombination to the origin genotype and the genotypes from a n-tuple \((1 \leq n \leq \rho)\) of the current population – although it might be possible to improve the test for the membership of the target in the set of possible offspring (i.e., the term corresponding to a time complexity of \(O(\lambda))\), it is emphasized that the set of possible offspring associated with one recombination operation cannot be assumed to provide any information about the set of possible offspring associated with a different recombination operation.

Consequently, in the worst-case, it would still be necessary to test each of the sets of possible offspring defined by a tuple of the population and, thus, there is no apparent approach by which the \(O(2^\rho)\) time complexity could be reduced, regardless of the manner in which the binary recombination operation is to be performed. Furthermore, even if apriori knowledge of the regularities introduced by a specific fitness function was available, it cannot be assumed that this knowledge could be used to reduce the number of population members below a level that would reduce the exponential component of the overall time complexity – such a claim would require further investigation.
3.10 Additional Time Complexity Considerations

It was previously noted that the approach to recombinational distance measurement that tests (with an $O(\lambda)$ time complexity operation) whether or not the target genotype is a member of each of the $2^\rho$ schemata that are defined by the population has a nonpolynomial time complexity of $O(\lambda 2^\rho)$. Thus, it remains to demonstrate the improvement in time complexity over the previous approach that can be achieved using the proposed greedy heuristic. Since the find-best-match algorithm must perform a Hamming distance calculation (i.e., a $\Theta(\lambda)$ operation) for each of the remaining population members (ranging from $\rho$ members at the first execution to 1 member for the final execution), find-best-match must have an $O(\rho \lambda)$ worst-case time complexity. Additionally, since specify-remaining-values has an $O(\lambda)$ time complexity, and both the find-best-match and specify-remaining-values algorithms are executed after each new population member is considered (up to $\rho$ times, in the worst-case), the overall time complexity of the proposed greedy approach is $O(\rho \lambda^2)$. Although this is a significant improvement over the time complexity of the previous approach, it is again emphasized that the greedy approach can be expected to overestimate the recombinational distance between genotypes. It should also be noted that the time complexity of the greedy heuristic approach to recombinational distance measurement can be more readily compared to the $\Theta(\lambda)$ time complexity of the mutational distance measurement approach based on the Hamming distance. Although the worst-case time complexity is increased by a factor of $\rho \lambda$, the proposed recombinational distance measure considers many more
traversals of genotypic space. Furthermore, it may be possible to reduce the worst and average-case time complexity of the greedy algorithm itself, through the use of optimizations such as the removal of those genotypes from the population that would contribute the same values to the remaining genotypic spatial distance to be traversed. Although this warrants further investigation, it should also be noted that since the greedy approach to recombinational distance measurement could not conceivably be measured without considering the value at each index of every population member, it can be conjectured that the complexity of such a heuristic could not be reduced past \( O(\rho \lambda) \), and the existence of such an approach remains an open problem.

### 3.11 Operator Representative Distances Summary

This chapter began by demonstrating that the development of a distance measure that is representative of the manner in which genotypic space is traversed by the unary mutation operator is not difficult. In fact, it has been demonstrated that the Hamming distance measure in widespread use is such a measure - the Hamming distance between genotypes accurately conveys the number of mutation operators that are necessary to transform either one into the other. Thus, the notion of operator traversal representativity has been introduced - for a measure (of the distance between genotypes \( \alpha \) and \( \beta \)) to be representative of the traversal specific to an operator (or a mechanism defined by a set of operators), the measure must reflect the number of applications of the operator (or mechanism) that are necessary to effect the transformation of \( \alpha \) into \( \beta \).
Although representativity in terms of traversals effected by a unary operator was not difficult to achieve, the genetic algorithm also employs a binary recombination operator. This operator is not reflected by the Hamming distance and, as a result, the improper use of the Hamming distance in constructing informative graphics can be detrimental. To develop a measure of the distance traversed by uniform recombination, it was necessary to redefine the binary recombination operator as a fixed set of unary recombination operators defined by the current population. This is conceptually similar to the use of a tangent in predicting the motion of an object following a curve - although the population of a genetic algorithm is expected to change over time, it is not unreasonable to expect the elements of a stable population to proceed along a direction in genotypic space that is similar to that defined if the current population was fixed. This follows from the fact that a genotype being recombined with elements of a converged population will proceed exactly along the direction in genotypic space defined by the current population until additional variation is introduced. The introduction of the unary recombination paradigm allows for both mutation and recombination to be treated as operators that traverse the same genotypic space - facilitating novel approaches to the integration of these measures.

Arguably the most important subtask of the recombinational distance measures proposed in this chapter was identified as the substantiation of whether or not one genotype can create another in the presence of a specific set of other genotypes with which recombination can occur. The approaches associated with this subtask were thoroughly explored and the low and high resolution recombinational distance measures were introduced. The low resolution measure would assign a recombinational distance value
from the set \( \{0, 1, \infty\} \) depending upon the number of recombination operation necessary to transform the origin genotype to the target genotype. The high resolution measure, on the other hand, would assign values from the set \([0, \lambda] \cup \{\infty\}\) that are based on the size of the subpopulation necessary to define a sufficient number of unique unary recombination operations to transform the origin genotype to the target genotype. The integrations of mutational distance with these approaches to recombinational distance measurement will result in two novel measures that are representative of the traversal of genotypic space by the complete explorative component of the evolutionary mechanism of the genetic algorithm. As indicated by Figure 31, the processes by which integrated measures are derived is the focus of the next chapter.

Figure 31. The upper portion of the summarizing chart from Figure 15. The next chapter will explore the integration of mutational and recombinational distances to derive a measure of the distance traversed by the complete explorative component of the evolutionary mechanism of the genetic algorithm.
Chapter 4

Mechanism Traversal Representativity

Overview

Although the Hamming distance is suitable as a measure of the mutational distance between genotypes, it is recognized that Hamming distance does not reflect the number of recombination operations necessary to produce one genotype as the descendent of another. To address this, the previous chapter established plausible approaches to recombinational distance measurement, and having defined both recombinational and mutational distance measures, it is the aim of Chapter 4 to explore approaches by which these measures can be combined. Since the variational component of the evolutionary mechanism is defined entirely by the mutation and recombination operators, the combination of these operator traversal representative measures (i.e. mutational and recombinational distance measures) could be used to define a mechanism traversal representative measure, herein referred to as an integrated variational distance measure.
This chapter begins by exploring the move set and transition probability approaches by which the conceptualizations of distance that facilitate the creation of adaptive landscapes are defined. It is demonstrated that the task of combining move set and transition probability conceptualizations of distance is nontrivial, and is further complicated by the difficulty associated with defining a move set conceptualization of mutational distance or a transition probability conceptualization of recombinational distance. It is also demonstrated that the standard approaches to combination (i.e. arithmetic, geometric, harmonic, quadratic, etc.) cannot be used to combine measures of mutational and recombinational distance into a value that is actually representative of the genotypic spatial distance that could be traversed using both mutation and recombination in tandem.

Instead, an alternative approach is proposed where recombination is conceptualized as the operator that provides a platform for the efficient traversal of genotypic space by the mutation operator. This approach is analogous to the traversal of a city using a mass transit system that is, itself, comprised of a subway system and a network of bus routes. A traveller wishing to traverse the distance between two locations will most probably require both the subway and the bus to efficiently traverse the distance. Consequently, a count of the number of subway stops between the two locations cannot be combined with a count of the number of bus stops between the two locations to compute an accurate count of the number of total stops necessary. Contrarily, it is expected that the traveller will use the subway system first to traverse as much of the distance as possible, and then use the network of buses to reach the destination.
Under the reasonable assumption that the genetic algorithm works similarly (with the subway system and bus networks corresponding to the recombination and mutation operators, respectively), an approach to computing the integrated variational distance between genotypes would consider the mutational distance from each genotype that can be reached by the recombination operator. This chapter proposes a possible approach by which a likelihood can be assigned to each path and ultimately combined to create a representative measure of integrated variational distance. A gross oversimplification of this approach is also introduced, wherein a mutational distance from the centroid of the population is computed as an approximation of the genotypic spatial distance that would be traversed using both operators. Although this approach was not expected to perform as well, it represents a natural and logical approach to measuring integrated variational distance, and since it can be evaluated very quickly it may prove to be more viable than the Hamming distance when complexity restraints preclude a fully representative approach. Since the problem of combining mutational and recombinational distance measures is in its infancy, and with the consideration that the primary motivation of this body of research was the development of a measure that could be employed to enhance the representativity of the adaptive landscape visualization, the novel measures proposed in this chapter were developed to assign ranks according to genotypic spatial distance, and not actual distance values.

The three novel approaches developed in this chapter (i.e. the mutational distance from the population centroid and the use of the low resolution and high resolution measures of the recombinational distance to the points in genotypic space from which mutation can be
used to traverse the remaining distance) are each an alternative to the use of the Hamming distance for genotypic spatial distance measurement. The novel approaches vary in complexity, computational expense, and achievable distance resolution, and although they are discussed extensively in this chapter, they will not be comparatively evaluated (against each other and against the Hamming distance) until Chapter 6, following the introduction of a novel evaluation methodology in Chapter 5.

4.1 Simplistic Distance Measure Combinations

As noted previously, the point mutation operator is specified by a single mutation rate parameter that determines the probability with which each binary value in a chromosome may (independently) be flipped into a complementary value. It was demonstrated by Vose (1999) that the use of this mutation operator definition entails that the mutation of two offspring chromosomes produced by recombining $\alpha$ and $\kappa$ is functionally equivalent to the recombination of two chromosomes $\alpha'$ and $\kappa'$ that were produced when chromosomes $\alpha$ and $\kappa$ were mutated, respectively. Equivalently, the probability that a chromosome $\beta$ will be produced as the offspring when recombining mutated chromosomes is the same as the probability that $\beta$ will be produced by mutating the offspring of a recombination operation. This entails that the mutation and recombination components of the explorative mechanism are actually commutative. This commutativity is fortunate because it does not preclude a general expression for the application of this explorative mechanism and, furthermore, it should be possible to measure the distance between chromosomes with respect to the traversals of this complete mechanism.
It was also noted previously that a conceptualization of distance is typically specified using either a set of transition probabilities or a set of move sets. Consequently, it is reasonable to assume that a complete explorative mechanism traversal distance measurement could be constructed by combining either move set or transition probability defined measures of mutational distance with either move set or transition probability defined measures of recombinational distance. However, since a move set conceptualization of the neighbourhood of a chromosome can only be used to determine whether the transition probability from that chromosome to another is zero or nonzero, attempting to combine a move set conceptualization of mutation with a transition probability conceptualization of recombination (or, a move set conceptualization of recombination with a transition probability conceptualization of mutation) would not be produced. Instead, it suffices to explore approaches to the combination of two move set defined conceptualizations or two transition probability defined conceptualizations.

If a measure of the distance in genotypic space between chromosomes \( \alpha \) and \( \beta \) that would be traversed by operator \( \omega_1 \) is represented using a move set conceptualization of distance, then \( \beta \) has membership in the move set of \( \alpha \) associated with operator \( \omega_1 \) if and only if the operator \( \omega_1 \), applied to chromosome \( \alpha \), could produce chromosome \( \beta \). Furthermore, the membership of \( \beta \) in the move set of \( \alpha \) associated with operator \( \omega_1 \) has no bearing whatsoever on the membership of \( \beta \) in the move set of \( \alpha \) associated with a different operator \( \omega_2 \). Thus, a move set representation of the distance in genotypic space between chromosomes \( \alpha \) and \( \beta \) that could be traversed using either operator \( \omega_1 \) or operator \( \omega_2 \) would naturally be the union of the move set representation of the distance between \( \alpha \)
and $\beta$ as traversed by $\omega_1$ with the move set representation of the distance between $\alpha$ and $\beta$ as traversed by $\omega_2$. With the operators denoted $\omega_1$ and $\omega_2$ being the mutation and recombination operators, respectively, it is obviously possible to derive a move set representation of the explorative mechanism traversal distance by applying a union to the both the move sets defined by mutation and the move sets defined by recombination.

Unfortunately, it is easily demonstrable that a move set conceptualization is not suitable for the conceptualization of the distance between chromosomes as would be traversed by the typical point mutation operator discussed previously. It was noted that a move set representation can be treated as a relaxation of a transition probability representation such that, for ordered chromosome pairing $(\alpha, \beta)$, $\beta$ is a member of the move set of $\alpha$ if and only if the transition from $\alpha$ to $\beta$ is associated with a nonzero probability. Since the transition probability associated with the mutation operator has already been defined by the positive definite function $(\lambda^{-1})^\delta \cdot (1 - \lambda^{-1})^{\lambda - \delta}$, between two binary strings of length $\lambda$ between which there is a Hamming distance of $\delta$, there does not exist an ordered chromosome pairing $(\alpha, \beta)$ associated with a transition probability of zero. As such, for every ordered pair of chromosomes $(\alpha, \beta)$, $\beta$ is always assigned membership to the move set of $\alpha$ associated with the mutation operator. It is obviously that this is completely unsuitable as a useful representation of the traversable distance between chromosomes. Fortunately, the relationship between move sets and transition probabilities can be 'tightened' such that, for an ordered pair of chromosomes $(\alpha, \beta)$, $\beta$ is assigned membership to a set denoted the $\theta$-likely move set of $\alpha$ if and only if the transition from $\alpha$ to $\beta$ is associated with a probability greater than the threshold value $\theta$. This would
indicate that it should be possible to effectively combine a move set defined conceptualization of the recombinational distance between two chromosomes with the $\theta$-likely move set defined conceptualization of the mutational distance between the same two chromosomes.

No transition probability defined conceptualizations of the genotypic spatial distances defined with respect to the recombination operator have yet been established, but the justification for this omission will not be introduced until Chapter 5. Two alternative approaches, however, were introduced in the previous chapter for computing the relative distance (and, it follows, a relative likelihood) for a traversal genotypic space by recombining a chromosome $\alpha$ with a chromosome $\kappa$ from the current population to produce a chromosome $\beta$. Although the range of possible values associated with these approaches can of course be normalized to the interval (0, 1), facilitating combination with the transition probability values associated with mutation, it is not difficult to demonstrate that a representative measure of the distance between two chromosomes that would be traversed by the explorative component of the evolutionary mechanism cannot be achieved through a simple (i.e. arithmetic, geometric, harmonic, or quadratic) combination of the measured recombinational distance and the measured mutational distance. Figure 32 depicts the traversal of the genotypic spatial distance between chromosomes ■■■ and □□□ with respect to an operator that flips exactly one bit with each application and a recombination operator defined by a singleton population containing chromosome ■□□. Although the mutational distance (i.e. the Hamming distance, in this case) between chromosomes ■■■ and □□□ is three, any of the
recombinational distance measurement approaches introduced in the previous section would yield a conclusion that the distance between these chromosomes is infinite, as recombination with chromosome ■□□ cannot effect the transformation of ■■■ into □□□, regardless of the number of applications of the operation. However, Figure 32 clearly indicates that only two operations (one recombination and one mutation) are actually needed to traverse the genotypic spatial distance from ■■■ to □□□.

Figure 32. Although the mutational and recombinational distances from ■■■ to □□□ (defined by a singleton population {■□□}) are three and ∞, respectively, only one mutation and one recombination operation are actually required to effect the transformation from ■■■ to □□□.
It can be deduced (in part from the example in Figure 32) that the mutational and recombinational distances cannot be considered independently for the development of a representative measure of the distance that is traversed through genotypic space by the complete explorative mechanism. It is necessary, instead, to consider the offspring chromosomes of one of the operators as providing a set of possible origin chromosomes for the second operator, thus deriving a set of paths that represent the traversal of genotypic space by the complete explorative mechanism. As mutation has a considerably larger distribution of possible offspring chromosomes, and the recombinational distance measures introduced in the previous chapter are (comparatively) more computationally expensive, the proposed complete explorative mechanism traversal distance measures would take the form of the mutational distance from those chromosome that could be the result of a recombination operation, adjusted with respect to the recombinational distance to each such chromosome.

### 4.2 Simplistic Population Centroid Approach

Although the previous chapter thoroughly explored conceptualizations of both mutational and recombinational distance, with varying degrees of resolution, with respect to the genetic algorithms traversal of genotypic space by recombination with one \((\Delta_{x_1})\) or more \((\Delta_{x_2})\) chromosomes in the current population, there is a simplistic approach to the measurement of genotypic spatial distances by the complete explorative mechanism that, although not a derivative of any approach introduced in the previous chapter, should not be omitted. Since, in its most basic form, recombination is an operator that separates a
population of chromosomes from an offspring chromosome, an intuitive alternative to
exploring every recombination operation defined by the population is to replace the
population with a single chromosome most representative of the elements of the
population. This would, naturally, be equivalent to first defining the notion of a
population centroid and then treating recombination as a binary operator with one of the
operands always having the value of the centroid of the population. To clarify, although
the recombinational distance $\Delta_{\xi\chi}$ between $\alpha$ and $\beta$ with current population $P$ was
computed in the previous sections by searching the results of recombination operations
$\chi(\alpha, p_1), \chi(\alpha, p_2), ..., \chi(\alpha, p_\rho)$ (for each of the $\rho$ unique population members in $P$), in
order to determine the likelihood with which chromosome $\beta$ might appear, a population
centroid approach would need only search recombination operation $\chi'(\alpha, \bar{p})$, where $\bar{p}$ is
the centroid of population $P$.

There are four obvious methodologies for defining the centroid of a population of
chromosomes from a binary alphabet, inspired by the three most common statistical
approaches to the representation of a population using a single value – the mean, the
median, and the mode. The mode centroid $\bar{p}_{\text{mode}}$ is, naturally, the most frequently
occurring member of the population, and the median centroid $\bar{p}_{\text{median}}$ could be easily
determined by sorting the binary chromosomes by the actual numerical values they
represent (as opposed to their phenotypic values). The remaining two methodologies for
computing the centroid of a population would entail the computation of the mean value of
the population for every possible index of a genotype, and expressing these mean values
either as a string of binary digits or as a string of real values, median centroids $\bar{p}_{\text{mean-B}}$ and $\bar{p}_{\text{mean-R}}$, respectively.

Although each of the preceding definitions could be reasonably applied to determine the centroid of a population, it is relatively simple to demonstrate that a recombinational distance computing using a centroid as a substitute for the actual elements of the population can yield results that are counter-intuitive. For a population of ten binary chromosome comprised of ■■■□, ■■□■, ■□■■, and □■■■, in concentrations 4/10, 3/10, 2/10, and 1/10, respectively, the four centroids, $\bar{p}_{\text{mode}}$, $\bar{p}_{\text{median}}$, $\bar{p}_{\text{mean-B}}$, and $\bar{p}_{\text{mean-R}}$, would have the values ■■■□, ■■□■, ■□■■, and 0.1 0.2 0.3 0.4, respectively. It is immediately obvious that the $\bar{p}_{\text{mode}}$ and $\bar{p}_{\text{median}}$ centroids are ignoring 60% and 70% of the population, respectively, and that $\bar{p}_{\text{mean-B}}$ is not at all representative of the population. Although centroid $\bar{p}_{\text{mean-R}}$ would appear to have retained more information than the other centroids, a simple sum of the chromosome component differences from the centroid suggests that an approach employing this centroid will also yield results that are counter intuitive.

Although the distance measurement result that chromosomes ■■■□ and ■■□■ are considered at recombinational distances 1.2 and 1.4 from the centroid, respectively, is not counterintuitive, other chromosomes ■□■■ and ■■□□ are assigned distances 1.8 and 1.6, respectively. This result is definitely contrary to the desired behavior, since chromosome ■□■■ is actually contained in the population while chromosome ■■□□ is not, and yet the former is being assigned a larger recombinational distance than the latter. Although the centroid derived approaches to recombinational distance measure were originally
included only to ensure that recombinational distance measurement was thoroughly investigated, these approaches can be adapted to include consideration for traversal by the mutation operator.

Although the representativity of the recombinational distance component of this centroid approach is overly simplistic, it must be observed that the typical Hamming distance approach to genotypic spatial distance measurement includes no consideration for recombination whatsoever. Consequently, the use of the centroid of the population to estimate the function of recombination may yield an approach to measurement that is more representative of the traversal of genotypic space by the explorative mechanism than any approach currently being used.

4.3 Integration of Recombinational Distance

It is most reasonable to begin the development of a measure expected to be representative of both mutation and recombination by exploring those measures designed to be representative of either mutation or recombination. Essentially, this is the development of an approach representative of traversals by the complete mechanism by exploring approaches representative of traversals by the component operators. To this end, Figures 33 and 34 depict the assignment of ranks to the genotypes of four binary digits according to their distance from a designated origin chromosome (■■■■), as computed by the Hamming distance measure and the two recombinational distance measures introduced in Chapter 3, respectively. Since these measures had distance resolution values (as defined
in Section 3.9) of 3 and 2 + λ respectively, they will be referred to as the low resolution and high resolution recombinational distances measures, respectively.

As it was noted previously that no reasonable measure of recombinational distance can be made without considering those chromosomes that comprise the current population, a nonuniform distribution of two unique chromosomes in a population of cardinality three is assumed. The population \{■■■□, ■■□■, ■■■■} is of sufficient diversity to ensure that the differences between the rankings generated by each of the approaches are evident.

![Diagram of chromosome ranking](image.png)

**Figure 33.** When chromosomes are ranked according to their Hamming distance from the origin (■■■■), those genotypes that can be most easily reached from the origin by the application of mutation are assigned the best ranks.
Figure 34. When chromosomes are ranked according to the low and high resolution recombinational distances (top and bottom, respectively), only those chromosomes that can be produced by recombining the origin (■■■) with one or more members of the population are assigned finite distance values.
As noted previously, the application of the typical mutation and recombination operators of the genetic algorithm have been demonstrated (Vose 1999) to be commutative operations. It was also noted that the combination of the relatively lesser range and relatively greater computational expense associated with the recombination operation over the mutation operation would indicate that a measure of the distance traversed by the complete explorative mechanism would be most practically determined using a measure of the mutational distance from each chromosome that is not considerable unreachable with respect to the recombination operation, scaled with respect to the finite recombinational distance assigned.

To this point there have been three novel approaches proposed for measuring recombinational distance: the simplistic centroid approach described in earlier in this chapter and the low and high resolution recombinational distance measurement approaches (with differing resolutions and orders of complexity), first described in Chapter 3. It stands to reason, then, that given the unary arity of the mutation operation and the apparent ease with which accurate and representative measures of mutational distance can be calculated, there are three approaches to the measurement of the distances traversed by the complete explorative component of the evolutionary mechanism that must be explored and evaluated – one for each of the recombinational distance measures proposed. However, the approaches developed for the calibration and comparative evaluation of different measures of genotypic spatial distance (by which the merits and drawbacks of each approach could be demonstrated) were sufficiently complex to warrant separate (and extensive) treatment, and are thus deferred to Chapter 5.
Of the proposed approaches to the measurement of the genotypic spatial distances traversed by the recombination, the use of a centroid is the most simplified and computationally inexpensive approach. As such, the first approach to genotypic spatial distance measurement (with respect to the complete explorative mechanism) to be investigated is the mutational distance from the population centroid. However, as the Hamming distance is not defined between one binary and one real valued string, the mutational distance from any origin to the centroid of the population $\bar{p}_{\text{mean-R}}$ cannot be computed using the Hamming distance. Instead, this distance is computed using the well-known Euclidean measure. This approach is depicted in Figure 35.

![Diagram]

**Figure 35.** The representation (from Figure 28) for the recombination of ■■■■ with population {■■□, ■□■}, with the Euclidean distance from the population centroid.
Although this approach would address some of the issues associated with the Hamming distance, it is not difficult to recognize that if the centroid is highly dissimilar from the origin chromosome (from which distances are measured), the distance from the centroid is not representative of short traversals. To clarify, consider the traversal from an origin to itself – although the distance should be assigned a value of zero, the Euclidean distance from a centroid that is highly dissimilar from the origin would not be representative of the actual, zero-length traversal required. This particular case is easily addressed by including the origin in the computation of the centroid (as depicted in Figure 36), but it does not address the lack of representation for very short (but nonzero) length traversals.

Figure 36. The weighted digraph from Figure 35 with the position of the centroid adjusted to include the origin chromosome ■■■■.
It is difficult to conceptualize further improvements to an approach that computes recombinational distance using a centroid alone. The low and high resolution approaches to recombinational distance measurement, on the other hand, do not represent a population as a single point in the continuous space with which genotypic space is contained. As a result, each of the traversals of genotypic space that are possible with the mutation operator (originating from each chromosome that can be reached by recombination) can be combined without dismissing the perceived recombinational distance to each reachable point, as depicted in Figure 37.

Figure 37. The Euclidean distance from each chromosome that can be reached by recombining the origin with members of the population is depicted using the undirected dotted lines. The solid lines represent traversals by recombination alone.
The proposed integration of mutational and recombinational distance measurement approaches would then be the combination of the recombinational distances from the origin chromosome to any chromosome reachable by recombination with the mutational distances from each of these reachable chromosomes to the target chromosome. Such an approach would reflect that recombination with chromosomes that are very different from the origin chromosome often result in offspring chromosomes that are a greater mutational distance from their parent chromosomes than could be easily achieved with mutation alone. In this way, recombination could be conceptualized as providing a possible shortcut for the traversals of genotypic space that use the mutation operator.

It follows logically that there is also a complementary conceptualization of mutation as an operation that increases the number of possible operand chromosomes for the recombination operator. This would be equivalent to the conceptualization that it is mutation that provides a possible shortcut for traversals that use the recombination operator alone. However, the relative ease with which mutational distance can be calculated (over recombinational distance) would suggest that the approach described with the former conceptualization (i.e. that recombination provides a platform for the traversals of mutation) would be more practical than the latter at this time.

4.4 Integrated Variational Distance Definitions

It is implicitly assumed by the proposed measurement approach that the ranking of a chromosome \( \beta \) according to its genotypic spatial distance from a chromosome \( \alpha \) can be
expressed using a weighted average of a measure of the ease with which chromosome $\alpha$ can reach an intermediate chromosome $\kappa$ using recombination and a measure of the ease with which that same chromosome $\kappa$ could then be mutated into chromosome $\beta$, for every chromosome $\kappa$ that can be reached through recombination alone. The weightings used for this average are derived in an effort to represent the viability of traversing the distance from $\alpha$ to $\beta$ using any number of recombination operations (between zero and the number of unique chromosomes in the population).

It is crucial to emphasize that the approaches described using the low and high resolution recombinational distance measures represent only an initial approach to the measurement of the genotypic spatial distances that are traversed by the explorative mechanism. There is little doubt that a more complex calculation could achieve a more representative measure, but time complexity remains a consideration of significance. At this time the objective is only to develop integrated measures (based on the low and high resolution recombinational distance measure) that are more representative than a Hamming distance approach and at least as representative as the approach introduced previously using the Euclidean distance from the population centroid. A comparative evaluation of these measures according to their relative representativity will be addressed in Chapter 6 after the evaluation methodology in Chapter 5 has been introduced and thoroughly explored.

To illustrate the process by which the proposed measures will be calculated, consider the genotypic spatial distance from one chromosome $\alpha$ to another chromosome $\beta$. For this example, this distance will be traversed using point mutation and uniform recombination
with members from a population of three chromosomes \(\{\kappa_1, \kappa_2, \kappa_3\}\). As noted previously, the use of the Hamming distance between chromosomes \(\alpha\) and \(\beta\) is equivalent to the assumption that the recombination operation will never be applied in the traversal of the genotypic space between \(\alpha\) and \(\beta\). Although this is certainly not a reasonable assumption to be made in general, it is possible (in any specific instance) that the recombination operator will not be applied at all in the traversal of the space. However, if a single unary recombination operation (defined by unique population members \(\kappa_1, \kappa_2, \text{and } \kappa_3\)) is applied to chromosome \(\alpha\), the offspring will belong to one of the three highest order schemas defined by the pairings of chromosomes \(\{\alpha, \kappa_1\}, \{\alpha, \kappa_2\}, \text{or } \{\alpha, \kappa_3\}\). In this case it could be concluded that the remaining distance to be traversed would be the minimum of the mutational distances from each of these schemata.

It must be noted that that the originating chromosome A may not have many values in common with any chromosome that belongs the population. Consequently, the highest order schema defined by the pairing of chromosomes \(\{\alpha, \kappa_1\}\) may, in fact, have relatively low order overall. As this would considerably weaken the ability of the proposed approach in determining a reasonable ranking of chromosomes according to their genotypic spatial distance from a single point, the mutational distances from each of the schemata defined by \(\{\alpha, \kappa_1\}, \{\alpha, \kappa_2\}, \text{or } \{\alpha, \kappa_3\}\) are replaced by the mutational distance from each of the schemata defined by only \(\{\kappa_1\}, \{\kappa_2\}, \text{or } \{\kappa_3\}\). Furthermore, since the concept of the distance between a schema and a chromosome is somewhat vague, a simplistic approach was developed - with each schema identified by a ternary string from the alphabet \(\{\text{■}, \square, \ast\}\), the distance from a schema to a chromosome can be reasonably
computed as the sum of the number of indices at which the chromosome does not match the schema.

To this point, under the assumption that no more than a single recombination operation will be employed, the proposed approach will compute the genotypic spatial distance between two chromosomes \( \alpha \) and \( \beta \) as the weighted average of the mutational distance from \( \alpha \) to \( \beta \) and the minimum of the mutational distances to \( \beta \) from \( \kappa_1 \), \( \kappa_2 \), and \( \kappa_3 \). To reduce this assumption (i.e. that no more than one recombination will occur) to an assumption that no more than two recombinations will occur, the minimum of the mutational distance to \( \beta \) from each of the highest order schemata defined for each unique subset of the population with cardinality two (i.e. \( \{ \kappa_1, \kappa_2 \} \), \( \{ \kappa_1, \kappa_3 \} \), or \( \{ \kappa_2, \kappa_3 \} \)) must also be included. It follows that the assumption can be subsequently weakened using each of the minima of the mutational distances to \( \beta \) from each of the highest order schemata defined for all possible sizes of population subset. This approach is depicted graphically in Figure 38.

As stated previously, the approach described in the preceding paragraph represents only one possible approaches to the measurement of the genotypic spatial distances traversed when both variational operators are employed. Operating under the assumption that the proposed approach will achieve more representative results than the techniques described previously (an assumption that will be challenged in Chapter 6), it remains now to derive the weightings by which the minimum distance traversals will the averaged.
Before any recombination is applied, chromosomes cannot accurately be described as offspring proximate to schemata defined by the population.

After one recombination...

After two recombinations, they may be nearer schemata defined by a pair of members.

After three recombinations (or more), chromosomes may be nearest the schema defined by the largest subset of the population.

Figure 38. The schemata to which chromosomes would be expected to be proximate (in genotypic space) following the application of one or more recombination operations defined with respect to the population \{κ₁, κ₂, and κ₃\}.

Contrary to the standard implementation of the point mutation operator, where the value at each index of a chromosome is independently and probabilistically mutated, the likelihood with which recombination is applied to a pair of chromosomes in typically parameterized at a value less than one. Consequently, although the mutation operator is applied each generation, the recombination operator might not be. If the likelihood with which recombination is applied is specified by \(p_{\text{recombination}}\), then the fraction of all possible traversals from chromosome A to chromosome B that will include at least one application recombination is also \(p_{\text{recombination}}\).
Since it is also possible for any offspring of a recombination operation to be recombined with one of its two parent chromosomes in a subsequent generation, the number of possible recombination operations that could be used during the traversal of the genotypic space from \( \alpha \) to \( \beta \) can range between 0 and \( +\infty \). Equivalently, it is possible that the offspring of a recombination of chromosome \( \alpha \) and chromosome \( \kappa \) from the current population might itself be recombined with the chromosome \( \kappa \) in a future generation. This entails that although a single recombination operation specified by the previous example population of chromosomes \( \{\kappa_1, \kappa_2, \kappa_3\} \) will produce offspring at relatively close proximity to only those schemata defined by \( \{\kappa_1\}, \{\kappa_2\}, \) and \( \{\kappa_3\} \), two consecutive recombination operations (specified by the same population) might produce offspring in close proximity to the schemata defined by chromosomes \( \{\kappa_1\}, \{\kappa_2\}, \) and \( \{\kappa_3\} \) or chromosomal pairings \( \{\kappa_1, \kappa_2\}, \{\kappa_1, \kappa_3\}, \) or \( \{\kappa_2, \kappa_3\} \). Similarly, three or more consecutive recombination operations might produce offspring in close proximity to the schemata defined by chromosomes \( \{\kappa_1\}, \{\kappa_2\}, \) and \( \{\kappa_3\} \), chromosomal pairings \( \{\kappa_1, \kappa_2\}, \{\kappa_1, \kappa_3\}, \) or \( \{\kappa_2, \kappa_3\}, \) or the chromosomal triple \( \{\kappa_1, \kappa_2, \kappa_3\} \). Each of these schemata represent a possible location in genotypic space from which the traversal by the mutation operator to the chromosome \( \beta \) must be computed.

To compute a set of appropriate weights for averaging the traversal lengths (including the traversal from \( \alpha \) to \( \beta \) by the mutation operator alone), the set of schemata to which offspring may belong (following the application of recombination) can be conceptualized as a state transition graph. Revisiting the population \( \{\kappa_1, \kappa_2, \kappa_3\} \) of the previous examples in this chapter, it has been established that the sets \( \{\}, \{\kappa_1\}, \{\kappa_2\}, \{\kappa_3\}, \{\kappa_1, \kappa_2\}, \)
\{\kappa_1, \kappa_3\}, \{\kappa_2, \kappa_3\}, \text{ and } \{\kappa_1, \kappa_2, \kappa_3\} \text{ define each of the highest order schemata to which future offspring might belong.}

The weights associated with the matrix for the state transition graph can be computed using the distribution of the population. If chromosomes \(\kappa_1, \kappa_2, \text{ and } \kappa_3\) comprise the complete current population in the ratio 4:2:1 (for example), then if a possible offspring chromosome is a member of the schemata defined by \{\kappa_1, \kappa_3\} was to be subjected to another recombination operation, the offspring could belong to either the schemata defined by \{\kappa_1, \kappa_3\} or the lower order schemata defined by \{\kappa_1, \kappa_2, \kappa_3\}. Since the chromosome \(\kappa_2\) accounts for 2/7 of the population, the offspring would become a member of the former schemata (i.e. defined by only \{\kappa_1, \kappa_3\}) with a probability 5/7 or the latter schemata (i.e. defined by \{\kappa_1, \kappa_2, \kappa_3\}) with a probability 2/7.

Since a chromosome could be recombined such that it becomes a member of any of the highest order schemata defined by a power set of the \(\rho\) unique members of the current population (including the empty set for situations where recombination is not applied at all), there are \(2^\rho\) possible schemata to which an offspring chromosome might belong. Thus, the state transition matrix would have \(2^\rho\) rows and \(2^\rho\) columns. Continuing with previous example population \{\kappa_1, \kappa_2, \kappa_3\}, since an offspring that has been recombined with population members \(\kappa_1\) and \(\kappa_3\) (for example) cannot be subjected to a recombination operation to produce an offspring that is guaranteed to be a member of the schema defined by \(\kappa_1\) alone, the state transition matrix is upper triangular.
For this example, if recombination will not be applied, the column associated with no operand schemata (i.e. the leftmost column) would be an absorbing state, and since no recombination operations have been performed, the matrix has but a single nonzero value. This state transition matrix for zero applications of recombination is depicted below, but if no recombination operations have yet been applied, then no chromosome can be described as the offspring belonging to the schema defined by an operand of the recombination operation by which it was generated. Consequently, the rows associated with the schemata defined by the sets \{κ₁\}, \{κ₂\}, \{κ₃\}, \{κ₁, κ₂\}, \{κ₁, κ₃\}, \{κ₂, κ₃\}, and \{κ₁, κ₂, κ₃\} cannot be assigned meaningful values. The row associated with the schemata defined by the set \{\}, on the other hand, is the only state to which chromosomes can be defined if a recombination operator is not applied.

If exactly one recombination was applied, then offspring could belong to only those schemata defined by \{κ₁\}, \{κ₂\}, or \{κ₃\}, in a ratio determined by the distribution of the population itself. This state transition matrix (for a single application of recombination) is depicted on the following page. Again, the rows associated with schemata to which chromosomes cannot be assigned (before the first recombination) are undefined.
If exactly two recombination operations were applied, then chromosomes that were previously members of the schemata defined by \( \{ \kappa_1 \}, \{ \kappa_2 \}, \) or \( \{ \kappa_3 \} \) would be recombined such that their offspring could belong to the highest order schemata defined by any of the sets \( \{ \kappa_1 \}, \{ \kappa_2 \}, \{ \kappa_3 \}, \{ \kappa_1, \kappa_2 \}, \{ \kappa_1, \kappa_3 \}, \{ \kappa_2, \kappa_3 \}\). Thus, the state transition matrix (for two applications of recombination) would be as follows:

\[
\begin{bmatrix}
0 & 0.57143 & 0.28571 & 0.14286 & 0 & 0 & 0 & 0 \\
0 & 0.57143 & 0 & 0 & 0 & 0.28571 & 0.14286 & 0 \\
0 & 0 & 0.28571 & 0 & 0.57143 & 0 & 0 & 0.14286 \\
0 & 0 & 0 & 0.14286 & 0 & 0.57143 & 0.28571 & 0 \\
\end{bmatrix}
\]

If exactly three recombination operations were applied, then chromosomes that were previously members of the schemata defined by \( \{ \kappa_1 \}, \{ \kappa_2 \}, \) or \( \{ \kappa_3 \} \) could also be recombined such that their offspring belong to the highest order schemata defined by any of the sets \( \{ \kappa_1 \}, \{ \kappa_2 \}, \{ \kappa_3 \}, \{ \kappa_1, \kappa_2 \}, \{ \kappa_1, \kappa_3 \}, \{ \kappa_2, \kappa_3 \}\). On the other hand, chromosomes that were previously members of the schemata defined by \( \{ \kappa_1, \kappa_2 \}, \{ \kappa_1, \kappa_3 \}, \) or \( \{ \kappa_2, \kappa_3 \} \)
could be subjected to their third recombination operation such that their offspring belong to the schemata defined by any of the sets \( \{ \kappa_1, \kappa_2 \} \), \( \{ \kappa_1, \kappa_3 \} \), \( \{ \kappa_2, \kappa_3 \} \), \( \{ \kappa_1, \kappa_2, \kappa_3 \} \). Thus, the state transition matrix (for three applications of recombination) would be as follows:

\[
\begin{bmatrix}
0 & 0.57143 & 0.28571 & 0.14286 & 0 & 0 & 0 & 0 \\
0 & 0.57143 & 0 & 0 & 0.28571 & 0.14286 & 0 & 0 \\
0 & 0 & 0.28571 & 0 & 0.57143 & 0 & 0.14286 & 0 \\
0 & 0 & 0 & 0.14286 & 0 & 0.57143 & 0.28571 & 0 \\
0 & 0 & 0 & 0 & 0.85714 & 0 & 0 & 0.14286 \\
0 & 0 & 0 & 0 & 0 & 0.71429 & 0 & 0.28571 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.42857 & 0.57143 \\
\end{bmatrix}
\]

This process could be repeated indefinitely to compute the likelihood associated with member of offspring in any possible schemata, for any number of recombination operations. Unfortunately, the state associated with the schemata defined by the entire population \( \{ \kappa_1, \kappa_2, \kappa_3 \} \) is actually an absorbing state. Furthermore, the likelihood of a population remaining unchanged decreases as more generations occur. Consequently, developing an estimate of the number of recombination operations that could be applied during traversal (in order to derive the weightings using the state transition matrix) is quite complex. The application of fewer than three recombinations would ignore some of the paths created by recombination, but as the number of recombinations applied varies indirectly with the likelihood that the current population will remain the same, this approach assumes that only one recombination operation is applied for each unique population member. This could certainly be tuned further, but for this investigation the assignment of a value equal to the number of unique population members will suffice.
Finally, with the probably with which recombination is applied being parameterized (by the value $p_{\text{recombination}}$), Figure 39 demonstrates the transitions of the offspring through the space of schemata to which those offspring might belong. The estimated probabilities associated with each transition can be determined using the preceding transition matrix and the parameter $p_{\text{recombination}}$.

Figure 39. The proposed approach to integrating mutational and recombinational distances; with the highest order schemata defining regions of genotypic space that can be ranked by their recombination distance from a fixed point, measures of the minimum mutational distance from each region can be combined (using a weighted average) to estimate the distance between an origin and a destination chromosome as might be traversed by the complete explorative mechanism.
Thus, the proposed approach to the integration of mutational distance has been defined for both the low resolution and high resolution recombinational distances, to form both low resolution and high resolution measures that are representative of traversals by the complete explorative mechanism. The integration of the high resolution recombinational distance is described in pseudocode below (including pseudocode for the proposed derivation of the weights).

The variable `<population>` is the multiset of chromosomes that makes up the population; it contains the set `<unique>` of the unique chromosomes of the population. The variable `<origin>` is the chromosome from which the distance is measured. The variable `<target>` is the chromosome to which the distance is measured. The variable `<schemata>` is the list of the highest order schemata defined by the various n-tuples of the population; each member has integer attribute `<recombine_distance>` for the number of unique chromosomes used for its definition, and integer attribute `<mutate_distance>` for the perceived mutational distance from the nearest element of the schema to the chromosome that has been designated as the `<target>`. The variable `<weights>` is an array of weightings computed (using the proposed technique) for the likelihood of offspring belonging to each set specified by `<schemata>`. The variable `<transition_matrix>` is a matrix of length(schemata)$^2$ real values with each cell containing the probability that the schema corresponding to the row must be expanded to the schema corresponding to the column after an additional recombination. The variable `<initial_state>` is a row vector defining the membership of offspring to the schemata – initially it represents the state for the schema defined by the empty set.
measure-complete-distance

(<population>, <origin>, <target>)

if <origin> = <target>
    return 0

for i = 0 to size(<population>).<unique>)
    for j ∈ the-set-of-i-tuples-of(<population>)
        <schemata>.add(j)
        <schemata>.lastElement.<recombine_distance> = i

for i ∈ <schemata>
    <schemata>(i).<mutate_distance> = 0
    for j = 0 to length(<origin>)

        <schemata>(i).<mutate_distance> +=
        <schemata>(i).<recombine_distance>

        if <origin>.charAt(j) ≠ <schemata>(i).charAt(j) &&
           <schemata>(i).charAt(j) ≠ '*'
           <schemata>(i).<mutate_distance>++

<weights> = compute_weights(<population>)
return (<schemata>(i).<mutate_distance>
    + <schemata>(i).<recombine_distance>) * <weights>

compute_weights

(<population>)

n = size(<population>.<unique>)

for i = 0 to n
    <weights> += (1 - p_recombination)^n-i
        * (p_recombination)^i
        * <initial_state>
        * <transition_matrix>i

return <weights>

### 4.5 Time Complexity Analysis

Since the proposed measure of integrated variational distance requires that a large number of mutational distances (i.e., between the target genotype and several locations in genotypic space) be computed, it is not difficult to surmise that the computational expense necessary for this measurement is considerable. Under the assumption that the
weights associated with each of the mutational distance have already been calculated, the integrated variational distance from genotype \( \alpha \) to genotype \( \beta \), given a population of \( \rho \) genotypes with which \( \alpha \) could be recombined, the first step in the algorithm would be to find all of the \( \rho \binom{C}{1} \) schemata defined by a 1-tuple of the population, all of the \( \rho \binom{C}{2} \) schemata defined by a 2-tuple of the population, all of the \( \rho \binom{C}{3} \) schemata defined by a 3-tuple of the population, etc. Those schemata defined by 1-tuples of the population can be derived with a constant time complexity operation (since the schemata are, themselves, equivalent to the unique population members), and each of the schemata defined by \( x+1 \)-tuple of the population can be derived from at least one of the schemata defined by an \( x \)-tuple with an \( O(\lambda) \) operation (using the expand-schema algorithm described previously). Since there are, thus, \( 2^\rho - \rho - 1 \) operation of \( O(\lambda) \) time complexity required to define each of the locations from which the mutational distance must be measured, this phase of the algorithm is associated with a \( O(\lambda 2^\rho) \) time complexity.

Having defined each of the \( 2^\rho \) required locations in genotypic space corresponding to possible traversals made by the recombination operator, it remains to compute the mutational distance (\( O(\lambda) \) time complexity) from each of these locations to the target. This entails that the time complexity associated with the second phase of the algorithm is also \( O(\lambda 2^\rho) \), indicating that the overall time complexity of the integrated variational distance measure is \( O(\lambda 2^\rho) \).

Conceivably, for a genotypic space of \( 2^\lambda = S \) elements, the distance between each of the \( S(S-1) \) unique pairs of genotypes could be computed by this approach (for an overall time
complexity of $O(\lambda^2 \rho S^3)$). However, it should be noted that the $2^\rho$ schemata defined by the tuples of the population need only be identified once (and would, thus, constitute a one-time preprocessing expense), and the mutational distance from each of these locations to the target genotype varies only with the target genotype (and not with the origin genotype). Although it might seem counterintuitive that many of the individual measurements used to compose the integrated measure can be computed without considering the origin explicitly, consider that the shortest traversable distance from a location in Manhattan (New York) to London is defined almost entirely by the distance between New York’s John F. Kennedy International Airport and the London Heathrow Airport. This distance (between the major airports) would also be highly representative of the shortest traversable distance to London from Queens (New York), Brooklyn (New York), etc, indicating that a measure that is more concerned with the location of the intermediate points (defined, in this case, by the members of the current population) is not an unreasonable approach.

By the proposed approach, the overall time complexity can be reduced to an $O(\lambda 2^\rho)$ preprocessing phase that is followed by a $O(\lambda 2^\rho)$ complexity operation for each of the maximum possible number of target genotypes, $S$. Thus the computational expense associated with the proposed measure of the integrated variational distance between all possible pairs of genotypes would have a worst-case time complexity of $O(\lambda 2^\rho S)$ – nonpolynomial with respect to the population size and genotype length (since $S = 2^\lambda$), but not, thankfully, with respect to the cardinality of genotypic space.
Since it is not unreasonable, when attempting to construct a representative depiction of a vast genotypic space, to use a uniform sampling of the space to reduce the computational expense associated with constructing a data visualization, it is similarly possible to use a sampling of the population to reduced the computational expense associated with the integrated variational distance calculation. At the lower extrema, this may be conceptualized as a measure where the mutational distance to a target genotype need only be computed from those schemata defined by a 1-tuple of the population (and, of course, from the origin genotype). There would, consequently, be only $\rho+1$ mutational distance measurements necessary. Since each of those measurements is made using an operation that has a worst-case time complexity of $O(\lambda)$, the time complexity associated with measuring the distance from one genotype to another is bounded below by an $O(\lambda\rho)$ operation.

Although this must still be performed for each of the possible target genotypes before a pairwise distance matrix can be generated, the overall time complexity is reduced to $O(\lambda\rho S)$ (i.e., polynomial time complexity). If the measure is then complicated (to improve the representativity of the result) by including consideration for those $\rho C_2$ schemata that are defined by the 2-tuples of the population, the time complexity is increased to $O(\lambda\rho^2 S)$ overall – although the worst-case time complexity grows very quickly, it is, at least, conceivable that the phases of the algorithm associated with the nonpolynomial time complexity (with respect to the population size) could be tuned towards a more reasonable calculation (at the expense of the representativity of the approach).
4.6 Mechanism Representative Distance Summary

With the development of three novel approaches to the measurement of genotypic spatial distances as they would be traversed by the complete explorative component of the evolutionary mechanism of the genetic algorithm, it remains to assess the suitability of each. Under the (safe) assumption that computational expense will be a consideration in every possible application, the relative complexity of each approach must be weighed against the representativity of the achievable results. This is especially true with the proposed approach to the integration of mutational and recombinational distances, for which the worst-case time complexity was shown to be $O(\lambda^2pS)$ and, thus, nonpolynomial. Although this chapter also introduced a greedy measurement heuristic that could be computed in polynomial time, this approach was shown to overestimate the recombinational distance between genotypes.

More importantly, this chapter demonstrated that the task of measuring recombinational distance is, essentially, equivalent to the NP-hard set cover optimization problem. This entails that an approach that considers every possible tuple that is defined by the current population must be associated with a nonpolynomial worst-case time complexity. While it might seem dissuasive that the proposed technique must always deal with an exponential term in the worst-case time complexity, it should be noted that complete depictions of genotypic spaces also must invariably deal with exponential time complexities in the genotype length $\lambda$ – a point that will be revisited in Chapter 7.2. Furthermore, the equivalence between recombinational distance measurement and the set
cover optimization problem also entails that there are several other efficient heuristics that might be serviceable and should be the subject of future investigation.

Since it has also been demonstrated that nontrivial traversals of genotypic space can be quite complex, it is necessary to propose a fair and unbiased approach to calibration and comparative evaluation. Analogously, if two measurement devices are to be compared (with the goal of identifying which device is more accurate), it is necessary to provide either a calibration approach (and the necessary materials) or a substantial set of traversals against which the accuracy of each device can be assessed. As no such set exists (at this time) that is representative of the traversals of genotypic space by the explorative mechanism of the genetic algorithm, an empirical approach to the measurement of genotypic spatial distances has been developed.

In the next chapter, an approach to empirical distance measurement that is inspired by a simplistic geometric construction in two dimensions is introduced and developed. As this approach will not be a genetic algorithm itself, the approach is then validated using the infinite population model of the genetic algorithm (Vose 1999). As indicated by the fragment of the research overview chart that is depicted in Figure 40, this empirical distance measure is foundational to the calibration and comparative evaluation of the approaches presented in this chapter. Following the introduction and validation of this empirical measurement approach, it becomes possible to evaluate the measures introduced in this chapter according to their performance at generating ranks for the chromosomes of genotypic space according to their distance from a fixed point. It can
then be conclusively determined if considering the traversals of genotypic space by the complete explorative mechanism yields more representative results than only considering traversals made by the mutation operator alone.

Figure 40. The lower portion of the summarizing chart from Figure 15. The next chapter will discuss the development of an empirical approach to genotypic spatial distance measurement. This approach was designed for the task of calibrating and comparatively evaluating the measures introduced in this chapter, and is validated against the infinite population model of the genetic algorithm before it is applied in Chapter 6.
Chapter 5

Empirical Distance Measurement

Overview

The last chapter introduced three approaches to integrated variational distance measurement (the mutational distance from the population centroid and the mutational distance from genotypes at measured low resolution and high resolution recombinational distance from a point of origin). Although the computational expense and achievable distance resolution associated with these measures has been discussed, for the enhancement of the adaptive landscape it is necessary to conclusively determine the relative accuracy with which each approach represents the actual traversal of genotypic space by the genetic algorithm. However, in the absence of a representative genotypic spatial distance measure against which approaches can be compared, differences in performance can be observed (and tested for statistical significance) but a comparative evaluation of the three novel approaches and the Hamming distance is not possible.
To address this complication, this chapter describes the development of an empirical approach to genotypic spatial distance measurement. By generating actual traversals of genotypic space (by the explorative component of the evolutionary mechanism), the accuracy with which a novel approach generates the same ranking observed in the actual traversals is used to assess relative performance. This chapter describes the process by which the actual traversals are generated, by exploring an analogous geometric construction of a series of evenly spaced concentric circles using only a drafting compass with a fixed radius.

By fixing a drafting compass to a point of origin in two dimensional space, a circular subplane can be generated that will contain every point in the space that is separated from the origin by a distance less than or equal to the width of the compass. If each of those points is then used to construct another circular subplane of the same width, the union of the new subplanes contains every point that is separated from the origin by a distance less than or equal to twice the compass width. This process (by which evenly spaced, concentric, circular subplanes can be generated) can be used to assign points a rank according to their distance from the origin. Similarly, the explorative component of the evolutionary mechanism can be used to generate concentric subspaces (referred to as zonal neighbourhoods) that are truly representative of the manner in which genotypic space is traversed.

Since the variational operators of the genetic algorithm are applied probabilistically, it is necessary to employ a Monte Carlo approach to the task of identifying concentric zonal
neighbourhoods. This chapter addresses each of the considerations necessary for adapting the analogous geometric construction process to the task of empirical genotypic spatial distance measurement. Furthermore, since the traversals are being generated by an abstraction of the simple genetic algorithm, this chapter verifies the validity of the proposed empirical distance measurement against the infinite population model of proposed by Vose in 1999.

5.1 Analogous Geometric Construction

Geometric construction techniques that employ a drafting compass to construct a bounding circular subplane associated with the fixed compass radius serve as an effective analogy for the empirical approach to distance measurement. Just as the act of turning a drafting compass is an exhaustive (continuous) exploration of all possible angles at which a mechanism capable of two-dimensional motion from the origin might proceed, a model of the explorative component of the evolutionary mechanism of the genetic algorithm might similarly be used to identify the subspace of genotypes that can be reached from any origin genotype by the application of the variational operators. As such, the geometric construction analogy is a valuable tool for comprehensively exploring the empirical distance measure.

For this analogy, a drafting compass has one leg fixed to a point on the surface of a finite two-dimensional plane, with that point being designated as the origin. The angle between the legs of the compass is fixed, in the context of this example, such that the compass
radius (i.e. the distance between the two points of intersection between the compass legs and the plane) is representative of the distance that could be traversed (over a fixed interval of time $T$) by the mechanism in question. By turning the compass, a circular subplane $C_1$ centered at the origin point is identified such that the distance between the origin and every point within the subplane is no more than the compass radius and, thus, each point in the subplane can be reached by the mechanism in question in, at most, $T$ units of time. If the compass is used to identify additional subplanes of the same radius $C_2$, centered at every point $i$ within the subplane $C_1$, the union of the additional subplanes $\bigcup C_2 \forall i \in C_1$ also identifies a circular subplane, denoted $C_2$ and centered at the origin point, such that the distance between the origin and every point within the subplane $C_2$ can be reached by the mechanism in question in, at most, $2T$ units of time. Furthermore, the distance between the origin and every point contained in the set difference between these subplanes, $C_2 - C_1$, must also be greater than a single compass radius and, thus, requires at least $T$ units of time to be reached by the mechanism in question. Consequently, every point that can be reached from the origin requiring either $T$ or $2T$ units of has been identified. Naturally, this approach can be repeated until the amount of time required by the mechanism (expressed as an integer multiple of $T$) to traverse each of the distances between the point of origin and another point in the finite two-dimensional plane has been determined. This empirical process, illustrated in Figure 41 on the following page, can serve as a model upon which a measure of the genotypic spatial distance traversed by the variational operators that comprise the explorative component (i.e. mutation and recombination) of the evolutionary mechanism of the genetic algorithm can be constructed.
Figure 41. The geometric construction analogy. With a fixed radius, the compass can be used at a point to define a two-dimensional circular subspace, and if each point in this subspace is also used to generate a two-dimensional circular subspace of the same radius, the union of the set of subspaces will define a two-dimensional circular subspace of twice the original compass radius.

Although the technique described by the preceding analogy must be heavily adapted before it would be suitable as an approach to the measurement of the genotypic spatial distances traversed by the genetic algorithm, it is not difficult to extend the analogy from a continuous, two-dimensional space to a discrete, binary, four-dimensional hypercubic space that is more easily equated with a genotypic space. Assuming a genotypic space of four bit candidate solutions, a mechanism defined entirely by a single bit flipping operator and origin chromosome ■■■■, the first subspace to be identified is depicted in Figure 42.
Figure 42. A genotypic space of hypercubic topology. It will be demonstrated that a genotypic space traversed by both a mutation and a recombination operator should not be treated as though it were constructed with a hypercubic topology.

The turning of the drafting compass is a trivially simple method by which the set of possible directions for an object traversing two-dimensional space can be exhausted, but it is clear that the set of possible 'directions' for the motion through a space of higher dimensionality can be more difficult to conceptualize. However, as evidenced by the Figure 42, there is a direct relationship between the Hamming distance that separates each chromosome from the origin and the membership of that same chromosome in one of the indexed subspaces. It should also be clear that the proposed methodology can be used to effectively measure genotypic spatial distances. Unfortunately, as the explorative component of the evolutionary mechanism uses both probabilistic application and a binary operator, the possible 'directions' through genotypic space that are associated with the explorative component are more difficult to define.
With the proposed mutation and recombination operators of this investigation (probabilistic and independent bit flipping mutation and uniform recombination, respectively), any member of the current population of the genetic algorithm can be considered as having been subjected to both a mutation operation and a recombination operation. For a population of $\rho$ binary string candidate solution representations of length $\lambda$, there are as many possible mutation operations as there are members of the genotypic space $S$ of candidate solutions, which, assuming the absence of any infeasible genotypes, is of cardinality $2^\lambda$. Furthermore, as every uniform recombination operation between chromosomes of length $\lambda$ can be uniquely identified by a binary crossover mask of length $\lambda$, there are also $2^\lambda$ possible recombination operations between each pair of chromosomes. Thus, with the current population containing $\rho$ members with which a specific individual chromosome could be recombined, there are $\rho 2^\lambda$ possible recombination operations of which some may be functionally equivalent. There are, thus, a total of $2^\lambda \cdot \rho 2^\lambda (= \rho 2^{2\lambda})$ possible applications of the explorative component of the evolutionary mechanism (i.e. unique combinations of a mutation operation with a recombination operation) that could be applied to any individual genotype. The $\rho 2^{2\lambda}$ possible operations entail that there may be as many possible 'directions' for the motion of each population member through genotypic space that would be effected by the explorative component of the evolutionary mechanism each generation. As this would suggest an intractable number of calculations to fully exhaust the set of possible directions, a Monte Carlo approach is favoured. Such an approach would repeatedly subject a chromosome to random mutation and recombination operations. The set of chromosomes resulting from these repeated random operations would then be considered
representative of the set of possible chromosomes that can be reached by a single application of the explorative component of the evolutionary mechanism. It is then possible, through repeat application of this approach, to establish a sequence of contoured zones, equivalent to the contoured representation of land relief often employed in two-dimensional topographic maps. Paralleling the circular subplanes from the earlier drafting compass analogy, the first contoured zone (designated $C_0$) would contain only the origin, while contoured zones $C_1$, $C_2$, ... $C_n$ would contain those chromosomes designated the first nearest, second nearest, and $n^{th}$ nearest chromosomes to the origin, respectively.

Having established the process by which the set of possible 'directions' for genotypic space traversals by the evolutionary mechanism can be exhausted, it remains to observe those properties of the approach that, although sufficient for the geometric construction, must be challenged and adapted for use by the genetic algorithm.

5.2 Important Considerations

Unsurprisingly, the technique suggested by the preceding geometric construction analogy must be heavily adapted before it would be suitable as an approach to the measurement of the genotypic spatial distances traversed by the genetic algorithm. Having already addressed (in the preceding section) the determination of the set of possible 'directions' for the motion of the evolutionary mechanism through genotypic space, it remains to consider the ramifications of the three observations of the bulleted list on the following page, in the order with which they will be addressed.
• Each stage of the geometric construction applies a union operation on a set of subplanes, each of which being centered at a unique point in the subplane identified in the previous stage. This observation would suggest a uniform distribution from which each point (or, equivalently, 'direction') might be selected randomly.

• The geometric construction analogy (and, specifically, the depiction in Figure 41) would seem to suggest that only points contained at the extrema of an identified subplane are actually needed for the next iteration. This assumption that is not expected to hold for a traversal by two distinct operators.

• Finally, because a ray emitted from the origin in a two-dimensional plane is expected to proceed in only a single direction, the geometric construction analogy did not require the specification of an upper limit on the compass radius. Since the genetic algorithm is expected to 'change direction' frequently, the possibility of an upper limit to the parameter corresponding to the radius must be investigated.

The previously described analogy results in the delineation of a circular subplane that suggests the set of possible points from which any 'neighbour' (at a distance less than or equal to a single compass width) could be randomly selected using a uniform distribution. It has, however, been demonstrated that the mutation and recombination operators of the genetic algorithm do not result in the uniformly random selection of one of the $2^λ$ candidate solution configurations that could be considered neighbours. It is, in fact,
crucial to the evolutionary mechanism that, for any current population, some chromosomes are more likely to appear in the next generation than others. Thus, the possible neighbours of a chromosome subjected to a single application of the explorative component of the evolutionary mechanism will be generated at different degrees of likelihood. This would be most accurately reflected by the algorithm if chromosomes more likely to appear than others (after a single application of the explorative component) are assigned membership to a contoured zone that is designated a lesser distance from the origin. As a clarifying example, consider a simplified mechanism defined entirely by a point mutation operator that independently and probabilistically (with a probability of 0.25) flips the bit at each index of a four bit binary chromosome. From an origin chromosome of $\text{■■■■}$, the four chromosomes for which only a single bit differs from the origin (i.e. $\text{■■■□}$, $\text{■■□■}$, $\text{■□■■}$, and $\text{□■■■}$) would each be the result of a single application of the explorative component of the evolutionary mechanism with a probability of $(0.25)(0.75)^3$. Accordingly, the six chromosomes for which exactly two bits differ (i.e. $\text{■■□□}$, $\text{■□■□}$, $\text{■□□■}$, $\text{□■■□}$, $\text{□■□■}$, and $\text{□□■■}$) would be the result of a single application with the lesser probability $(0.25)^2(0.75)^2$. It should be noted that the proposed experimental approach will only yield observations that can be used to approximate these probabilities and thus it is will be necessary to use significance testing to determine the set of chromosomes that should be assigned to each contoured zone.

Also contrary to the drafting compass analogy is the observation that, following the determination of the set of chromosomes contained in contoured zone $C_1$, it is not immediately obviously which chromosomes should be subjected to the mutation and
recombination operators to determine contoured zone $C_2$. To clarify, although it was true in the drafting compass analogy that the first circular subplane $C_1$ was centered upon the origin point $C_0$ and the circular subplane that would be created by the union of subplanes constructed using the points of $C_1$ (with $C_0 \subseteq C_1$) as central points would, in fact, be equivalent to the union of subplanes constructed using the points of set difference $C_1 \setminus C_0$ as possible points of origin. Similarly, in determining $C_3$, the union of subplanes created center points from the one of the sets $C_2$, $C_2 \setminus C_0$, and $C_2 \setminus C_1$ would all be identical.

Although this result could be construed as indicating that that the set of operators need only be applied to those chromosomes belonging to the furthest contoured zone from the origin chromosome, this is not the case for the evolutionary mechanism of the genetic algorithm. This result is, in fact, a consequence of the subadditivity property (i.e. the satisfaction of the triangle inequality) of the two-dimensional Euclidean distance associated with the geometric construction analogy. Since the distance between the origin and a point in contoured zone $C_{i+2}$ cannot exceed the sum of the distance between a point in $C_i$ and a point in $C_{i+1}$ and the distance between that point in $C_{i+1}$ and the point in $C_{i+2}$, there is no reason to include the points in $C_i$ in the determination of $C_{i+2}$. It must, however, be emphasized that the triangle equality does not hold for a distance defined as a function of both mutational-defined and recombination-defined measures of distance.

Finally, although the geometric construction analogy does not indicate the presence of any constraints on the compass width, it must also be stressed that the evolutionary mechanism of the genetic algorithm typically does not traverse genotypic space in what
would be considered a single direction. Mutation and recombination operations are
determined probabilistically each generation and the state of the population in the current
generation determines the possible offspring a chromosome might produce if subjected to
recombination. It could then be concluded that, since a single generation of the genetic
algorithm represents the longest period of time that the evolutionary mechanism can be
guaranteed not to change direction, the width of the drafting compass (following the
geometric construction analogy), in the context of empirical distance measurement for the
genetic algorithm, must never exceed a single generation.

5.3 Adaptation to the Genetic Algorithm

With respect to the considerations and constraints presented in the previous paragraphs,
two empirical approaches to the measurement of genotypic spatial distance traversed by
the evolutionary mechanism of the genetic algorithm are summarized in pseudocode on
the following page. The algorithm requires, for initialization, the candidate solution space
$S$, the current population $P$, and the singleton set $C_0$ containing the chromosome
designated as the origin. Up to $|S|$ sets of chromosomes, referred to as neighbourhoods, of
the form $C_i$ for $0 < i < |S|-1$, are identified by the algorithm such that the elements of the
$i^{th}$ neighbourhood $C_i$ contains only those chromosomes assigned a relative distance of $i$
from the origin chromosome. Chromosomes are assigned membership to neighbourhood
$C_i$ if the likelihood with which they appear is deemed significant, after repetitive and
probabilistic application of the variational operators to each chromosome contained in
every neighbourhood $C_j$ where $j < i$. 
The variable `<genotypes>` is the array of genotypes; it has Boolean attribute `<measured>` for whether or not a genotype has been assigned membership to a contoured zone, and integer attribute `<frequency>` to record the number of appearances.

The variable `<zone>` is the array of contoured zones; it has real attribute `<quotient>` for the fraction of the Monte Carlo trials using the chromosomes in that zone as the origins.

The variable `<reachable>` is the set of chromosomes with nonzero appearance frequencies that are not assigned to a zone.

// the origin is assigned membership to zone 0

<genotypes>[0].<measured> = true
<zone>[0].<quotient> = 1.0

i = 1
until |<zone>[i-1]| = 0

<genotypes>[1 to |S-1|].<frequency> = 0

// for each chromosome in an established zone

for j = 0 to i

    for <origin> ∈ <zone>[j]

        for k = 1 to <zone>[j].<quotient> ÷ |<zone>[j]|

            // apply the operators and record the result
\[<\text{result}>= \text{mutate}(\text{recombine}(<\text{origin}>))\]

\[<\text{genotypes}>[<\text{result}>].<\text{frequency}>++\]

\[<\text{reachable}> = \text{descending-sort-by-frequency}(<\text{genotypes}>)\]

for \(j\) where \(<\text{genotypes}>[j].<\text{included}> = \text{true}\)

\[<\text{reachable}> = <\text{reachable}> \setminus <\text{genotypes}>[j]\]

// add most frequent chromosomes to the current zone

for \(j = 1\) to \(|<\text{reachable}>|\)

until significant_frequency_difference(1, \(j+1\))

\[<\text{zone}>[i] = \text{zone}[i] \cup <\text{reachable}>[j]\]

// the trial quotient for each zone is recalculated

for \(j = 0\) to \(i\)

\[<\text{zone}>[j].<\text{quotient}> = 0\]

for \(k = 0\) to \(|S-1|\)

if \(<\text{genotypes}>[k] \in <\text{zone}>[j]\)

\[<\text{zone}>[j].<\text{quotient}>+= <\text{genotypes}>[k].<\text{frequency}>\]

normalize \(<\text{zone}>[0\rightarrow i].<\text{quotient}>\)
Significance could be determined through the specification of a threshold for the minimum likelihood with which a chromosome must appear to be assigned membership to the current neighbourhood or through the use of statistical significance testing. Although both approaches have been explored, the former is easier to implement and less computationally expensive, while the latter is more suitable for mapping the relative likelihood associated with a neighbourhood to an actual transition probability.

A clarifying example of the process by which the preceding approaches empirically measure the distance between chromosomes by identifying concentric neighbourhoods is included in Figure 43. This figure depicts the ranking of chromosomes according to their empirically measured distance from an origin \( C_0 = \{\text{■■■■}\} \), for the set of genotypes of length \( \lambda = 4 \) and a current population of two members, \( \square\square\square\square \) and \( \text{■■□□} \). As the latter approach associates each chromosome in the candidate solution space with a probability-defined measure of the genotypic spatial distance between the chromosome and a point of origin, as traversed by both variational operators of the genetic algorithm, it can be applied to compute a distance between two chromosomes. The result, rather than having been computed with an arbitrarily selected measure (such as the Hamming or Euclidean distance), was instead computed through the empirical observation of actual variational operator applications and, as such, can be considered an undistorted measure of the genotypic spatial distance traversed by those operators. The empirical measure is, thus, an excellent tool for the calibration and evaluation of the novel measures introduced in Chapter 4 (i.e., the Euclidean distance from the centroid and the integrations of mutational distance with low and high resolution recombination distance, respectively).
Figure 43. The ranking of genotypes of length $\lambda = 4$ by empirical distance from the origin chromosome. Contrary to the rankings that would be established by a Hamming distance measure, chromosomes □□■■ and ■■□□ are assigned ranks nearer the origin than the other chromosomes at a Hamming distance of two.

Before proceeding with the evaluation and comparison of the possible measures of genotypic spatial distance, it is important to emphasize that the processes described previously are not to be confused with the mean first hitting time statistic. Although the mean first hitting time that is associated with a random walk could conceivably be employed as a measure of proximity and an alternative to the explicit measurement of
distance, it is crucial to observe that a random walk is expected to change directions frequently, thus following a longer path than is usually necessary. Contrarily, the approach to empirical distance introduced previously always identifies (at least with respect to the probability constraints outlined previously) the smallest number of variational operations of the genetic algorithm necessary to transform one chromosome (in the current generation) into another (in a future generation). Moreover, it is emphasized that the evolutionary mechanism of the genetic algorithm, although Markovian, is guided in its traversal of genotypic space by the mechanism of selection, and is, thus, certainly not a random walk.

5.4 Empirical Approach Validation

It should be evident from the pseudocode introduced in the previous section that the proposed approach to empirical distance measurement was not meant to actually parallel a generalized instance of the genetic algorithm. It was, instead, designed to repeatedly emulate the application of the exploratory component of the evolutionary mechanism to a specific chromosome, ultimately predicting the neighbourhood of the chromosome with respect to the application of the variational operators defined by the current population. As explicitly noted, this is not entirely unlike computing the transition probability between all possible pairs such that sets of chromosomes associated with transformations of decreasing degrees of likelihood are mapped to neighbourhoods that are assigned representative distance values from the origin. There are, however, subtle differences that preclude the use of transition probabilities alone for a representative distance measure.
As noted several times previously, the distance measures associated with genotypic spatial topology are typically conceptualized as either sets of move sets or sets of transition probabilities. Although it was previously demonstrated that simple move sets are not suitable for representing genotypic spatial distance traversal by mutation, without the specification of an arbitrary threshold parameter, no justification for the lack of a measure of recombinational distance comprised simply of a set of transition probabilities has yet been introduced. The discussion was deferred here because this chapter explores the relationship between an actual Markov model of the evolutionary mechanism and the empirical approach to genotypic spatial distance measurement that was introduced in the previous chapter, in an attempt to validate the legitimacy of the empirical approach.

Vose's infinite population model of the genetic algorithm was used previously to demonstrate how the exploitative selection operator can be analyzed separately from the explorative variation operators. The evolutionary mechanism is represented, in this model, as the product of a transition matrix associated with the selection operator and another transition matrix associated with the variation operators, while abstracting a finite population of chromosomes into an infinite population that can be effectively represented as a row vector. Although the selection mechanism was noted previously to be evaluation function dependent (necessitating that it be excluded from any approach for which evaluation independence is deemed necessary), the infinite population model does facilitate investigation of the probabilities associated with random traversals of the genotypic space by the variation operators, provided, of course, that the available current population of chromosomes defining the range of the recombination operator is known. It
was noted previously that the probabilities for the mutational distance between chromosomes $\alpha$ and $\beta$, of length $\lambda$ and separated by Hamming distance $\delta$, is defined by:

$$(\lambda^{-1})^\delta \cdot (1-\lambda^{-1})^{\lambda - \delta}.$$ 

If, however, the Hamming distance $\delta$ between chromosomes is not known beforehand, the computation of the transition probability associated with the mutation of chromosome $\alpha$ into chromosome $\beta$ requires the bitwise logical comparison of the binary values, for each index $i$, $0 \leq i < \lambda$. The likelihood, then, of chromosome $\alpha$ being mutated into chromosome $\beta$ in a single generation is the product, for each index, of either the probability parameter associated with mutation ($p_{\text{mutation}}$) if $\alpha$ and $\beta$ are complementary at that index, or one minus the probability parameter associated with mutation if $\alpha$ and $\beta$ have the same value at that index. This formula, equivalent to the previous formula in terms of both functionality and time complexity, is defined:

$$i < \lambda \prod_{i=0}^{\lambda} \left( p_{\text{mutation}} \cdot (\alpha_i \neq \beta_i) + (1 - p_{\text{mutation}}) \cdot (\alpha_i = \beta_i) \right).$$

Having redefined the binary recombination operation into a set of unary recombination operators defined by the current population of chromosomes, the transition probability associated with the recombination of chromosome $\alpha$ with a member of the population to produce chromosome $\beta$ as an offspring becomes the sum of each probability associated with the likelihood that $\alpha$ could be recombined with a unique chromosome $\kappa$ from the
population to produce $\beta$ as an offspring, for each of the unique chromosomes in the population. Unsurprisingly, since this determination would also require bitwise comparisons to determine which indexed values of chromosome $\alpha$ should be exchanged with those of chromosome $\kappa$ to produce chromosome $\beta$ as an offspring, the formula describing the transition probabilities that could be used to define a measure of recombinational distance is structurally similar to the formula described previously for a measure of mutational distance. It should be noted that the transition probability that a recombination operation applied to chromosomes $\alpha$ and $\kappa$ could produce chromosome $\beta$ as an offspring will only be nonzero if and only if $\neg(\alpha_i \neq \beta_i \land \alpha_i = \kappa_i)$ for each index value $i$, $0 \leq i < \lambda$. Otherwise, at every index $i$ for which $\alpha_i = \beta_i$, the crossover mask must have a value of 0 if $\alpha_i \neq \kappa_i$, and at every index $i$ for which $\alpha_i \neq \beta_i$, the crossover mask must have a value of 1 if $\alpha_i \neq \kappa_i$. With the likelihood that an arbitrary index of the crossover mask will have a value of 1 defined by the uniform recombination probability parameter $p_{\text{recombination}}$, the transition probability that a recombination operation applied to chromosomes $\alpha$ and $\kappa$ could produce chromosome $\beta$ as an offspring is defined:

$$
\prod_{i=0}^{\lambda} \left( (\alpha_i = \kappa_i) \land (\alpha_i \neq \beta_i) \right) \prod_{i=0}^{\lambda} \left( (\alpha_i \neq \kappa_i) \land (\alpha_i = \beta_i) \right) p_{\text{recombination}} + \left( (\alpha_i = \kappa_i) \land (\alpha_i = \beta_i) \right) (1 - p_{\text{recombination}})
$$

The structural similarity between this formula and the previous formula, for the transition probabilities associated with mutation and recombination, respectively, is not surprising given that a mutation operator applied with probability $p_{\text{mutation}}$ to a chromosome $\alpha$ is equivalent to a uniform recombination operation applied with probability $p_{\text{recombination}}$ to complementary chromosomes $\alpha$ and $\kappa = \neg\alpha$ if the probability parameter $p_{\text{mutation}}$ is equal
to the probability parameter $p_{\text{recombination}}$. As $\kappa = \neg \alpha \leftrightarrow \alpha_i \neq \kappa_i \forall i, \ 0 \leq i < \lambda$, the equivalence can be easily demonstrated and is included below.

1. $\forall \ 0 \leq i < \lambda \ \ i(\kappa_i = \neg \alpha_i)$

2. $P_{\text{mutation}} = P_{\text{recombination}}$

\[
\begin{align*}
  &\prod_{i=0}^{i<\lambda} \neg((\alpha_i = \kappa_i) \land (\alpha_i \neq \beta_i)) \prod_{i=0}^{i<\lambda} (((\alpha_i \neq \kappa_i) \land (\alpha_i \neq \beta_i)) \cdot p_{\text{recombination}} + ((\alpha_i \neq \kappa_i) \land (\alpha_i = \beta_i)) \cdot (1 - p_{\text{recombination}})) \\
  &\prod_{i=0}^{i<\lambda} \neg(\text{false} \land (\alpha_i \neq \beta_i)) \prod_{i=0}^{i<\lambda} (((\text{true} \land (\alpha_i \neq \beta_i)) \cdot p_{\text{recombination}} + ((\text{true} \land (\alpha_i = \beta_i)) \cdot (1 - p_{\text{recombination}})) \\
  &\prod_{i=0}^{i<\lambda} ((\alpha_i \neq \beta_i) \cdot p_{\text{mutation}} + ((\alpha_i = \beta_i) \cdot (1 - p_{\text{mutation}})) \\
  &\prod_{i=0}^{i<\lambda} ((\alpha_i \neq \beta_i) \cdot p_{\text{mutation}} + ((\alpha_i = \beta_i) \cdot (1 - p_{\text{mutation}}))
\end{align*}
\]

With the preceding formula it is possible to construct the complete transition probability matrix associated with the set of unary recombination operations. It should be noted that
although the point mutation and uniform recombination operations are known to be commutative, the transition probability matrix associated with recombination must be generated using the current population. If the application of mutation precedes the application of recombination, the transition probability matrix associated with recombination that should be generated will be different. Thus, for the example in the following paragraph, the complete transition probability matrix associated with the two variational operators (point mutation and uniform recombination) that comprise the entire explorative component (of the evolutionary mechanism of the genetic algorithm) is actually computed as the product of the transition probability matrices associated with recombination and mutation, in that order.

The transition probability matrices constructed using the previous formulae closely match the intuitive conceptualizations about the distances in genotypic space that could be traversed in the application of the variational operations of a single generation. The transition probability matrices associated with point mutation and uniform recombination, and the product of the latter and the former, are included for comparison, between a nonexistent population (for which the transition probability matrix associated with the recombination operation is the identity matrix) and a population comprised of only a single member – chromosome □□□. Naturally, as the mutation operation can be applied independent of the genotypes in the population, the matrix associated with mutation when the population is empty is identical to the matrix associated with mutation when the population is {□□□}. 
state transition matrix for point mutation, specified by $p_{\text{mutation}} = 0.33333$:

$$
\begin{bmatrix}
0.29630 & 0.14815 & 0.14815 & 0.07407 & 0.14815 & 0.07407 & 0.07407 & 0.07407 & 0.03703 \\
0.14815 & 0.29630 & 0.07407 & 0.14815 & 0.07407 & 0.14815 & 0.03703 & 0.07407 & 0.07407 \\
0.14815 & 0.07407 & 0.29630 & 0.14815 & 0.07407 & 0.03703 & 0.14815 & 0.07407 & 0.07407 \\
0.07407 & 0.14815 & 0.14815 & 0.29630 & 0.03703 & 0.07407 & 0.07407 & 0.14815 & 0.07407 \\
0.14815 & 0.07407 & 0.07407 & 0.29630 & 0.14815 & 0.14815 & 0.07407 & 0.29630 & 0.14815 \\
0.07407 & 0.03703 & 0.14815 & 0.07407 & 0.07407 & 0.14815 & 0.29630 & 0.07407 & 0.14815 \\
0.03703 & 0.07407 & 0.07407 & 0.14815 & 0.07407 & 0.14815 & 0.14815 & 0.14815 & 0.29630 \\
\end{bmatrix}
$$

state transition matrix for recombination with an empty population:

$$
\begin{bmatrix}
1.00000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.00000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.00000 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.00000 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.00000 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.00000 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1.00000 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.00000 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.00000 \\
\end{bmatrix}
$$

state transition matrix for recombination with population $\{□□□\}$, $p_{\text{recombination}} = 0.75$:

$$
\begin{bmatrix}
0.34375 & 0.09375 & 0.09375 & 0.09375 & 0.09375 & 0.09375 & 0.09375 & 0.09375 & 0.09375 \\
0 & 0.43750 & 0 & 0.01875 & 0 & 0.01875 & 0 & 0.01875 & 0 \\
0 & 0 & 0.43750 & 0.01875 & 0 & 0 & 0.01875 & 0.01875 & 0 \\
0 & 0 & 0 & 0.62500 & 0 & 0 & 0 & 0.37500 & 0 \\
0 & 0 & 0 & 0 & 0.43750 & 0.01875 & 0.01875 & 0.01875 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.62500 & 0 & 0.37500 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.62500 & 0.37500 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.00000 & 0 \\
\end{bmatrix}
$$
state transition matrix for mutation and recombination with the empty population:

\[
\begin{bmatrix}
0.29630 & 0.14815 & 0.14815 & 0.07407 & 0.14815 & 0.07407 & 0.07407 & 0.07407 & 0.03703 \\
0.14815 & 0.29630 & 0.07407 & 0.14815 & 0.07407 & 0.14815 & 0.03703 & 0.07407 & 0.14815 \\
0.14815 & 0.07407 & 0.29630 & 0.14815 & 0.07407 & 0.03703 & 0.14815 & 0.07407 & 0.14815 \\
0.07407 & 0.14815 & 0.14815 & 0.29630 & 0.03703 & 0.07407 & 0.14815 & 0.14815 & 0.14815 \\
0.14815 & 0.07407 & 0.07407 & 0.07407 & 0.14815 & 0.14815 & 0.14815 & 0.07407 & 0.29630 \\
0.07407 & 0.14815 & 0.03703 & 0.07407 & 0.14815 & 0.14815 & 0.14815 & 0.07407 & 0.14815 \\
0.03703 & 0.07407 & 0.07407 & 0.14815 & 0.07407 & 0.14815 & 0.14815 & 0.14815 & 0.29630 \\
\end{bmatrix}
\]

state transition matrix for mutation and recombination with population \{□□□\}:

\[
\begin{bmatrix}
0.16782 & 0.13079 & 0.13079 & 0.11227 & 0.13079 & 0.11227 & 0.11227 & 0.11227 & 0.10301 \\
0.09954 & 0.19907 & 0.08102 & 0.16204 & 0.08102 & 0.16204 & 0.07176 & 0.16204 & 0.14352 \\
0.09954 & 0.08102 & 0.19907 & 0.16204 & 0.08102 & 0.07176 & 0.16204 & 0.07176 & 0.14352 \\
0.06019 & 0.12037 & 0.12037 & 0.24074 & 0.05093 & 0.10185 & 0.10185 & 0.20370 & 0.20370 \\
0.09954 & 0.08102 & 0.08102 & 0.07176 & 0.19907 & 0.16204 & 0.16204 & 0.20370 & 0.14352 \\
0.06019 & 0.12037 & 0.05093 & 0.10185 & 0.12037 & 0.24074 & 0.10185 & 0.20370 & 0.20370 \\
0.06019 & 0.05093 & 0.12037 & 0.10185 & 0.12037 & 0.10185 & 0.24074 & 0.20370 & 0.20370 \\
0.03704 & 0.07407 & 0.07407 & 0.14815 & 0.07407 & 0.14815 & 0.14815 & 0.14815 & 0.29630 \\
\end{bmatrix}
\]

Although recombination is an ineffective operator when no population is present (as can be deduced from the equivalence to the identity matrix), recombination between chromosome ■■■ and population member □□□ (with which it is complementary) can produce any chromosome (with uniformly distributed probabilities), while recombination between chromosome ■■■ and population member ■■■ can produce no other offspring than ■■■ itself. Finally, the two matrices associated with the use of both variational operators (for the empty population and the population \{□□□\}, respectively) indicate that
the presence of population member □□□ effectively increases the likelihood (reduces the distance) between every chromosome in the population and □□□.

Although the preceding example matches our intuition about the likelihood of one chromosome transforming into another in a single generation, the evolutionary mechanism is not expected to function in a single iteration. Ideally, a measure of the distance between two chromosomes that must be traversed (through genotypic space) by the explorative mechanism should reflect more than simple traversals of length 0 and 1. As a clarifying example, the transitional probability associated with the transformation of chromosome ■■■ into chromosome □□□, using those variational operators defined by a nonexistent population and a point mutation rate of zero would, naturally, be the identity matrix. If the chromosomes ■□□ and □□■ were then added to the population the transition probability matrix associated with recombination (computed using the approach detailed previously) would reflect the possible recombination operations that could be applied.

It is important to note that the transition probability matrix value for the transition from ■■■ to □□□ (by recombination with the population {■□□, □□■}) actually remains zero in spite of the fact that recombination alone can now achieve the traversal from ■■■ to □□□. This example (i.e. recombination with population {■□□, □□■} but a mutation rate of zero) is depicted in Figure 44. The absence of representation for the traversal from ■■■ to □□□ would be addressed by raising the transition matrix to an exponent whose value represents the number of iterations of the mechanism to be applied.
Figure 44. A chromosome ■■■ that is only being subjected to recombination with an empty population is not capable of any transitions other than the transition to itself. When chromosomes □□□ and □□□ are added to the population, recombination can effect transitions from ■■■ to any member of {■■■, □□□, □□□, □□□, □■□, □■□}. It is important to note that the transition probability from ■■■ to □□□ is still 0, because a single recombination cannot effect the transition from ■■■ to □□□.

Although the transition probability matrix for Figure 44 had a value of 0 for the transition from ■■■ to □□□, the square of the matrix has a nonzero value, indicating that the distance between ■■■ and □□□ is finite with respect to traversals at least two generations in length, as depicted in Figure 45. It is then possible to compute the probability that α can be transformed into β within a specified number of iterations of the mechanism.
Figure 45. The square of the transition matrix defines a much more connected digraph than Figure 44. This graph indicates that there are four paths from ■■■ to □□□ that can be traversed by recombination alone, with the population {■□□, □■□}. Each of these paths, {■■■→■□□→□□□, □□□→□■■→□□□, □■■→□□■→□□□, ■■■→□□□→□□□}, begins at chromosome ■■■, ends at chromosome □□□, and requires exactly two applications of the recombination operator.

Although it might be tempting to use this approach to compute transition probabilities associated with traversals of increasing lengths and attempt to compute an expected transition probability, it must be emphasized that transition probabilities are not
equivalent to distance measures. The simple, five state example depicted in Figure 46 can be used to effectively differentiate the two approaches.

Figure 46. A simple, five state example used to illustrate the difference between distances and transition probabilities.

Firstly, it is observed that even though the distance from A to itself should have a value of 0, the transition probability is 0.3. This point aside, it is also observed that the probability associated with the transitions from A to D and from A to E have the same probability, in spite of the fact that the distance from A to E cannot ever be traversed (since no path connects those states). Although this could be addressed by squaring the corresponding transition matrix, the product of a state vector corresponding to state A (i.e., [1.0, 0.0, 0.0, 0.0, 0.0]) with the squared transition probability matrix would be
[0.09, 0.78, 0.03, 0.10, 0.00], suggesting that the distance from A to C is somehow more
difficult to traverse than the distance from A to D, even though C must be reached before
D can be entered and must, therefore, be a lesser distance from A.

Of course transition probability matrices could be used (as part of a more complex
approach) to gather measurements that would be equivalent to those gathered by the
proposed empirical distance measure. This approach was investigated and it was
determined (using a correlation analysis) that each distinct iteration of the empirical
measurement algorithm does reproduce the behaviour expected by the product of an
initial population vector and the transition probability matrix, provided that any
nonmaximal values of the product vector are reduced to 0 (as would be expected if the
population size could only be guaranteed to support the most likely avenues of traversal).

The correlation between the distance rankings that are specified by adapted state
transition matrix approach and those that are specified using the proposed empirical
measure of genotypic spatial distance is depicted in Figure 47 for an origin genotype ■■■■■■■■
and the explorative component established by a population composed
entirely of genotypes ■■■■■■■■, ■■■■■■□, ■■■■■□□, and ■■■■□□□. The $R^2$
value of 0.9938 clearly validates the proposed empirical measure against this adapted
transition matrix model of the explorative component, and while it is thus possible to
derive similar rankings with either technique, the construction of the transition matrix for
the explorative component is nontrivial, and for conducting analyses more rapidly, the
empirical measure is preferable.
Figure 47. The correlation between the rankings generated using a state transition matrix against those generated by the proposed empirical approach to genotypic spatial distance measurement.

Furthermore, given that transition probabilities could conceivably be used to replicate the behaviour of the empirical distance measure, it should be stressed that the ability of the empirical distance measure to function in the absence of transition probabilities can often prove beneficial. Using the proposed approach, it will be possible for developers exploring new variational operators for the genetic algorithm to determine the neighbourhood structures that are induced by those operators without needing to determine the exact transition probabilities. It could also be used as a means to empirically verify proposed transition probabilities, which are often subtle and hard to calculate.
To further validate the effectiveness of the empirical approach, it is expected for a measure specified for mutation alone that the empirically determined ranking should be strongly correlated with Hamming distance determined ranking. In fact, the ranks assigned by each measure are identical (Figure 48), further emphasizing the representativity of the empirical approach.

Figure 48. The ranking of chromosomes by Hamming distance from the origin are identical to those generated empirically (for a mutation operator alone).

The comparison introduced in this chapter, between the transition probability defined approach and the empirical approach to genotypic spatial distance measurement, clearly demonstrates that the empirical approach does closely resemble the actual traversal of genotypic space by the explorative mechanism, under the constraints and conditions specified. The empirical distance measurement approach is, thus, ideally suited as a
baseline against which the measures of recombinational and explorative mechanism traversal distances can be evaluated.

Having established the validity of the proposed approach to empirical distance measurement, it is now possible to comparatively evaluate the representativity of those distance measurement approaches that require a lesser computational expense (relative to the computational expense of an empirical approach). The chart (from Figure 2) depicting the overview of the complete body of research and developed that preceded application is revisited in Figure 49. At this point, both mutational and recombinational distance measures have been explored thoroughly, and three integrations of mutation and recombination have been proposed (i.e. the Euclidean distance from the population centroid, and the two approaches that were derived from the measures of recombinational distance). Each of these novel measures considers the traversal of genotypic spatial distances by the complete explorative component of the evolutionary mechanism, and not simply those traversals possible with a mutation operator. As these novel measures are expected to outperform the Hamming distance approach used typically, it was necessary to develop an empirical approach for measuring actual traversals (by the evolutionary mechanism of the genetic algorithm) to ensure that relative performances could be fairly evaluated.

It is now possible to proceed, in Chapter 6, with the comparative evaluation of the novel measures of genotypic spatial distance measurement against an approach using the Hamming distance.
Figure 49. As depicted in the graphic overview first introduced in Chapter 1, the development of an empirical measure of the genotypic spatial distance traversed by the variational operators of the genetic algorithm can be used to conduct a fair evaluation of the different approaches to genotypic spatial distance measurement. Chapter 6 is the comparative evaluation of the three novel approaches and an approach that uses the Hamming distance measure.
Chapter 6

Comparative Evaluation of Measures

Overview

The empirical measure of distance was designed for conducting comparative evaluations of the proposed novel approaches to genotypic spatial distance measurement. Although it has already been demonstrated that the empirical measure is representative of the manner in which the explorative component of the evolutionary mechanism traverses genotypic space, it must again be emphasized that the empirical approach was developed to facilitate the calibration and evaluation of measurement approaches and was not intended for a data visualization application itself. Accepting that each of the four proposed approaches to determining the centroid of population (i.e. the use of the mode, the median, and both a binary and real-valued representation of the mean) are all highly similar, the comparative evaluations presented in this chapter are conducted between: the Hamming distance from the chromosome designated as the origin, the Euclidean distance
from the real-valued mean population centroid, and the two approaches that were derived from the measures of recombinational distance introduced in Chapter 3, respectively referred to as the low resolution and high resolution integrated distance measures.

Although it has been demonstrated that each of the four measures to be evaluated can be computed independent of the fitness function being optimized, the three novel measures (i.e. those that include consideration for the recombination operation) cannot be computed independent of the population. Consequently, when these measures are compared against a Hamming distance approach to genotypic spatial distance measurement, it is necessary to specify any significant properties that the current population may have. To ensure that the conclusions derived from these preliminary comparisons are generalizable, the comparisons in this chapter were conducted for three types of randomly generated population – populations that exhibit a uniform, linear, or exponential distribution, respectively.

This chapter conclusively demonstrated that the Hamming distance approach to measurement was not representative of the manner in which the explorative component of the evolutionary mechanism traverses genotypic space. Furthermore, it was demonstrated that each of the three novel measures (i.e. the distance from the centroid and the low resolution and high resolution integrated distance measures) is much more representative than the Hamming distance. Although these measures incur a greater computational expense, the increased representativity can be used to enhance the adaptive landscape visualizations significantly – a fact that is demonstrated in Chapter 7.
6.1 Experimental Design

To ensure that any conclusions based on observed differences in accuracy are generalizable, the empirical distance from an origin chromosome to every other chromosome in a genotypic space defined for binary candidate solution configurations of length six is computed used the approach described in the previous chapter, for operators defined by several unique populations. It is again stressed that every approach that has been proposed thus far was fitness function independent. Furthermore, since there is no reason to believe that the application of the mutation and recombination operators discussed would yield any observations for relatively long chromosome lengths that would not also be observable with shorter lengths, the results presented here for binary chromosomes of size six are expected to be generalizable to larger chromosomes.

The first five populations to be examined were developed specifically to illustrate certain properties or shortcomings specific to each approach, and the remaining 60 populations were generated randomly, based on several observations of the distribution of population as it evolves with a genetic algorithm. Although it is certainly possible to compare the different approaches on a uniform random sample the space of possible populations, it cannot be guaranteed that a uniform sample of space of possible populations would accurately represent a sample of the populations that would be typical of actual instances of the genetic algorithm. Since the distribution of the population is certainly not independent of the function being optimized, the results in this section only generalize to the space of populations that exhibit particular distributions. That being said, the different
distributions were derived from observations of the population of a genetic algorithm as it was used to search the space of a two-dimensional Rastrigin's function. Of the 60 populations generated randomly for this analysis, the first 20 exhibit a uniform distribution - each was constructed from six unique chromosomes that were randomly sampled from the solution space according to a uniform distribution. The next 20 populations exhibit a linear distribution - five unique chromosomes were randomly sampled from the solution space according to a uniform distribution, and were placed in the population in the ratio 5:4:3:2:1. The final 20 populations exhibit an exponential distribution - five unique chromosomes were randomly sampled from the solution space according to a uniform distribution, and were placed in the population in the ratio 8:4:2:1. Consequently, following a comparative evaluation of the four approaches on each of the populations selected for illustrative purposes, three distinct evaluations of the approaches will be performed, with results generalizable to the space of possible populations that exhibit uniform, linear \((y = x + 1)\), and exponential distributions \((y = 2^x)\), respectively.

To evaluate the performance of each of the proposed genotypic spatial measures, it is necessary to quantify the degree to which the ranking of genotypes according to the proposed measure corresponds to the ranking established by the empirical distance measure. The performance measure selected for this task was the Spearman's rho rank correlation coefficient (Spearman 1987), which is a nonparametric measure of statistical dependence used to compare the degree to which two ranked variables are correlated. Since each genotype can be assigned ranks according to its measured distance from a designated point (using either the empirically measured distance or one of the four
proposed distance measures), then, in the context of this investigation, Spearman's rho would provide a means by which the representativity of a proposed distance measure could be evaluated. Should the measure being evaluated assign the same ranks as those assigned by the empirical measure, it would be an indication that the proposed measure is maximally representative of the genotypic spatial distances that are actually traversed.

The Spearman's rho rank correlation coefficient ($\rho$), is evaluated using the formula:

$$\rho = 1 - \frac{6 \sum (d_i)^2}{n(n^2 - 1)}$$

where $n$ is the number of assigned ranks and $d_i$ is the difference between the ranks assigned to the $i^{th}$ element by the two measures. The formula yields a value between -1 and +1, inclusively, with a value of +1 indicating that the two rankings assigned are in perfect agreement, and a value of -1 indicating perfect disagreement.

The first population used in the comparative evaluation had a cardinality of seven and was comprised of genotypes ■□■■■■, ■■■□■, and ■■■■□. The population was exponentially distributed, with four copies of ■□■■■■, two copies of ■■■□■, and a single copy of ■■■■□. From the genotype ■□■■■■ that was designated as the origin, each of the four approaches to genotypic spatial distance measurement (the Hamming distance from the origin, the Euclidean distance from the population centroid, and the low resolution and high resolution integrated distance measures) were used to rank (in
increasing order with tie-breaking) every chromosome in genotypic space according to the distance from the origin. Each set of results is compared with the ranking of chromosomes assigned by the empirical distance measure, which, under the conditions specified in Chapter 5, is recognized to be sufficiently representative of the actual traversal of genotypic space by the explorative mechanism.

In spite of the fact that the Hamming distance measure does not include consideration for the traversal of genotypic space by the recombination operator, the Hamming distance defined ranks correspond reasonably well to those values specified to the standard established by the empirical distance measure (with $\rho = 0.856044$). Notwithstanding this adequate performance of the Hamming distance measure on this population, the Euclidean distance from the centroid exhibits a significant improvement in performance over the Hamming distance, as indicated by a reduction in increase in the rank correlation coefficient $\rho$ from 0.856044 to 0.996498. This is undoubtedly because the use of the centroid as a representation of the population does permit consideration of the effect of the recombination operator, whereas the Hamming distance from the origin offers no consideration for the effect of any possible recombination operation. The improvement in representativity of the Euclidean measure over the Hamming measure is evident.

It is important to recognize that the assumption (required by the unary recombination paradigm) that the population and, thus, the operators, are essentially fixed by the current population, entails that the observed traversals of genotypic space are, in a sense, anchored by the current members of the population. This is depicted in Figure 50 and
suggests that the close genotypic spatial proximity of the target to one or more population members often results in shorter traversals than a close proximity to the origin.

Figure 50. As a consequence of the assumption that the population will remain fixed for a period of time, the empirically generated traversals of genotypic space are effectively anchored to the subspaces defined by the chromosomes of the population.

Although this property is far from obvious, it is easily demonstrated using the space of genotypes of length $\lambda = 4$ and operators defined for a singleton population \{□■■■\}. Since only a subset of the values from the transition probability matrices associated with mutation, recombination, and the combination of the two (i.e. the complete explorative component of the evolutionary mechanism) are necessary to demonstrate this property, these matrix fragments are presented as Tables 2, 3, and 4. The matrix fragment associated with mutation indicates (as expected) that the likelihood of mutation transforming ■■■■ into ■■■□ is equivalent to the likelihood of mutation transforming ■■■■ into ■□■■.
Table 2. Part of the transition matrix associated with mutation, for $\lambda = 4$. The probability that mutation will transform ■■■■ into ■■■□ (i.e., 0.1055) is contained in the cell at row ■■■■, column ■■■□, and is equivalent to the probability that mutation will transform ■■■■ into ■■■□.

However, it is clear from the matrix fragment in Table 3 below (for recombination with genotype □□■■) that recombination alone not capable of producing ■■■□ as an offspring of ■■■■, but is capable of producing ■□■■ (provided the second index is recombined).

Table 3. Part of the transition matrix associated with recombination, for $\lambda = 4$ and a population containing only the genotype □□■■.
It should be noted that, for the previous example (wherein the population contains only the genotype □□■■), the recombination of a genotype □□■■ with an element of the population (of which only □□■■ is available) must always produce the genotype □□■■ as the offspring. This would entail that the genotype □□■■ is actually an absorbing state for the transition diagram that would be associated with the recombinational operator alone. Furthermore, this reinforces that a simulation where only the chromosome □□■■ is available for recombination would be expected to converge on that chromosome quickly, if not for the effect of the mutation operator.

When the effect of the two operators are combined to comprise the complete explorative component of the evolutionary mechanism, it is clear that the probability of ■□■■ being produced as after a single application of the explorative component is significantly greater than the probability that ■■■□ will be produced.

<table>
<thead>
<tr>
<th></th>
<th>■■■■</th>
<th>■■■□</th>
<th>■■□■</th>
<th>■□■■</th>
<th>■□□■</th>
<th>...</th>
<th>■□■■</th>
</tr>
</thead>
<tbody>
<tr>
<td>■■■■</td>
<td>0.1846</td>
<td>0.0615</td>
<td>0.0615</td>
<td>0.0205</td>
<td>0.1318</td>
<td>...</td>
<td>0.0127</td>
</tr>
<tr>
<td>■■■□</td>
<td>0.0967</td>
<td>0.1494</td>
<td>0.0322</td>
<td>0.0498</td>
<td>0.0791</td>
<td>...</td>
<td>0.0264</td>
</tr>
<tr>
<td>■■□■</td>
<td>0.0967</td>
<td>0.0322</td>
<td>0.1494</td>
<td>0.0498</td>
<td>0.0791</td>
<td>...</td>
<td>0.0264</td>
</tr>
<tr>
<td>■■■□</td>
<td>0.0557</td>
<td>0.0732</td>
<td>0.0732</td>
<td>0.1260</td>
<td>0.0498</td>
<td>...</td>
<td>0.0557</td>
</tr>
<tr>
<td>■■□■</td>
<td>0.0791</td>
<td>0.0264</td>
<td>0.0264</td>
<td>0.0088</td>
<td>0.2373</td>
<td>...</td>
<td>0.0205</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>■□□■</td>
<td>0.0127</td>
<td>0.0146</td>
<td>0.0146</td>
<td>0.0205</td>
<td>0.0381</td>
<td>...</td>
<td>0.1846</td>
</tr>
</tbody>
</table>

Table 4. Part of the transition matrix associated with the application of the complete explorative component of the evolution mechanism (i.e. both mutation and recombination) to genotypes of length \( \lambda = 4 \) when the population is the single chromosome □□■■.
Furthermore, as the approach makes no assumptions as to the size of the population (to remain maximally general), the minimum population assumption permits the assignment of only those nearest chromosomes (in this case, ■□■■ and □■■■) to the next concentric zone. Thus, except for those traversals of genotypic space that are of a trivial length, the proximity of a chromosome to a member of the population becomes more significant than the proximity of the chromosome to the origin.

The low resolution and high resolution integrated distance approaches were both observed to exhibit a much more representative ranking of chromosomes than the Hamming distance approach, with rank correlation coefficient $\rho$ values of 0.991164 and 0.99675, respectively. This would indicate that, for this population of chromosomes ■■■■■□, ■■■■□■, and ■■■□■■, in the ratio 4:2:1, the high resolution integrated distance is more representative than the approach that uses the Euclidean distance from the centroid, which is, itself, more representative than the low resolution integrated distance approach.

The second population used in the comparative evaluation (constructed to be highly dissimilar from the previous) was comprised of candidate solution configurations □□□□□■, □□□□■□, and □□□■□□ in the ration 4:2:1 (for a total population cardinality of 7). Contrary to the first population, the elements of the second population are, at least with respect to mutational distance, very far from the origin chromosome.

The Hamming distance approach fares very poorly, exhibiting a rank correlation coefficient $\rho$ of -0.054624. Although the poor performance of the Hamming distance is
not entirely unexpected since virtually every recombination operation is capable of producing each offspring, the degree to which the ranks generated using the Hamming distance measure are uncorrelated with those ranks generated by the empirical distance measure is quite significant. To further demonstrate the significance of this result, consider that an approach whereby the ranking of genotypes is accomplished using a random permutation of possible ranks would be expected to yield a rank correlation coefficient $\rho$ of 0.0. This entails that, for the population in question, a random permutation of the possible ranks to which genotypes could be assigned could be expected to be more representative of the manner in which genotypic space is traversed than an assignment of rankings using the Hamming distance measure.

It is noteworthy that some of the short traversals (in terms of the empirically measured distance) do correspond to some of the shorter Hamming distance assignments, such as the distance from the origin chromosome ■■■■■■ to chromosomes ■■□■■■, ■□■■■■, and □■■■■■. These traversals of genotypic space can be accomplished using either the mutation and recombination operation, and do not necessarily require the application of both. For most chromosome pairings however, the Hamming distance rank assigned is quite different from the rank assigned by the empirical distance measure.

Unfortunately, the rankings derived from the Euclidean distance from the centroid are also unrepresentative of the ranks determined empirically, with a rank correlation coefficient $\rho$ of 0.52555. This is obviously a consequence of the population centroid calculation, which (using a real-valued representation of the mean value for each
chromosome index) is computed to be at a great Euclidean distance from the origin chromosome. Consequently, short traversals (such as the traversal of genotypic space from chromosomes □□■□□ to chromosome □□□■□) are assigned large distance values because the centroid is so mutationally distance from the origin.

The low resolution and high resolution distance approaches suffer similarly, exhibiting ρ values of 0.315052 and 0.587146, respectively. Notwithstanding any room for improvement that might be possible with a more specialized approach, the high resolution integrated distance measure is most representative, followed by the approach that uses the Euclidean distance from the centroid, which is itself followed by the low resolution integrated distance measure.

The next three populations were constructed such that the frequency of a genotype value of one at each index would be the quotient of one over the size of the population, for population sizes 6, 3, and 2. Thus, these populations were constructed from the candidate solution configuration sets: {□□□□□□, □□□□□□, □□□□□□, □□□□□□, □□□□□□, □□□□□□}, {□□□□□□, □□□□□□, □□□□□□}, and {□□□□□□, □□□□□□}, respectively.

Additional comparative evaluations were conducted with these artificial populations because they were constructed such that there would be no advantage to the application of the recombination operator over simply using the mutation operator – since all offspring are equally probable and reachable with a single application of the recombination operator, the mutational distance alone is expected (largely) to provide the ranking. It must be emphasized that these populations do not necessarily entail that the
empirical distance measure will assign the same ranks as if the recombination operator was not applied – the recombination operator is still applied by the empirical approach, suggesting that traversals of genotypic space will once again be anchored by the current chromosomes of the population, as described previously.

The ranking of chromosomes according to increasing Hamming distance from the origin is highly correlated to the ranking derived from the empirical distance measure, and, as expected, the ranking of chromosomes according to increasing Euclidean distance from the centroid produced nearly the same rankings as those derived from the Hamming distance approach. For population sizes 2 and 3 the rank correlation coefficient between both the Hamming and the Euclidean derived rankings and the empirically derived rankings were 0.867284 and 0.972746, respectively. For population size 6, the rank correlation coefficient value for the approach using the Hamming distance from the origin was 1.000000, while the rank correlation coefficient value for the Euclidean approach was 0.998798.

The integrated distance approaches, applied to the same populations, exhibited precisely the same rank correlation coefficient values as those exhibited by the Hamming distance approach. Thus, for these populations alone, there was no discernable advantage to the use of either the Euclidean distance from the centroid or the integrated distance measures over the Hamming distance approach. This is, of course, unsurprising as these populations were engineered specifically to ensure that the effect of the recombination operator would be indiscernible from the effect of the mutation operator.
Having explored five populations constructed specifically to explore certain properties of each approach to genotypic spatial distance measurement, it remains to perform comparative analyses using the randomly generated populations from the spaces of populations that exhibit uniform, linear \((y = x + 1)\), and exponential distributions \((y = 2^x)\), respectively. As noted previously, the selection of these distributions was based on observations of the population of a genetic algorithm that was being used to search the space of candidate solutions to a two-dimensional Rastrigin's function. However, to improve the generalizability of the comparative evaluations, randomly generated populations of similar distributions were used for the analysis, rather than using populations associated with a particular fitness function.

The random sampling of the different population spaces allows for the treatment of the sum of the squared error (between the ranks generated empirically and the ranks generated using one of the proposed measurement approaches) as though it were a random variable. It is thus possible to generate a confidence interval around each of the median rank correlation coefficients and, provided the values are normally distributed, use a paired student t-test of significance where necessary to determine which of the four approaches is most representative of the ranks generated by the empirical approach.

As it will be evident that 15 hypotheses are tested by the statistical analyses included, the significance level has been adjusted using the Bonferonni correction from \(\alpha = 0.05\) (for a 95.0000% level of confidence in a single test, for a net 39.7214% level of confidence...
over all 18 tests) to $\alpha = 0.002778$ (for 99.7222% level of confidence in a single test, for a net 95.1163% level of confidence over all 18 tests).

The degree to which each of the four approaches to genotypic spatial distance measurement could be used to generate ranks that are representative of those ranks achieved using the empirical distance measure was first assessed on the 20 population sampled from the space of all uniformly distributed populations of size six. The median values (and Thomson-Savur confidence intervals) for the Spearman's rho rank correlation coefficient between the empirically generated ranks and those generated by each of the Hamming distance from the origin, the Euclidean distance from the centroid, and the low and high resolution integrated distances, were 0.231233 [-0.013874, 0.390716], 0.911315 [0.865236, 0.938302], 0.817623 [0.756262, 0.871680], and 0.942868 [0.912614, 0.963198], respectively. These are depicted in Figure 51. Notwithstanding the considerable improvement exhibited by the low resolution integrated distance measure over the Hamming distance measure, the low resolution integrated distance and the Hamming distance exhibited the poorest performance. Although the median rank correlation coefficients indicated that the high resolution integrated distance measure was superior (in terms of representativity) to the Euclidean distance from the centroid, since the confidence intervals overlap further significance testing is required. This testing was performed using the student t-test for statistical significance, and although this is not the most powerful approach by which statistical significance could be evaluated for this investigate, the test is sufficiently powerful to demonstrate the desired properties for the novel measures of genotypic spatial distance.
Figure 51. The relative representativity of each of the approaches for populations of uniform distribution, as measured by the rank correlation coefficients with the rankings generated using the empirical distance measure. Clearly, the Hamming distance measure is outperformed by all three of the proposed novel approaches.

Although the normality plots of the rank correlation coefficient for the Euclidean distance exceeds an $r^2$ correlation coefficient of 0.95 (suggesting that the values are normally distributed, the $r^2$ correlation coefficient of the normality plot for the high resolution integrated distance does not ($r^2 = 0.921$), indicating that a nonparametric approach to significance testing is necessary – the significance testing will be performed on the data value rankings instead of the data values themselves.
Normality Plot of the Sum of Squared Error
(Euclidean Distance, Uniformly Distributed Populations)

$R^2 = 0.9585$

Normality Plot of the Sum of Squared Error
(High Resolution Integrated Distance, Uniformly Distributed Populations)

$R^2 = 0.921$

Figure 52. Two of the normality tests, indicating that nonparametric statistical significance testing must be used.
A paired student t-test on the ranks yields a p-value less than 0.0001 – since this value is less than significance level 0.002778, the difference between the mean sums of squared error is statistically significant and it can be stated that the ranks generated by the high resolution integrated distance measure are more representative of the ranks generated by the empirical distance measure than those generated using the Euclidean distance from the centroid.

The representativity of the ranks generated by each the four approaches to those ranks generated using the empirical distance measure was then assessed on the 20 population sampled from the space of linearly distributed populations of size six. The rank correlation coefficients between the empirically generated ranks and those generated by each of the Hamming distance from the origin, the Euclidean distance from the centroid, and the low and high resolution integrated distances, had values of 0.084146 [-0.079350, 0.277644], 0.922631 [0.853674, 0.952404], 0.640351 [0.425126, 0.773546], and 0.926505 [0.881330, 0.957784], respectively. These are depicted in Figure 53.

Once again, the Hamming distance is substantially outperformed by each of the novel approaches proposed, and since the confidence interval around the median rank correlation coefficient associated with the low resolution integrated distance is not overlapped by any other interval, further statistical significance testing is not warranted. Thus, it remains only to determine whether the observed improvement of the high resolution integrated distance over the Euclidean distance is statistically significant. As in the previous significance test, the lack of a normal distribution to the rank correlation
coefficient values (normality plot $r^2$ values of 0.8556 and 0.8403 for the Euclidean and high resolution integrated distance, respectively) necessitates that a nonparametric paired student t-test be performed. With a two-tailed p-value of 0.5018, the superiority of the high resolution integrated distance over the Euclidean distance is not statistically significance. Nevertheless, both approaches are superior to the low resolution integrated distance measure, which is, itself, dramatically superior to the Hamming distance measure that is typically used.

![Figure 53. The relative representativity of each of the approaches for populations of linear distribution, as measured by the rank correlation coefficients with the rankings generated using the empirical distance measure. Clearly, the Hamming distance measure is outperformed by all three of the proposed novel approaches.](image)
As the final evaluation of representativity, the rank correlation coefficients were assessed on the 20 population sampled from a space of exponentially distributed populations of size six. The coefficients between the empirically generated ranks and those generated by each of the Hamming distance from the origin, the Euclidean distance from the centroid, and the low and high resolution integrated distances, had values of 0.220914 [-0.043784, 0.301912], 0.909861 [0.850000, 0.966586], 0.596571 [0.381410, 0.757510], and 0.938044 [0.882166, 0.974450], respectively, as depicted in Figure 54.

![Representativity to Empirically Generated Values for Populations of Exponential Distribution](image)

**Figure 54.** The relative representativity of each of the approaches for populations of an exponential distribution, as measured by the rank correlation coefficients with the rankings generated using the empirical distance measure. Clearly, the Hamming distance measure is outperformed by all three of the proposed novel approaches.
As in both previous analyses, the Hamming distance is substantially outperformed by each of the novel approaches, and the confidence interval around the median rank correlation coefficient associated with the low resolution integrated distance is not overlapped by the intervals around the rank correlation coefficient values of either the Euclidean or high resolution integrated distance measures. It again remains to determine whether the observed improvement of the high resolution integrated distance over the Euclidean distance is statistically significant, and as the data values are not normally distributed (normality plot $r^2$ values of 0.7646 and 0.7488 for the Euclidean and high resolution integrated distance, respectively), a nonparametric paired student t-test must again be used to test the significance of this difference. With a two-tailed p-value of 0.0007, less than significance level 0.002778, the difference in performance between the high resolution integrated distance and the Euclidean distance is statistically significant. It can, thus, be concluded that, of the four approaches, the high resolution integrated distance measure generates the ranks that are most representative of those generated by the empirical distance measure.

### 6.2 Analytical Results

The comparative analyses presented in this chapter have yielded three conclusions, the first of which being the affirmation that the ranking of chromosomes according to their Hamming distance from the origin is not representative of the ranking of chromosomes according to their empirically determined genotypic spatial distance. Indeed, since the Thomson-Savur confidence intervals around the median rank correlation coefficient
values for the Hamming distance all contain the value 0.0, it cannot be stated that the
Hamming distance can be used to provide ranks that are any more representative than a
random permutation of the possible ranks. This clearly indicates that the Hamming
distance is a poor measure of the distances traversed through genotypic space by the
explorative component of the evolutionary mechanism.

The second conclusion that can be drawn from the results of the comparative analyses is
that the ranking of chromosomes according to their Euclidean distance from the centroid,
or either the low resolution or high resolution integrated distance from the origin, is much
more representative of the ranking of chromosomes according to their empirically
determined genotypic spatial distance. This clearly establishes that the use of any of these
measures would improve the representativity of measured distances over the Hamming
distance, which remains the most typical approach to genotypic spatial distance
measurement when binary chromosomes are used.

Finally, it can be concluded that, of the approaches investigated, the high resolution
integrated distance measurement can be used to generate the most representative
chromosomal rankings for populations that are either uniformly or exponentially
distributed. Furthermore, for populations exhibiting a linear distribution, the high
resolution integrated distance measurement is as effective as the Euclidean distance from
the centroid, and remains superior to approaches based on the low resolution integrated
distance measure and the Hamming distance measure.
Although the high resolution integrated distance measure does (overall) exhibit an improvement over the approach that uses the Euclidean distance from the centroid (on those populations from the set of uniformly, linearly, or exponentially distributed populations), the associated increase in computational expense is not trivial. As the weighting that defines the trade-off between representativity and computational expense should be defined by the user, neither the high resolution integrated distance nor the Euclidean distance from the centroid should be discounted entirely.

At this point it has been conclusively demonstrated that the Hamming distance is largely unsuitable as a measure of the genotypic spatial distance traversed if both mutation and recombination operators are employed by the evolutionary mechanism, and three alternative (and more suitable) measures have been proposed and thoroughly analyzed from both a theoretical and empirical perspective. These novel measures can now be applied to the task of constructing exploratory graphics, specifically the adaptive landscape data visualization technique, that are more representative of the manner in which the genotypic space would actually be traversed and thus, by the criteria established previously in Chapter 2, are more inherently more valuable.
Chapter 7

Applications and Future Work

The model proposed by Wijk (2005) established a methodology for assessing the value of a data visualization strategy, by computing the profit associated with the use of a visualization as the difference between the weighted improvement in the knowledge of the person viewing the visualization and the expenses associated with the development and use of the visualization. The appeal of the adaptive landscape technique to genetic algorithm researchers is in no small part due to the fact that the technique predates the genetic algorithm by approximately four decades, has no further development costs, and is highly intuitive, effectively minimizing any costs associated with training users in its usage. Furthermore, ignoring the exploration process from Wijk's model (by fixing the specification parameters associated with the visualization), the exploration cost term from Wijk's economic model can also be ignored and the cost associated with the technique can be assessed by examining the sessional costs alone. The formula for computing the profit of the adaptive landscape visualization can then be simplified, for this discussion, to $W(\Delta K) - nmC_s$. 

221
Although adaptive landscape construction does entail both a rendering of the discrete three-dimensional surface and a linear time complexity pass through the population to identify the points of the surface that correspond to the members of the population, it is the transformation of the chromosome space into the three-dimensional surface that, if properly performed (with the necessary application of an effective dimensionality reduction technique), incurs the highest computational expense. With the chromosome space employed by the genetic algorithm almost always having a dimensionality greater than two, (and once the evaluation function is considered, greater than three), the dimensionality of the chromosome space must be decreased before the landscape can be visualized. In order to properly reduce the high dimensionality typically present in a chromosome space, a multidimensional scaling technique requiring a significant computational expense must be employed. As noted previously, it is not uncommon for researchers to instead limit usage of the adaptive landscape visualization to instances where the domain of the objective function being optimized (the phenotypic space) is two-dimensional and then disregard the dimensionality of the chromosome space in favour of the dimensionality of the phenotypic space. It must be acknowledged that such an approach can significantly distort the perceived interchromosomal distances from the visualization and does not address the fact that the variational genetic operators employed by the genetic algorithm operate on representations (i.e. chromosomes) of a solution, and not upon a candidate solution itself. In order to properly construct an adaptive landscape visualization, the notion of distance or neighbourhood that is used to induce a structure on the chromosome set to create the chromosome space must convey the proximity of the genotypes of the candidate solutions, rather than the phenotypes. The high dimensionality
of the resulting chromosome space must then be properly reduced by constructing a lower dimensional representation that will not mislead the user, using a multidimensional scaling technique such as the approach developed by Sammon in 1969. Although the research presented in this thesis has conclusively demonstrated that the Hamming distance measure does not accurately convey the proximity of candidate solutions in genotypic space, the three novel approaches to genotypic spatial distance measurement, based on the Euclidean distance to the population centroid and the integration of the low resolution and high resolution measures of recombinational distance with a measure of mutational distance, are suitable for use in a dimensionality reduction technique.

Although there are numerous approaches to dimensionality reduction, the initialization of Sammon's nonlinear mapping (Sammon 1969) to an arbitrary two-dimensional set of points with the same cardinality as the higher dimensional chromosome space facilitates the comparison of the resulting lower dimensionality representation against simpler visualizations. Furthermore, the iterative nature of the technique facilitates the development of an animate extension of the multidimensionally scaled adaptive landscape visualization. As the set of points is iteratively transformed into a set more representative of the higher dimensional space (by repeatedly measuring the interchromosomal distances between every unique pairing in the chromosome space and incrementally adjusting the lower dimensionality representation), it is possible to determine the degree of accuracy with which the resultant lower dimensionality representation has preserved the interchromosomal distances of the higher dimensionality space.
As noted previously, since the image of the adaptive landscape (as defined in Wijk's model) is perceived as the set of relative positions between the points on the surface, it is absolutely essential that the interchromosomal distances of the lower dimensionality representation match those of the higher dimensionality representation. Furthermore, if the user perceives a feature in the lower dimensionality representation that is not indicative of a feature in the actual, high dimensionality chromosome space, the consequence can be a change $\frac{dK}{dt}$ to the knowledge of the user that includes a piece of incorrect information. As this would drive the $\Delta K$ component of the profit formula down (possibly even resulting in a negative $\Delta K$), failure to use a proper multidimensional scaling technique reduces the value of an adaptive landscape visualization to a point where it may provide only false insight.

### 7.1 Multidimensionally Scaled Visualizations

Sammon's nonlinear mapping algorithm was previously employed in constructing visualizations of the genetic algorithm by Dybowski, Collins, and Weller (1996), but where their approach was applied only to specific populations to indicate the presence of multiple solutions, this investigation was motivated to construct representations of the higher dimensional chromosome space that could be used to construct adaptive landscapes that were more representative of actions of the genetic algorithm. By the application of Sammon's nonlinear mapping to the minimization of the stress between the pairwise distance matrices of the higher dimensional genotypic space and the lower dimensional representation (the extrusion of which is designated as the adaptive
landscape) the planar arrangement of candidate solution configurations is measurably more representative of the actual topology of genotypic space. Furthermore, since the novel approaches to genotypic spatial distance measurement introduced in this body of research can be used to compute the pairwise distance matrix for the chromosomes in the higher dimensional space, and since it was demonstrated in the previous chapter that these distance measurement approaches exhibit an improvement in representativity that is statistically significant, a multidimensional scaling of a grid of points corresponding to the Cartesian co-ordinate plane typically used in adaptive landscape construction will yield a two-dimensional arrangement of points that is measurably and quantifiably more representative of the actual topology of genotypic space. This planar arrangement of points can be transformed (using the Delaunay triangulation, 1934) into a two-dimensional surface from which a quantifiably more representative adaptive landscape can be constructed.

It is not uncommon for researchers to extrude an adaptive landscape visualization from a two-dimensional collection of points, derived from phenotypic space and arranged using the Cartesian co-ordinate plane. For a phenotype of dimensionality two, the use the Hamming distance from the origin as the measure of genotypic spatial distance would effect a rearrangement of points configured using the Cartesian co-ordinate system (in Figure 55, left) to an arrangement of chromosomes representative of the manner in which genotypic spatial distances are traversed by the mutation operator alone (in Figure 55, right). The points in these figures have been coloured according to their respective Hamming distance from a genotype composed of all zeros to improve readability.
Figure 55. The arrangement of genotypes on a Cartesian co-ordinate plane by the phenotypes they encode (left) is a misleading practice. A rearrangement using the mutational distance (i.e. the Hamming distance) yields a more representative arrangement (right), in which known structural features are actually visible.

Although it was concluded that the Hamming distance from the origin is not a suitable measure of genotypic spatial distances that would be traversed by both mutational and a recombinational genetic operators, it should be noted that the rearrangement of points using Sammon's nonlinear mapping actually manifests a pattern of gaps that appear to recursively bisect the subspace both horizontally and vertically. This pattern of gaps corresponds to the known presence of the aptly termed Hamming cliffs in this binary encoding - although two values on an axis might be representative of positive integers that are separated by a difference of one, if the greater of the integers has a radix of two and an integer exponent, their binary representations could be considered complementary. Since the two-dimensional representation that is created when the multidimensional scaling technique is applied with the interchromosomal distances computed using the Hamming distance has introduced a visible feature into the image of the adaptive
landscape that can be perceived and will introduce a positive contribution to the ΔK change in knowledge of the user, the value of the adaptive landscape visualization has been increased through the use of multidimensional scaling.

Furthermore, the transformation from the square form of the Cartesian co-ordinate plane to a circular plane of points will also inhibit the false perception that there are somehow exactly four points in the space that are furthest from the centre. As the perception of this feature would have contributed negatively to the knowledge of the user, its removal could also be considered a net increase in the value of the visualization. The adaptive landscapes that might be constructed for Rastrigin's function from the arrangements of Figure 55, left and right, are included in Figure 56, left and right, respectively.

Figure 56. The landscapes that are extruded from the patterns in Figure 55 for Rastrigin's function. It should be noted that, due to the coarse treatment of phenotype space, the ruggedness associated with Rastrigin's function is not visible.
Having been determined entirely using the Hamming distance from each chromosome to the origin chromosome, the use of the arrangement of points in Figure 55, right, instead of the grid of points in Figure 55, left, to construct the adaptive landscapes of Figure 56 right, in contrast with the landscape depicted in Figure 56, left, resulted in a visualization that is measurably more representative of the traversal of genotypic space by the mutation operator alone, regardless of the population. However, as the Hamming distance from a candidate solution configuration to the origin is not the most suitable approach to the measuring the genotypic spatial distances as they would be traversed by the explorative component (i.e. both the unary mutation operator and the binary recombination operator) of the evolutionary mechanism of the genetic algorithm, adaptive landscapes constructed for specific populations are much more representative with respect to traversal by both the variational operators of the genetic algorithm.

The use of the Euclidean distance from the centroid of the current population and the low resolution and high resolution integrated distance measures of genotypic spatial distance that were introduced in Chapter 4 effect the rearrangement of points (initially configured using the Cartesian co-ordinate system) to the chromosomes depicted in Figures 57, 58, and 59, left, for the populations indicated. Using the Delauney triangulation (1934) for a fitness function based on a two-dimensional Rastrigin's function results in the creation of the three-dimensional surface depictions of the adaptive landscapes that are depicted in Figures 57, 58, and 59, right, in which the quantifiable improvement in representativity is achieved.
Figure 57. The two-dimensional genotype arrangement (left) from which the landscapes (multidimensionally-scaled to three dimensions, projections right) were extruded, using Euclidean distance from the centroid (top) and the high resolution integrated distance measure (bottom), with population members \{■□■■□□, ■□■□□□, ■□□■□■, □■■■□□\} in the ratio 3:1:1:1. These approaches to the representation of genotypic space yield a reduction in stress over the landscape that would be extruded from a Cartesian co-ordinate plane. The genotypes depicted left are either black (if they belong to the current population) or have been shaded according to their Hamming distance from genotype ■■■■■■.
Figure 58. The two-dimensional genotype arrangement (left) from which the landscapes (multidimensionally-scaled to three dimensions, projections right) were extruded, using Euclidean distance from the centroid (top) and the high resolution integrated distance measure (bottom), with population members {■□□□■■, ■□□□■□, ■□□□□■, □■■□■■, □■□□■□} in the ratio 2:1:1:1:1. These approaches to the representation of genotypic space yield a reduction in stress over the landscape that would be extruded from a Cartesian co-ordinate plane. The genotypes depicted left are either black (if they belong to the current population) or have been shaded according to their Hamming distance from genotype ■■■■■■.
Figure 59. The two-dimensional genotype arrangement (left) from which the landscapes (multidimensionally-scaled to three dimensions, projections right) were extruded, using Euclidean distance from the centroid (top) and the high resolution integrated distance measure (bottom), with population members \{□■■■□, □■□□□, □■■■■, □■□□□\} in the ratio 3:1:1:1. These approaches to the representation of genotypic space yield a reduction in stress over the landscape that would be extruded from a Cartesian co-ordinate plane. The genotypes depicted left are either black (if they belong to the current population) or have been shaded according to their Hamming distance from genotype ■■■■■.
It is emphasized that it may be possible to achieve an even greater reduction in the stress associated with these visualizations, but this only suggests a weakness of the multidimensionally scaling technique (i.e. Sammon's nonlinear mapping), and should not be considered an indication that the Euclidean distance from the centroid and the high resolution integrated distance measures are less representative than the analysis in Chapter 6 indicated.

### 7.2 Additional Considerations

It must be emphasized that Sammon's nonlinear mapping approach was designed with the assumption that the matrix of distance values was generated using a true distance metric – notably that the pairwise interchromosomal distances are symmetric. Furthermore, it is again noted that image of the adaptive landscape is perceived as the set of relative positions between the candidate solution configurations on the surface and, consequently, there are no directed edges between points on the adaptive landscape to indicate whether the distance between genotypes portrayed by the landscape is representative of the distance traversed in either or both directions. This is a failing of the adaptive landscape visualization technique - the assumption made is that, for any pair of genotypes, it is as easy to transform the first genotype into the second as it is to transform the second genotype into the first. Thankfully, it may be possible to circumvent this limitation (with the use of glyphs at each point on the adaptive landscape surface) and avoid any increase to the expense associated with user training that was specified in Wijk's model (2005).
It should also be noted that Sammon's nonlinear map, like many multidimensional scaling approaches, is attempting to solve an optimization problem (i.e. a minimization of the discrepancy between the actual genotypic space and the lower dimensional representation) and, as such, is susceptible to all the problems normally associated with optimization. These weaknesses are inherent to multidimensional scaling in general, and cannot be addressed by increasing the representativity of the genotypic spatial distance measures.

It is also very rare that any approach to multidimensional scaling can construct a lower dimensional representation (from a complex high dimensional object) without at least some error – there will almost always be some data distortion. This limitation should not be taken as an indication that data visualization techniques should not at least attempt to minimize this error – the less data distortion, the more valuable the visualization. It is, again, emphasized that, in spite of the increase in computational expense associated with the minimization of the data distortion in a visualization, the decision to incur the additional cost must be at the discretion of the user.

Finally, it is important to recognize that since the proposed approaches to the measurement of both recombinational and integrated variational distances each entail the processing of an exponential number of schemata, researchers seeking to apply these techniques may be forced into the use of heuristics and approximations. This should not be considered dissuasive – since the task of constructing a complete depiction of the genotypic space being investigated does require pairwise treatment of the elements of a
set of candidate solutions with a cardinality that is, itself, exponential in terms of genotype length. This entails that visualization construction will often require the use of techniques (such as uniform sampling) that provide an approximate (as opposed to an exhaustive) depiction of points in the search space. Similarly, although the novel distance measures introduced in this dissertation are associated with nonpolynomial time complexities, several heuristics have been introduced to improve the feasibility of representative genotypic spatial distance measurement, and more should be investigated in the future.

7.3 Future Work

Although the approaches to the measurement of genotypic spatial distance that have been developed for this dissertation were motivated for the purpose of enhancing the adaptive landscape visualization approach for use with the genetic algorithm, the increased representativity that is achieved with these approaches can be easily applied to the measurement of population diversity. As it was noted previously, many of the different approaches to population diversity measurement in genotypic space are essentially aggregations of the genotypic spatial distance separated all possible pairings of the genotypes in the population (Wineberg and Oppacher 2003). It follows, naturally, that the low resolution and high resolution integrated distance measures introduced in this dissertation can be performed between all possible genotype pairings to quantify the diversity of the population. Since it was conclusively demonstrated that these measures of distance are more representative (of the manner in which genotypic space is traversed)
than the Hamming distance, it follows that an approach to population diversity measurement using the proposed integrated distance measures would yield a more representative quantification of population diversity than the customary approach wherein the pairwise Hamming distance is used for the assessment. It also follows that any attempt to perform cluster analyses on the populations used by genetic algorithms can be expected to yield more representative results - both population diversity measurement and cluster analysis in genotypic space represent fruitful avenues for future investigation.

It should also be evident that the empirical approach to genotypic spatial distance measurement that was developed for this body of research could conceivably be used to establish traversals of other high dimensional spaces, by other complex mechanisms, and need not be restricted to evolutionary algorithms. These traversals could then be used (as demonstrated in this dissertation) for the calibration and comparative evaluation of other possible measures of distance. It should also be noted that the empirical distance measurement approach presented in this dissertation was designed to be both fitness function and population size independent – it must be emphasized that apriori knowledge of the function being evaluated or the size of the population that will be used by the genetic algorithm may be useful in improving the degree to which the neighbourhood structure derived by the empirical distance measure matches the manner in which the genotypic space will be traversed. Although the independence and generalizability of the techniques presented in this dissertation is most definitely a strength of the proposed
approaches, more specialized derivations of the techniques presented herein may yield additional applications.

Finally, although the time complexity associated with the high-resolution integrated variational distance measure might seem prohibitive, this dissertation also demonstrated that low-complexity distance measure approximations (such as the Euclidean distance from the population centroid) represent a significant improvement over the standard Hamming distance approach. Although there might not be any way to avoid a large computation expense if a user decides that he or she requires a data visualization constructed from a maximally representative measure of distance, there may be several heuristics that would improve (i.e., reduce) the time complexity of the visualization process at the expense of the representativity (with this dissertation proposing techniques derived from the Euclidean distance from the population centroid and from a sampling of the schemata that represent the regions of genotypic space that can be reached by recombination). These could be employed by the users that are faced with constraints on the computational expense that can be afforded to the task of data visualization.
Chapter 8

Conclusions

The open problem addressed by this body of research is relevant to the domains of both evolutionary computation and data visualization; the adaptive landscape, an established approach to the visualization of genotypic space, was observed to distort the distance (between genotypes) that would be traversed by the operators of the genetic algorithm. This observation was not particularly surprising (given that the adaptive landscape visualization technique predates the genetic algorithm by four decades), nor was it entirely unknown to the research community. However, in spite of the established weakness inherent to the adaptive landscape approach, it remains in widespread usage by the community.

A comprehensive investigation of the relevant background literature revealed that even though some researchers are aware of the shortcomings of an adaptive landscape visualization approach to genotypic space, the inherent weaknesses are far from common knowledge; the typical approach to the generation of an adaptive landscape is the
To minimize the discrepancy between the space of candidate solution genotypes and the two-dimensional representation from which the landscape is extruded, a multidimensional scaling technique is required. These techniques begin with the construction of pairwise distance matrices for both the higher dimensional space and the lower dimensional representation. By adjusting the representation such that the error between these distance matrices is minimized, the representativity of the two-dimensional arrangement is maximized. Although the process entails an increase in computational expense, the resulting visualization is more accurate and thus (by conventional criteria) inherently more valuable. This approach can be applied to the visualization of the genotypic spaces traversed by the genetic algorithm, but the need for a representative measure of the genotypic spatial distances (traversed by the operators) cannot be overemphasized. The widespread use of the Hamming distance as a measure of genotypic
spatial distance is only valid if the space of candidate solution genotypes is to be traversed using a unary point mutation operator alone - since the genetic algorithm also employs a binary recombination operator, the Hamming distance dramatically distorts the genotypic spatial distances traversed by the evolutionary mechanism.

Faced with the difficulty of having to integrate measures of the distance traversed by both a unary operator and a binary operator, it might be concluded that the adaptive landscape visualization is not suited for depictions of complex genotypic spatial chromosomes. However, the adaptive landscape visualization technique remains highly intuitive for conceptualizing the processes of optimization, and since the domain of data visualization prefers an incremental model of visualization development over the replacement of established techniques with novel approaches, the adaptation and advancement of the adaptive landscape visualization is well justified. Furthermore, an accurate adaptive landscape is predicated on a more accurate understanding of the neighbourhood structure induced by the reproductive operators.

The principle requirement for the advancement of the adaptive landscape visualization is a representative measure of the distance traversed by the explorative component of the evolutionary mechanism. Having assessed this need, the task of developing such a measure can be further subdivided into the development of a representative measure for each of the explorative operators and a solid approach to the combination of the two. It was previously noted that representativity with respect to mutation is achieved by the Hamming distance measure, but representativity with respect to recombination is
significantly more difficult. Although a number of abstractions of the recombination operator have been explored and the concept of a recombinational landscape has been (somewhat) explored, the problem of developing representative measures of genotypic spatial distance is rarely approached from the perspective of data visualization or population diversity applications.

For any attempt to integrate two or more measures of distance (each of which being representative of the distance traversed by a distinct operator), it is crucial that the domain and range of each operator be equivalent. In the context of the genotypic space traversed by the genetic algorithm (to be depicted in the adaptive landscape visualization), the domain and range must each be the space of chromosomes that represent candidate solution configurations. However, since recombination is typically conceptualized as an operation that requires a pair of candidate solutions (i.e. a population of size two) to produce a single offspring chromosome, the domain of the recombination operator is the space of populations (of candidate solution configurations, with cardinality two) and not the space of individual genotypes. To facilitate the eventual integration of a measure of recombinational distance with a measure of mutational distance, it was necessary to develop a unary recombination paradigm – completely preserving the functionality of the recombination operator while redefining the domain to that of the mutation operator. In establishing this paradigm, it was important to recognize that the unary recombination operations applied to the population of a genetic algorithm are determined by the state of the current population but remain independent of the evaluation function being optimized. The preservation of this generalizability (i.e.
evaluation function independence) was a high priority objective throughout this research. It was also critical that the properties of a recombinational measure were sufficiently explored, and it was determined that a recombinational distance measure would not satisfy the conditions for classification as a metric (and, instead, would be classified as a quasimetric). Although this would complicate any depiction of genotypic space, it does not preclude the improvement of the adaptive landscape visualization technique.

Initially, the notion of the recombinational distance from one genotype to another was only assigned a value of zero if the two genotypes were equivalent or a value of $+\infty$ if no subset of the current population could be recombined with the first genotype to produce the second genotype as an offspring. If, on the other hand, it was possible (by recombining the first genotype with any subset of members of the current population) to produce the second genotype as an offspring of the first, the distance from the first genotype to the second was expressed as an unknown, but finite, constant value. This led to the operational definition of the distance resolution of a measure: the level of precision with which the distance from one genotype to another could be expressed. Following an extensive analysis, it became possible to increase the distance resolution of the initial measure of recombinational distance, and, subsequently, the two proposed measures of recombinational distance were denoted the low resolution and high resolution recombinational distance measures, respectively.

A preliminary investigation indicated that the integration of a proposed recombinational distance measure with a mutational distance measure would be problematic, and that a
representative measure of the genotypic spatial distance traversed by both operators together could not be achieved using a simplistic approach (i.e. an arithmetic, geometric, harmonic, or quadratic averaging approach). An abstraction of the evolutionary mechanism was derived in which recombination was treated as the operation whereby the mutational distance between chromosomes could be shortened. Equivalently, the recombination operator would (conceptually) be providing a platform from which the mutation operator could be applied to greater effect.

Although the reverse (i.e. that mutation can provide a platform from which recombination could be applied) would not be unreasonable, the ease with which the former can be computed (relative to the latter) lead to the definition of an integrated distance measure as a conventional mutational measure from points that are at established recombinational distances from the chromosome of origin. From the two novel measures of recombinational distance, the low resolution and high resolution measures of the genotypic spatial distance traversed (by the explorative component of the evolutionary mechanism) were defined.

Even with proposed measures that were developed from a strong theoretical foundation, there was not, unfortunately, a methodology by which these approaches could be calibrated or comparatively evaluated. Without known (and nontrivial) genotypic spatial distances against which the accuracy of a proposed measure can be tested, it is impossible to determine whether or not the measure is superior (i.e. more representative) than an existing technique. To this end, an empirical measure that determines the distance
between genotypic chromosomes by observing actual traversals (by the operators in question) was developed. This approach would permit researchers to establish the distance between chromosomes using the mechanism of traversal itself. Although an empirical approach would entail a computational expense significant enough to preclude its use in data visualization and population diversity applications, an empirical distance measure is invaluable for establishing test cases and for the evaluation of novel measures of genotypic spatial distance.

Naturally, the proposed approach to empirical distance measurement was not meant to emulate a genetic algorithm exactly. Instead, a genotype designated as the origin is subjected to operations that precisely mirror the exploratory component of the evolutionary mechanism of a simple genetic algorithm. With each application, a set of genotypes identified as nearest the origin genotypic are identified, and, thus, a set of concentric contoured zonal neighbourhoods around the origin can be constructed. The membership of a genotype to a neighbourhood can be interpreted as the ranking of the distance between this genotype and the origin, with respect to all finite genotypic spatial distances. Within the constraints specified, the validity of this approach was established using Vose's infinite population abstraction of the genetic algorithm as a model.

Having investigated and proposed novel approaches to the measurement of the genotypic spatial distances traversed by the explorative component of the evolutionary mechanism, and having established and validated an empirical distance measurement technique that can be used for comparative evaluation, it remained only to assess the relative
performance of the novel measures against the Hamming distance measure that is employed typically. The results of this analysis conclusively demonstrated that an approach to genotypic spatial distance measurement that uses the Hamming distance between genotypes is not representative of traversals made by the genetic algorithm and, consequently, the Hamming distance is not suitable for data visualization applications. Of the proposed measures, the integration of a mutational distance measure and the high resolution recombinational distance measure outperformed the Hamming distance approach dramatically, and was demonstrated to be the most representative approach for the sample populations of uniform and exponential distribution.

Finally, the novel measures of genotypic spatial distance were integrated into an implementation of Sammon's nonlinear mapping approach to multidimensional scaling, and this approach was used to enhance the standard adaptive landscape visualization approach by increasing representativity and decreasing data distortion. Although the resultant visualizations are more expensive to produce, computational expense must always be incurred at the discretion of the user, and the enhanced adaptive landscape visualizations remain highly intuitive. With insight accepted as the principle objective of data visualization, it is absolutely critical that graphical depictions not distort the data or otherwise mislead the user, and since it has been quantifiably demonstrated that these visualizations are more representative of the actual traversal of genotypic space by the genetic algorithm, the initial task of enhancing the adaptive landscape visualization technique by improving representativity has been achieved.
References


Mazza R. (2009), "Introduction to Information Visualization", Springer.


