NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

TESTING EFFECTIVENESS OF GENETIC ALGORITHMS FOR EXPLORATORY DATA ANALYSIS

by

Jason W. Carter

September, 1997

Thesis Advisor: Hemant K. Bhargava

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE September 1997	3. REPORT TYPE AND DATES COVERED Master's Thesis

4. TITLE AND SUBTITLE : TESTING EFFECTIVENESS OF GENETIC ALGORITHMS FOR EXPLORATORY DATA ANALYSIS

5. FUNDING NUMBERS

6. AUTHOR(S)
Carter, Jason W.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

Heuristic methods of solving exploratory data analysis problems suffer from one major weakness – uncertainty regarding the optimality of the results. The developers of DaMI (Data Mining Initiative), a genetic algorithm designed to mine the CCEP (Comprehensive Clinical Evaluation Program) database in the search for a Persian Gulf War syndrome, proposed a method to overcome this weakness: reproducibility -- the conjecture that consistent convergence on the same solutions is both necessary and sufficient to ensure a genetic algorithm has effectively searched an unknown solution space. We demonstrate the weakness of this conjecture in light of accepted genetic algorithm theory. We then test the conjecture by modifying the CCEP database with the insertion of an interesting solution of known quality and performing a discovery session using DaMI on this modified database. The necessity of reproducibility as a terminating condition is falsified by the algorithm finding the optimal solution without yielding strong reproducibility. The sufficiency of reproducibility as a terminating condition is analyzed by manual examination of the CCEP database in which strong reproducibility was experienced. Ex post facto knowledge of the solution space is used to prove that DaMI had not found the optimal solutions though it gave strong reproducibility, causing us to reject the conjecture that strong reproducibility is a sufficient terminating condition.

14. SUBJECT TERMS Gulf War Illnesses, Genetic Algorithm, Data Analysis, Data Mining	15. NUMBER OF PAGES 76
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFI- CATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

Approved for public release; distribution is unlimited

TESTING EFFECTIVENESS OF GENETIC ALGORITHMS FOR EXPLORATORY DATA ANALYSIS

Jason W. Carter
Lieutenant, United States Navy
B.S., University of Missouri - Rolla, 1990

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
September, 1997

Author: _____
Jason W. Carter

Approved by: _____
Hemant Bhargava, Thesis Advisor

_____
William Haga, Associate Advisor

_____
Reuben T. Harris, Chairman
Department of Systems Management

ABSTRACT

Heuristic methods of solving exploratory data analysis problems suffer from one major weakness – uncertainty regarding the optimality of the results. The developers of DaMI (Data Mining Initiative), a genetic algorithm designed to mine the CCEP (Comprehensive Clinical Evaluation Program) database in the search for a Persian Gulf War syndrome, proposed a method to overcome this weakness: reproducibility -- the conjecture that consistent convergence on the same solutions is both necessary and sufficient to ensure a genetic algorithm has effectively searched an unknown solution space. We demonstrate the weakness of this conjecture in light of accepted genetic algorithm theory. We then test the conjecture by modifying the CCEP database with the insertion of an interesting solution of known quality and performing a discovery session using DaMI on this modified database. The necessity of reproducibility as a terminating condition is falsified by the algorithm finding the optimal solution without yielding strong reproducibility. The sufficiency of reproducibility as a terminating condition is analyzed by manual examination of the CCEP database in which strong reproducibility was experienced. Ex post facto knowledge of the solution space is used to prove that DaMI had not found the optimal solutions though it gave strong reproducibility, causing us to reject the conjecture that strong reproducibility is a sufficient terminating condition.

TABLE OF CONTENTS

I. INTRODUCTION 1
II. BACKGROUND 3
A. INTRODUCTION TO GENETIC ALGORITHMS 3
B. INTRODUCTION TO DATA MINING INITIATIVE 5
1. Introduction 5
2. Design 7
a. Genetic Algorithm 7
b. Statistical Analysis Algorithm 8
3. Results 11
III. REPRODUCIBILITY CONJECTURE 13
IV. LITERATURE REVIEW 17
V. METHODOLOGY 23
A. REPRODUCIBILITY AND GENETIC ALGORITHM THEORY 23
B. EXPERIMENTAL DESIGN 26
1. Testing the Necessity of Reproducibility 27
2. Testing the Sufficiency of Reproducibility 28
B. ANALYSIS STRATEGY 29
VI. FINDINGS 31
A. EXPERIMENTAL RUNS 31
1. Testing the Necessity of Reproducibility 31
2. Testing the Sufficiency of Reproducibility 35
B. VERIFICATION RUNS 42
VII. CONCLUSIONS AND RECOMMENDATIONS 45
APPENDIX A. REPRODUCIBILITY AS DEFINED BY JACOBSON 47
APPENDIX B. TOP 20 SOLUTIONS FROM EXPERIMENTAL RUN 201 49
APPENDIX C. FINDINGS APPLICABLE ONLY TO DAMI RESEARCH 51
A. FITNESS LANDSCAPE CONSIDERATIONS 51
B. DESIGN CONSIDERATIONS 55
C. CONCLUSIONS AND RECOMMENDATIONS 58
LIST OF REFERENCES 61
BIBLIOGRAPHY 63
INITIAL DISTRIBUTION LIST 65

Soli Deo Gloria!

I. INTRODUCTION

In 1996, Bhargava and Jacobson developed a genetic algorithm application designed to mine the database holding the medical records of over 19,000 Persian Gulf War (PGW) veterans in search of a syndrome responsible for their medical complaints. As part of this study, Bhargava and Jacobson introduced the idea of reproducibility as a quality metric to the well-established field of genetic algorithm theory. (Bhargava and Jacobson, 1997)

This thesis examines their conjectures concerning reproducibility, both from a theoretical and a practical standpoint. Specifically, it examines the following questions:

- Is strong reproducibility either a necessary or a sufficient metric for measuring the effectiveness of a genetic algorithm discovery session?
- What testing method can be used to measure the effectiveness of a genetic algorithm search on an unknown solution space?

First, a review of accepted genetic algorithm theory to date is performed. Then, a new methodology for the testing of genetic algorithms on unknown solution spaces is developed. In this scheme, an interesting solution of known quality is inserted into the database. A discovery session is then performed on the modified database to determine with what effectiveness the algorithm locates the seeded solution.

Using this methodology, we have shown that strong reproducibility is neither a necessary nor a sufficient metric for determining the effectiveness of a genetic algorithm discovery session. Because of the probabalistic nature of genetic algorithm searches, there remains no objective certainty of the optimality of the results. However, the testing

method devised in this thesis does offer subjective criteria for measuring the algorithm's adeptness at locating solutions of interest to the developer.

The results of this study contribute both to the growing body of genetic algorithm theory and to the medical practitioners in search of a PGW syndrome. Specific recommendations applicable only to DaMI research are made in Appendix C.

This thesis is divided into seven chapters:

- Chapter I: Introduction.
- Chapter II: Background. Includes introduction to genetic algorithms and to DaMI.
- Chapter III: Reproducibility Conjecture. A summary of the conjecture made by Bhargava and Jacobson.
- Chapter IV: Literature Review.
- Chapter V: Methodology.
- Chapter VI: Findings.
- Chapter VII: Conclusions and Recommendations.

II. BACKGROUND

This chapter provides necessary background material for the rest of the thesis, including a general introduction to genetic algorithms and an introduction to DaMI (Data Mining Initiative).

A. INTRODUCTION TO GENETIC ALGORITHMS

A genetic algorithm is an automated, adaptive search technique modeled after the Darwinian principles of natural selection and 'survival of the fittest.' Genetic algorithms grew out of the study of adaptation in artificial and natural systems by Holland (1975) in the early 1970's. By using this method, a genetic algorithm can search the problem space in a general manner .

The genetic algorithm is designed to operate on a population of candidate solutions analogous to the chromosomes of a biological system. Each solution is modeled as a chromosome, and is evaluated by an objective function. It is the value returned by this objective function, called the fitness measure, which determines the probability of each chromosome reproducing offspring to pass on to the next generation. Each chromosome consists of a string of genes, whose values are called alleles. These genes are typically represented as a string of bits, though floating point numbers and integers may be used. (Holland, 1975)

A typical genetic algorithm is illustrated in Figure 2.1. The genetic algorithm begins by selecting an initial population, P(t), at time t=0. This initial population is

usually selected randomly, but may be selected deterministically if the situation warrants. Each of the members of the initial population is then evaluated by the objective function. While the terminating condition is not satisfied, the results of these evaluations are used as inputs in probabilistically determining which members reproduce for the next generation, according to the Darwinian principle of survival of the fittest. This reproduction is accomplished by a process called crossover, which may be further supplemented by mutation. These offspring are used as the inputs to the next generation, and the process repeats itself. A generational genetic algorithm stores the offspring in a temporary location until the end of the generation, when they replace the entire parent generation. In a steady-state genetic algorithm, the offspring immediately replace the parents in the current generation. (Corcoran and Wainwright, 1995)

```
procedure GA
begin
  t = 0;
  initialize P(t)
  evaluate structures in P(t);
  while termination condition not satisfied do
  begin
    t = t + 1;
    P(t) = select from P(t-1)
    alter structures in P(t);
  end
end.
```

Figure 2.1: Typical Genetic Algorithm
From Corcoran and Wainwright (1995)

The genetic algorithm uses three genetic operators to mimic genetic recombination in the production of offspring: reproduction, crossover, and mutation. Solutions from the current generation are preferentially selected according to the relative

value of the objective function, and then operated on by one of these genetic operators, as described below:

- **Reproduction:** Asexual reproduction of single parent rule to single offspring rule without modification
- **Crossover:** Sexual reproduction involving the exchange of chromosomes between two parents producing two different child rules
- **Mutation:** Asexual reproduction of single parent rule with random modifications resulting in a different child rule

(Holland, 1975)

While the basic principles and operations of a genetic algorithm are simple and straightforward, there are numerous variations and options which can be implemented to customize a genetic algorithm for a specific task. The modeling of hypotheses into chromosomes, the methods of selecting hypotheses for reproduction, crossover, and mutation, and the specific methods of introducing random mutations into the chromosomes are some of the ways that a genetic algorithm can be individualized. A particular genetic algorithm developed at the Naval Postgraduate School is the focus of this study.

B. INTRODUCTION TO DATA MINING INITIATIVE

1. Introduction

DaMI is a genetic algorithm developed by Jacobson to assist the Department of Defense (DoD) in the effort to define and localize a PGW syndrome. Since the gulf war, over 27,000 PGW veterans have presented health complaints which they attributed to their service in the region (CCEP, 1996a). Many of these veterans reported nonspecific

symptoms not directly attributable to a specific disease or syndrome (group of commonly occurring symptoms/conditions) (CCEP, 1996a). The large number of PGW veterans presenting health complaints sparked an effort by the DoD to attempt to discover if these non-specific symptoms could be correlated with any "clusters" of PGW veterans. The theory of this approach is that a PGW syndrome will be characterized by a "cluster" or group of individuals sharing some common trait(s) (demographics, location, action, exposures, etc.) who also share a similar group of symptoms. (CCEP, 1996b)

DaMI was developed as a search algorithm designed to locate these clusters within the Comprehensive Clinical Evaluation Program (CCEP) database. With few variations, it is a conventional generational genetic algorithm designed to mine the CCEP database to aid the search for a PGW syndrome (Jacobson, 1996). A syndrome is defined by a unique series of symptoms and/or ailments which are shared by a specific group of individuals (Jacobson, 1996).

A genetic algorithm was chosen because of the large search space resident in the CCEP database. DaMI examines the association between a large number of variables. In one of Jacobson's studies, there were 15 standard symptoms (LHS) and 21 possible diagnoses (RHS) (Jacobson, 1996). The attributes were represented as Boolean variables and were not limited in the number of possible combinations (i.e. any or all combinations of symptoms and diagnoses could be simultaneously present or "true"). This resulted in a search space of 2^{36} or 6.8×10^{10} possible hypotheses. To analyze this search space using simple "brute force" methods (i.e. testing every possible combination exhaustively) on a typical 486DX/66 Mhz personal computer would require ~315 years, based on an analysis

rate of 600,000 analyses per day (Jacobson, 1996). A genetic algorithm was chosen to analyze this search space because of its ability to effectively search a database in considerably less time than the brute force approach.

2. Design

a. Genetic Algorithm

The DaMI data structure was designed such that each chromosome consisted of a number of genes, where each gene was encoded as a Boolean attribute representing some piece of medical information for each service member. Over 19,000 DoD personnel were represented in the CCEP database, with each person's record encoded into this chromosomal format. The first runs performed by Jacobson (1996) involved chromosomes with 53 genes that were divided into left-hand-side (LHS) and right-hand-side (RHS) attributes, where the LHS consisted of 32 possible exposures/demographics and the RHS consisted of 21 possible diagnoses. An individual who reported 10 different exposures and was diagnosed with 3 different diagnoses might have a chromosome that looked like the following (where each 'Y' represents a positive report of a specific exposure/demographic or the presence of a specific diagnosis, and each 'N' represents a negative report of a specific exposure/demographic or the absence of a specific diagnosis. The first three genes, '1MC' may represent demographics such as '1' = 'army', 'M' = 'male', 'C' = 'Caucasian'):

1MCNNNYYNYNNYYYNYYYNNNNNNNYNNNNN | YNNNNNNNNYNYNNNNNNNNN

32 exposures | 21 diagnoses

DaMI is designed to search the CCEP database, which consists of 19,000 chromosomes of this type. Its basic architecture is modeled after Goldberg (1986), with the exception that DaMI stores rules as strings of Boolean attributes ('T' = consider the attribute; 'F' = don't consider the attribute). In this manner, DaMI can examine the associations between risk factors (exposures/demographics) and outcomes (symptoms/diagnoses) in aggregate before competing for selection and genetic recombination (Jacobson, 1996). Figure 2.2 illustrates the difference between the Goldberg model and that used in the DaMI architecture.

b. Statistical Analysis Algorithm

The DaMI statistical package in use is a fairly simple algorithm. Given a set of dependent attributes (RHS) and independent attributes (LHS), the statistical package is designed to return a value representing the "interest" of the given combination. "Interesting" is defined as "combinations of RHS attributes (dependent variables) which are highly dependent on combinations of LHS attributes (independent variables), or in other words, the candidate dependent variables are truly determined (not independent of) by the candidate independent variables." (Jacobson, 1996)

Conventional Genetic Algorithm Representation (Goldberg, 1989)

	Demographics		*Reported Exposures*				*Outcome Diagnoses*		
Rule	**Gender**	**Service**	**Uranium**	**Oil**	**Smoke**	**Combat**	**Anthrax**	**Fatigue**	**Depression**	**Memory Loss**
1	male	Navy	Yes	*	*		No	*	Yes	*

Rule 1 indicates a relationship between **Male Navy** personnel who reported exposure to **Uranium** but **not Anthrax** and an outcome diagnosis including **Depression**

***DaMI* Genetic Algorithm Representation**

	Demographics		*Reported Exposures*				*Outcome Diagnoses*		
Rule	**Gender**	**Service**	**Uranium**	**Oil**	**Smoke**	**Combat**	**Anthrax**	**Fatigue**	**Depression**	**Memory Loss**
2	TRUE	TRUE	TRUE	FALSE	FALSE		TRUE	FALSE	TRUE	FALSE

Rule 2 indicates a relationship between **gender, service, reported exposure to Uranium and/or Anthrax** and **whether or not** the patient was diagnosed with **Depression**

Figure 2.2: Conventional and DaMI Algorithm Representations From Jacobson (1996)

To determine the fitness measure of each attribute combination, DaMI uses what Jacobson described as a modified j-measure value (Jacobson, 1996). In classical epidemiology, a test is evaluated in terms of four variables which describe how successfully the test predicts the actual presence (or absence) of a particular disease. These four variables are computed using a two-by-two matrix, or contingency table, of test results and actual disease presence. These four variables are represented by {a,b,c,d} in Figure 2.3.

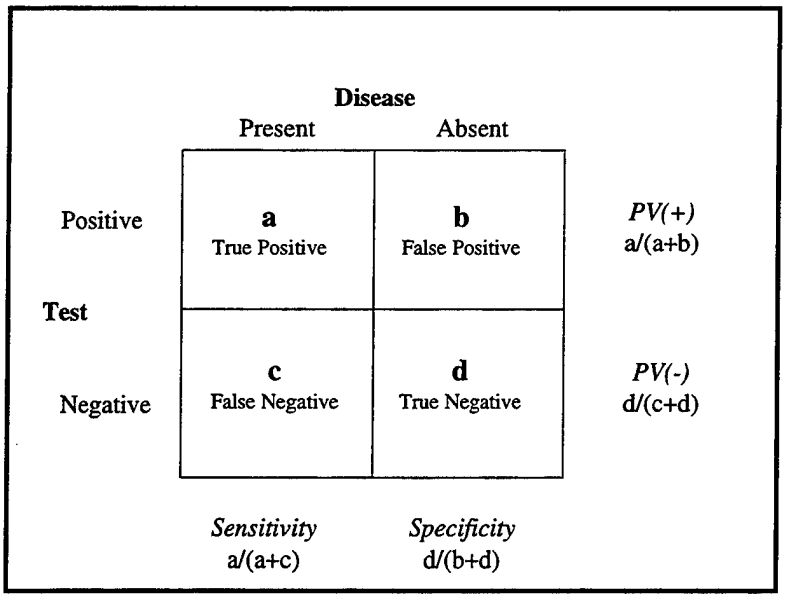

Figure 2.3: Classical Epidemiological Measures From Jacobson (1996)

From these four variables, four quality values are computed. These values are:

- **Positive Predictive Value:** Indicates the ability of a positive test to accurately identify the presence of a disease in a patient. It is indicated as PV(+) in Figure 2.3
- **Negative Predictive Value:** Indicates the ability of a negative test result to accurately determine the absence of a disease in a patient. It is indicated as PV(-) in Figure 2.3
- **Sensitivity:** The proportion of subjects with a disease who have a positive test for the disease.
- **Specificity:** The proportion of subjects without the disease who have a negative test.

(Jacobson, 1996)

The goal in DaMI research was to create a measure which was "suitably large when any of the four measures [PV(+), PV(-), sensitivity, and specificity] were large

and suitably low when none of the measures were relatively large—in effect an aggregate fitness measure." (Jacobson, 1996) The following measure was developed:

$$if\left(\frac{a \times d}{b \times c}\right) \geq 1, mod\_j = \frac{a \times d}{b \times c}$$
$$if\left(\frac{a \times d}{b \times c}\right) < 1, mod\_j = \frac{b \times c}{a \times d}$$

A natural log function was used to shape the fitness function for better genetic competition, such that the actual fitness measure becomes:

$$modified\ j\text{-}measure = 1 + ln[(a*b)/(c*d)]$$

A sample calculation of the modified j-measure is shown in Figure 2.4.

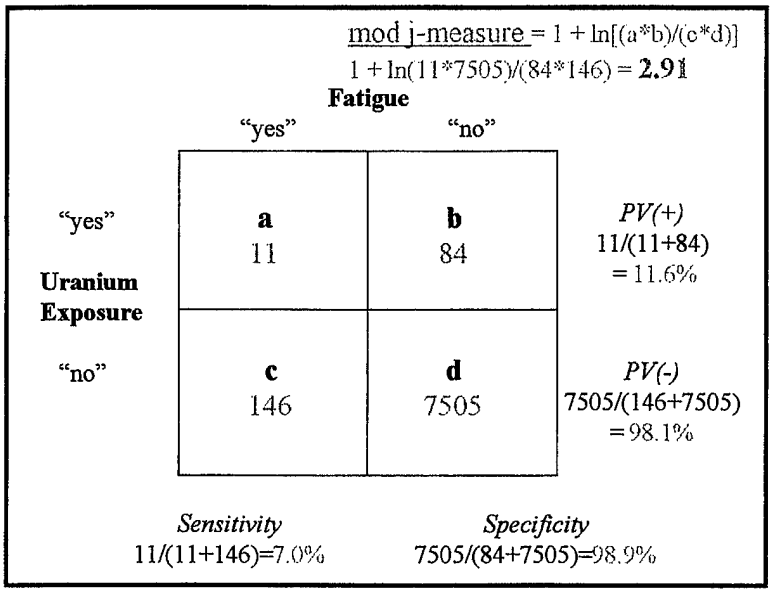

Figure 2.4: Modified J-measure Calculations From Jacobson (1996)

3. Results

Twenty-five discovery sessions (runs) were conducted by Jacobson (1996), of which six production runs were discussed. Earlier runs were used to test the performance

of DaMI during development and to refine the settings of tunable parameters for optimal discovery. Three of these six runs searched for associations between the gender, service, race, and reported exposures of PGW participants (LHS) and the diagnoses that were assigned by the CCEP medical examination process (RHS). They are referred to as *exposure-to-diagnosis* runs. (Jacobson, 1996) The other three production runs (*exposure-to-symptom* runs) were not addressed by this thesis.

In addition to these runs, a series of specialized analyses was performed relating to an oil fire in Khamisayah, Iraq. This study involved correlations between range (in miles) from Khamisayah and combinations of 15 standard symptoms and/or 60 diagnoses categories. (Bhargava and Jacobson, 1997) This study is referred to as the Khamisayah study, and was also used as a part of this thesis.

While the results produced by DaMI are impressive, the authors raised a paradoxical question: How can we be assured that the results produced by DaMI are the best possible results? (Bhargava and Jacobson, 1997) It is impossible to *prove* that DaMI's results are the best results without exhaustively testing every hypothesis, yet it was the impracticality of doing this that facilitated the use of a genetic algorithm as a search tool in the first place. Not only does this have an important bearing on the confidence placed in the algorithm's results, but an even more fundamental question must be answered: What terminating condition is necessary to declare that a discovery session is complete and no more runs need be performed? The next chapter will address the developers' proposed answer to that question.

III. REPRODUCIBILITY CONJECTURE

In the last chapter, we discussed the paradoxical situation inherent in any heuristic search – the uncertainty regarding optimality of the results. The developers of DaMI offered a proposed solution: reproducibility. Specifically, they looked for evidence in successive runs that a genetic algorithm started (in generation 0) from radically different points in the fitness landscape, yet converged (in the last generations) to the same solutions. This evidence, termed *reproducibility*, was offered as strongly suggesting that the "optimal values are indeed global." (Bhargava and Jacobson, 1997) To make these pair-comparisons, a graph was made to show that a very small percentage of low fitness measure (1.0-3.0) hypotheses was duplicated from run-to-run, while near-complete duplication of high fitness measure (>8.01) hypotheses was experienced (see Figure 2.5).

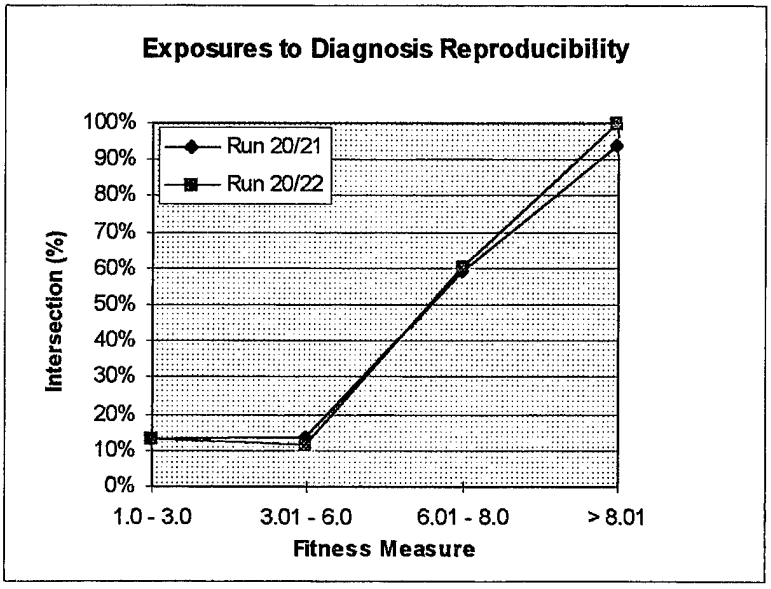

Figure 2.5: Exposure-to-diagnosis Reproducibility. From Jacobson (1996)

This strong reproducibility was considered enough evidence for the authors to claim that they "feel strongly that any rule of interest will be in DaMI's output hypothesis set." (Bhargava and Jacobson, 1996) A reproduction of the relevant section of Jacobson's thesis is included as Appendix A of this study. Figures 2.6 and 2.7 were included to describe what they considered the two possible outcomes of a genetic search.

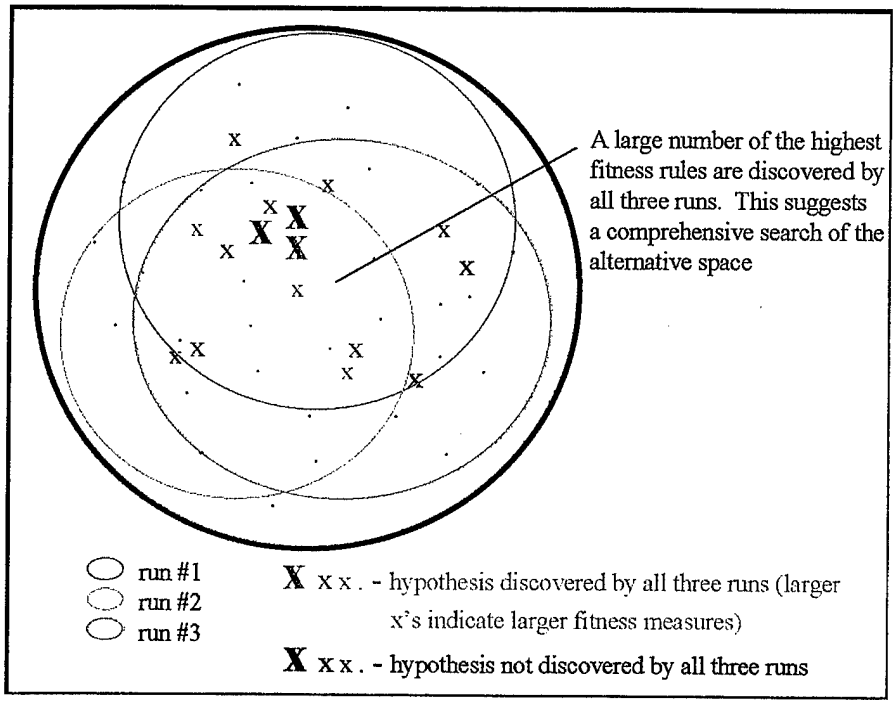

Figure 2.6: Strong Reproducibility in GA Search From Jacobson (1996)

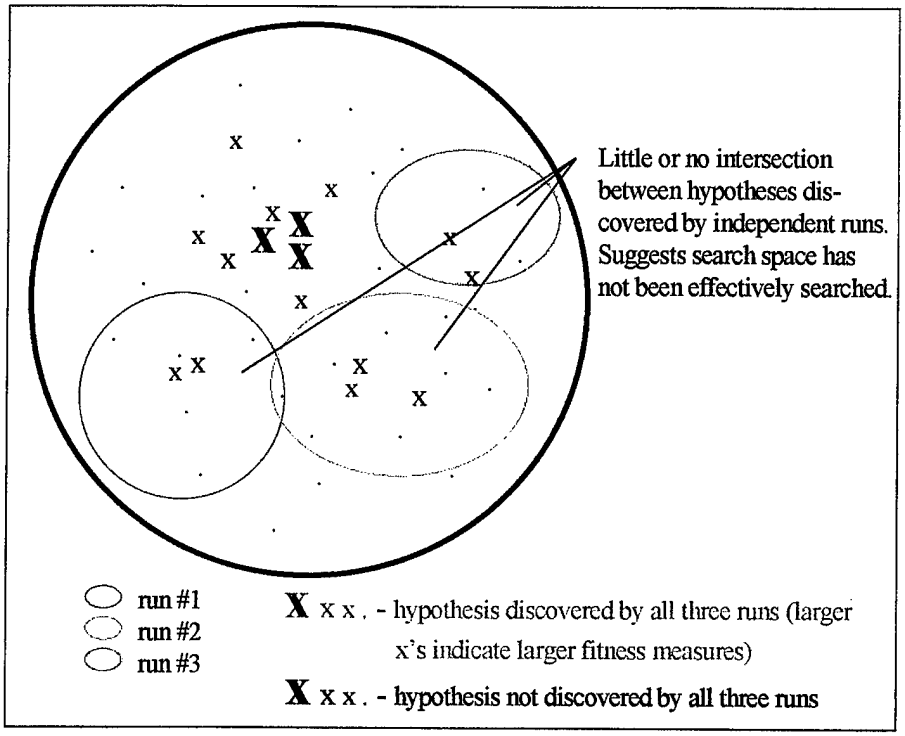

Figure 2.7: Weak Reproducibility of GA Search From Jacobson (1996)

It is believed that Jacobson was the first to specifically propose reproducibility as a metric to determine when a discovery session should be terminated. The next chapter in this thesis will discuss the analysis of this claim from the standpoint of conventional genetic algorithm theory. Then we will discuss the procedure devised to specifically test the claim using the scientific method.

IV. LITERATURE REVIEW

A literature review was conducted on over 1500 titles related to genetic algorithms. Titles were examined for their relation to one of two criteria:

- Generic convergence theories
- Generic testing methods

Twenty-seven articles were reviewed whose titles appeared to suggest discussion of generic convergence theories. No articles were found that discussed generic testing methods for genetic algorithms. A legitimate effort was made to cover the spectrum of literature available on these topics, and it is believed that what follows is a good representation. The possibility remains, however, that some articles were missed. We apologize in advance if this is the case.

Genetic algorithms were first introduced by Holland in the early 1970's (Holland, 1975). With genetic algorithm theory still in its infant stages, Holland demonstrated that the "algorithm's power is most evident when it is confronted with problems involving high dimensionality (hundreds to hundreds of thousands of attributes, as in genetics and economics) and multitudes of local optima." (Holland, 1975) Holland recognized that convergence of a genetic algorithm on a solution is not a useful guide to its robustness because of the non-zero probability that the observed average performance of suboptimal structures in the domain will exceed the observed average performance of the optimal structure(s), leading to the possibility of the deletion of data concerning the optimal structure (Holland, 1975).

Holland goes on to say, however, that each structure must therefore be repeatedly tested, and that this repeated testing (and the law of large numbers) "assures that suboptimal structures which have a finite probability of displacing an optimal structure will do so with a limiting frequency approaching that probability." (Holland, 1975) Here may be found the genesis of the idea that reproducibility leads to strong assurance that the genetic algorithm has searched the solution space effectively. This is insufficient in and of itself, however, because by Holland's own claims, there is still a non-zero probability that the algorithm will converge on this suboptimal structure.

In 1983, Ermakov and Zhiglyavskij offered a convergence theory for random search techniques using probability analysis. This work was further tailored to evolutionary algorithms by Qi and Palmieri in 1994. By 1996, Weishui and Chen proved a convergence theorem of genetic algorithms with all three basic operators in the general sense (solution space is m-dimensional Euclidean space). It was the first convergence theorem in the strict sense (Weishui and Chen, 1996).

In 1988, Koza discussed a phenomenon in genetic algorithms termed premature convergence, in which the fitness measure of a mediocre rule is disproportionately larger than the other individuals of its generation, leading to the mediocre rule dominating the population too quickly and providing the only material for future rules.

It will be helpful at this point to discuss the concept of fitness landscapes, first introduced by Wright in 1932. This concept involves the mapping of an individual's genomes to its fitness, and a visualization of that mapping. The idea of genetic algorithms searching on a fitness landscape was introduced as early as 1989 (Kauffman,

1989). To understand a fitness landscape, first imagine the space of all possible hypotheses that could be generated by a particular search algorithm applied to a particular problem. Each particular hypothesis has a fitness measure associated with it. Now, imagine that the space of all possible hypotheses is mapped onto the x-y plane, and that the fitness of each particular hypothesis is plotted on the z-axis. This will create a surface where the peaks are the locations of the hypotheses with good fitness measures, and the valleys are the locations of the hypotheses with high fitness measures. Discovering the global optimum then becomes equivalent to searching over this landscape for the highest peak. (Kinnear, 1994)

It follows from the above description that the neighbors of any particular hypothesis on the x-y plane are those hypotheses that can be generated by a single operation of the genetic operators. A key aspect of the success of evolutionary adaptive techniques is now raised—the correlation between the parents' and the offspring's fitness. If there is no variation between parents and offspring, then no improvement is made in the genetic search. On the other hand, if there is no correlation at all between the parents and offspring, then a genetic search becomes of no avail because the preferential selection of parents yields no probabilistic improvement in the selection of offspring, making the genetic algorithm no better than a random search technique. (Kinnear, 1994)

Kinnear uses the term *ruggedness* to describe this correlation between parents and offspring. A genetic algorithm will have difficulty locating the highest peak in a fitness landscape with great ruggedness. Contrarily, a genetic algorithm will likely have little

difficulty locating the global optimum in a fitness landscape consisting of one large hill, the top of which represents the best solution. (Kinnear, 1994)

In 1989, Goldberg described a minimal deceptive problem, in which "short, low-order building blocks lead to incorrect (suboptimal) longer, higher order building blocks," causing the genetic algorithm to diverge from the global optimum. Though he predates Kinnear's discussion of the ruggedness of fitness landscapes, deception and high ruggedness can be considered almost synonymous. In terms of fitness landscapes, a deceptive problem can be viewed as a flagpole in a valley surrounded by rolling hills, where the tip of the flagpole is higher than any hill and represents the global optimum. The "neighbors" of the flagpole (on the x-y plane) would all be located in the valley, and would be preferentially passed over by the algorithm in favor of points on the surrounding hills, though these points tend to lead the algorithm to converge only to local optima. Goldberg (1989) also showed that a standard genetic algorithm would consistently converge to an incorrect solution of the deceptive problem.

Other authors offered solutions to overcome these deceptive problems. In 1994, Renders and Bersini proposed to combine genetic algorithms with more traditional hill-climbing algorithms in a hybrid computing environment. Also in 1994, Dasgupta reported success using what he termed a structured genetic algorithm which introduced hierarchy into the genome representation in order to overcome deceptive problems. In 1995, Kingdon and Dekker recommended random changes in the representation of the search space to prevent convergence on suboptimal solutions.

In all of the articles reviewed that discussed the issue, it was generally accepted that one must either possess *a priori* knowledge of the fitness landscape or have no measure of certainty associated with the algorithm's results. It is believed that Jacobson was the first to suggest reproducibility as an indication of the algorithm's robustness. In addition, it is believed that any testing of the genetic algorithm's performance was done on a fitness landscape of known quality. No articles were found that offered a solution to test a genetic algorithm's success on an impractically complex and unknown fitness landscape.

V. METHODOLOGY

We now have a sufficient theoretical base to relate this to the current study. In this chapter, we will first discuss the implications of Jacobson's conjecture on reproducibility in light of conventional genetic algorithm theory. We then discuss the specific procedure designed by the authors to test the conjecture.

A. REPRODUCIBILITY AND GENETIC ALGORITHM THEORY

Recall the two possible outcomes of a genetic algorithm discovery session proposed by Jacobson (1996) (see Figures 2.6 and 2.7) – either strong reproducibility (indicating a successful search) or weak reproducibility (indicating an unsuccessful search). Theoretically, there are four possible outcomes according to these criteria:

- **Strong/positive:** The algorithm produces strong reproducibility and locates the optimal solution.
- **Strong/negative:** The algorithm produces strong reproducibility and does not locate the optimal solution.
- **Weak/positive:** The algorithm does not produce strong reproducibility and locates the optimal solution.
- **Weak/negative:** The algorithm does not produce strong reproducibility and does not locate the optimal solution.

The claim of Jacobson (1996) is that the second and fourth criteria can be eliminated as possibilities. In classical philosophical terminology, this amounts to the assertion that strong reproducibility is both *necessary* and *sufficient* to ensure that the solution space has been effectively searched. It is the testing of this claim to which this research is directed. Specifically, is strong reproducibility a valid terminating condition for a discovery session?

Consider the hypothetical situation in which the CCEP database fitness landscape consists only of a single flagpole in the middle of a large valley adjacent to a single rolling hill (see Figure 5.1). As above, the tip of the flagpole is the maximum value on the landscape and represents the global optimum. Upon running DaMI on this solution space, it seemed reasonable to conclude that DaMI could consistently converge on the suboptimal peak at the top of the rolling hill, yielding strong reproducibility as described in Bhargava and Jacobson (1997), while at the same time failing to locate the global optimum because of the landscape's deceptive properties.

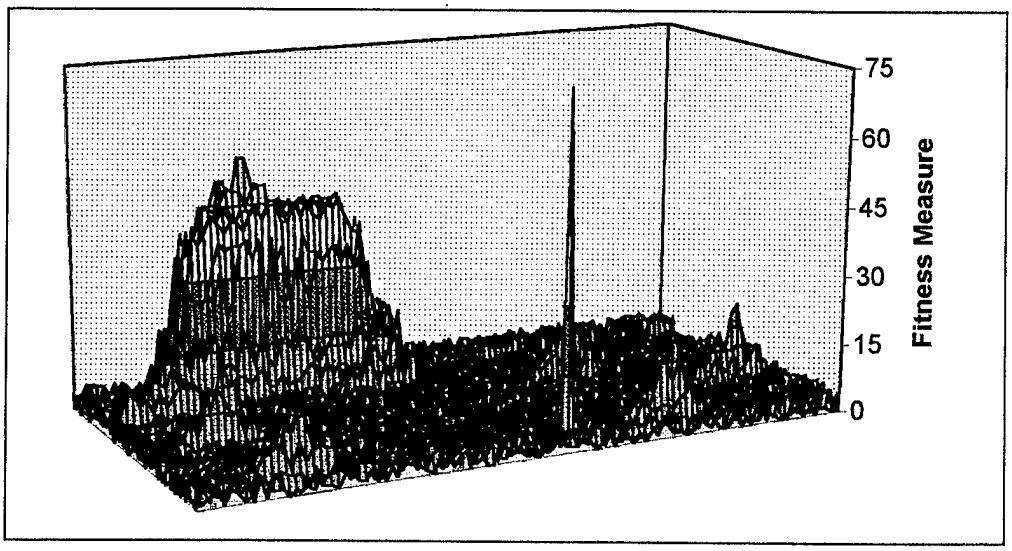

Figure 5.1: Hypothetical Deceptive Fitness Landscape

Because of the complexity of the CCEP database, however, it is unlikely that this simplistic version of a fitness landscape comes even close to representing that of the database. We now stretch the analogy further, and add more small hills to the picture, each with a much smaller base and a much lower peak than the first hill. It should be understood that we are starting to describe in graphical terms the non-zero probability of

suboptimal structures dominating the optimal structures described by Holland. In this particular situation, this probability is potentially large, and we may still expect DaMI to converge on the larger hill, while still leaving the flagpole undiscovered. Referring to the four possible outcomes of a discovery session, this outcome would be a *strong/negative*, eliminating the *sufficiency* of strong reproducibility to assure optimal results.

Visualize now a second fitness landscape consisting of many rolling hills of very nearly the same size and same shape, with one larger hill in the middle representing the global optimum (see Figure 5.2).

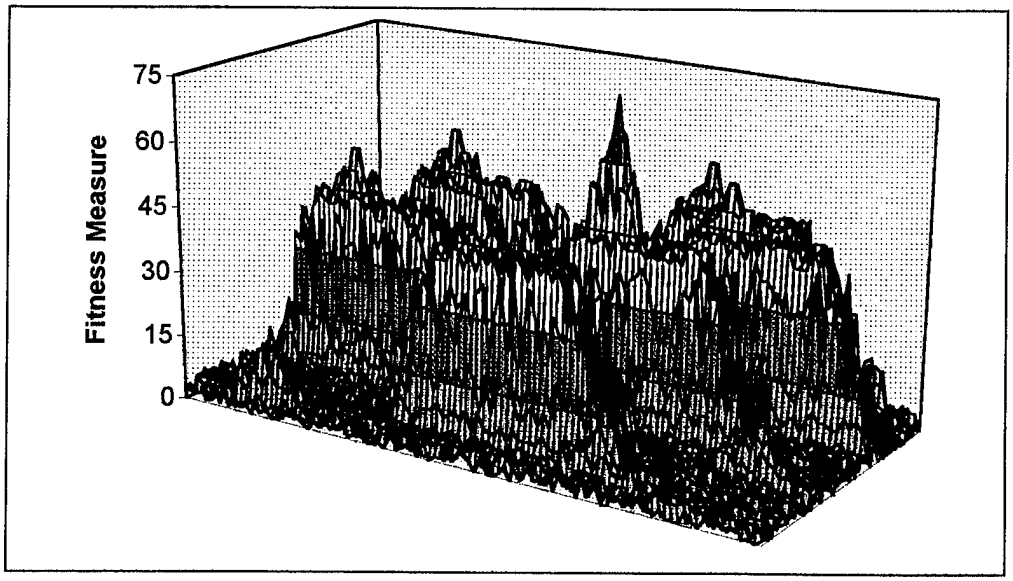

Figure 5.2. Hypothetical Fitness Landscape with Near-Uniform Solutions

Because of a genetic algorithm's inherent randomness and the large number of paths to climb the optimum hill, it seemed reasonable to conclude that a genetic algorithm could locate the optimal solution some of the time, yet fail to give indications of strong reproducibility as described in Bhargava and Jacobson (1997). This situation would be a

weak/positive, eliminating the *necessity* of strong reproducibility as a terminating condition.

B. EXPERIMENTAL DESIGN

The purpose of this study was to address these theoretical issues using the scientific method on the DaMI genetic algorithm. Specifically, could a testing scheme be devised which would falsify the above claims? There were essentially two independent hypotheses to test – the first that reproducibility was a *necessary* terminating condition, the second that reproducibility was a *sufficient* terminating condition. The inherent difficulty in testing these hypotheses lies in the paradox noted in Chapter II, i.e. that to *prove* how effectively a genetic algorithm had searched the solution space would require absolute knowledge of the fitness landscape. In the case of the CCEP database, an exhaustive analysis of the database was impractical because of its sheer size.

It was the design of this thesis, however, not to positively prove the above claims (which proof would be highly impractical), but to test the claims from a statistical perspective. The procedure devised was relatively simple. The solution space was deliberately altered in a very small way by the surgical insertion of "interesting" hypotheses, as defined by Jacobson (1996). These hypotheses were deliberately chosen to be more interesting than any hypotheses reported in Jacobson (1996), as measured both by the modified j-measure statistical analysis and by intuitive inspection of the contingency tables. After this seeding of interesting solutions, we ran DaMI on the modified database enough times to examine its performance. Specifically, we looked for

two things: DaMI's level of reproducibility and its adeptness at locating the seeded solutions.

1. Testing the Necessity of Reproducibility

The first hypothesis to be tested was the necessity of strong reproducibility as a terminating condition. In order to do this, it would be necessary to insert a solution that would be analogous to the large hill described above, yet higher than any solution found by DaMI in its prior runs. A program similar to the one in Figure 5.3 was used to seed solutions into the Khamisayah database, where *prim1* is the database table where the participants' medical records resided.

```
Select prim1
 scan
  if LHS attributes = desired conditions
   replace RHS attributes with desired condition
  endif
 endscan
```

Figure 5.3: Seed Code

A series of runs on the modified database would yield one of the four results described in Section V.A. A strong/positive or a weak/negative result would tend to confirm the conjecture made by Jacobson in his thesis, but would prove nothing. A strong/negative result would have no bearing on the necessity of strong reproducibility as a terminating condition, but would disprove the conjecture that strong reproducibility was a sufficient terminating condition. Only a weak/positive result would absolutely falsify

the conjecture that strong reproducibility was a necessary terminating condition, so a solution was seeded that was considered of a nature to best yield this result. The solution considered to best meet this criteria would consist of relatively few numbers of attributes and would affect a large number of records.

If the algorithm gave strong/positive, strong/negative, or weak/negative results, different solutions would be seeded in a further attempt to falsify this particular conjecture. If a large number of runs continually gave these results, a statistical analysis would be performed to determine the significance of the findings.

2. Testing the Sufficiency of Reproducibility

The more significant claim made by Jacobson was the sufficiency of strong reproducibility as a terminating condition. Not only was it the more significant claim, but it would also be more difficult to test. Two different solutions were seeded into the exposure-to-diagnosis database to test this conjecture. Again, four outcomes were possible. A strong/positive or a weak/negative result would tend to confirm the conjecture made by Jacobson (1996), but would prove nothing. A weak/positive result would have no bearing on the sufficiency of strong reproducibility as a terminating condition, but would disprove the conjecture that strong reproducibility was a necessary terminating condition. Only a strong/negative result would absolutely falsify the conjecture that strong reproducibility was a necessary terminating condition, so a solution was seeded that was considered of a nature to best yield this result. The solution considered to best

meet this criteria would consist of a complex combination of attributes and would affect a relatively small number of records.

If the algorithm gave strong/positive, weak/positive, or weak/negative results, different solutions would be seeded in a further attempt to falsify this particular conjecture. If a large number of runs continually gave these results, a statistical analysis would be performed to determine the significance of the findings.

B. ANALYSIS STRATEGY

All runs were analyzed for reproducibility in the same manner used by Jacobson. To generate the reproducibility graphs, the first run was compared individually to each subsequent run. The program used to perform these comparisons was identical to that used in Jacobson (1996). For these comparisons, strong reproducibility was defined as any series of runs in which all runs agreed on at least 90% of the solutions with fitness measure >8.01.

In addition, the output of each run was analyzed manually to determine if the seeded solution was located by the genetic algorithm. The output was inspected not only for solutions that exactly matched the seeded solution, but for patterns that would identify the seeded solution.

VI. FINDINGS

A. EXPERIMENTAL RUNS

1. Testing the Necessity of Reproducibility

In order to test the necessity of reproducibility, the program in Figure 6.1 was used to seed a relatively simple solution into the Khamisayah database. A total of 1074 (of 7746) records were affected.

```
Select prim1
 scan
  if fatig = "Y" and diarr = "Y"
   replace kin10 with "Y"
  endif
 endscan
```

Figure 6.1: Seed Code

The pre-seeded and post-seeded contingency tables are shown below:

	'T'	'F'
'T'	9	1065
'F'	43	6629

modified j-measure
1.26

before seeding

	'T'	'F'
'T'	1074	0
'F'	43	6629

modified j-measure
undefined

after seeding

It is of no consequence that the modified j-measure for the seeded solution was undefined. It is not the object of a genetic algorithm necessarily to find the *one* best solution, but to find a *range* of the best solutions. Because of the large number of records altered by the seed, this would have a large collateral effect on other potential solutions, as was intended.

Nine experimental runs were performed on this modified database. The results are shown in Figure 6.2.

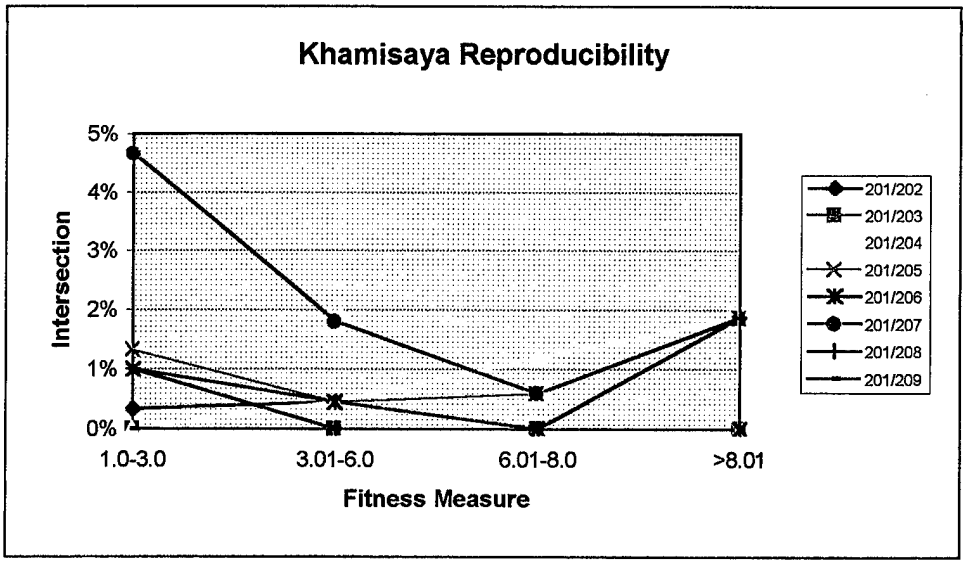

Figure 6.2: Testing the Necessity of Reproducibility

To interpret the chart, consider the point (1.0-3.0,4.7%) in series 201/207. An intersection value of 4.7% indicates that 4.7% of the hypotheses of fitness measure 1.0-3.0 in the first run (run 201, in this instance) are also located in the second run (run 207, in this instance). Note that by the standards described in Jacobson (1996), this represents weak reproducibility. Did DaMI locate the seeded solution? Upon answering this question, we will begin to see one weakness of reproducibility in describing the success of

question, we will begin to see one weakness of reproducibility in describing the success of a genetic algorithm. Remember that the seeded solution itself had an undefined fitness measure, so to find that one exact solution would not be feasible. However, to figure out that this was the best solution only took a cursory look at DaMI's results.

Consider only the first run, run 201. This run had 108 solutions with fitness measure >8.01. Of these solutions, all 108 included the RHS attribute, *kin10*, and at least one of the LHS attributes, *fatig* or *diarr*. In addition, 66 of the 108 solutions included all three of the seeded attributes. DaMI did, in fact, locate the seeded solution in this run (see Appendix B for the top 20 rules found by run 201). Six of the other tables yielded similar results. The reason the results in Figure 6.2 do not show strong reproducibility lies in the method used to calculate the intersection between two tables. Consider the four solutions below:

- LHS: *fatig, diarr, headache*; RHS: *kin10, kin30*
- LHS: *fatig, diarr*; RHS: *kin10, kin30*
- LHS: *fatig, diarr, headache*; RHS: *kin10*
- LHS: *fatig, diarr, backpain*; RHS: *kin10, kin30*

Visual observation of the four solutions quickly leads to the conclusion that there is a high correlation between the three affected attributes. However, when calculating intersection between two solution sets, the computer program used by Jacobson (1996) (and by this study) only counts an intersection if both the LHS text and the RHS text are exact duplicates. If the first two solutions above were found by one run and the second two by another, the computer program would yield 0% reproducibility, as there are no exact duplicates. Seven of the nine experimental runs yielded just this type of result. So,

in some sense, though these seven runs produced no strong reproducibility as defined by Jacobson, they had each converged on the same solution

Had this been the case for all the runs, perhaps the only thing needed would be an alteration of the definition of reproducibility. However, two of the runs (206 and 209) did not converge on the correct solution. An examination of the results showed that run 206 found strong associations between *diarr* and *kin10*, but did not locate the even stronger correlation when the symptom *fatig* was added. Run 209 did not yield an association between any of the three seeded attributes. This test proved that strong reproducibility was not a necessary terminating condition for a genetic algorithm, nor should the operator wait for strong reproducibility to be certain that the algorithm had effectively searched the solution space. With this in mind, the two figures (see Figures 2.6 and 2.7) used by Jacobson to represent the two possible outcomes of a genetic algorithm search may now be supplemented with Figure 6.3.

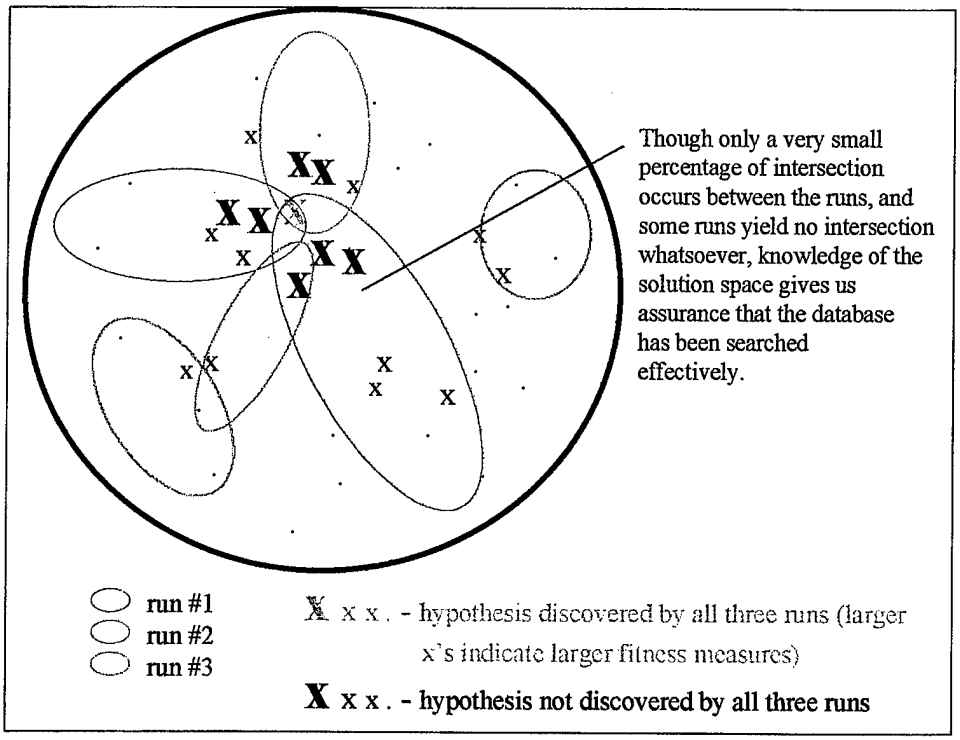

Figure 6.3: Alternate Explanation of Weak Reproducibility After Jacobson (1996)

2. Testing the Sufficiency of Reproducibility

In order to test the sufficiency of reproducibility, a program similar to that in Figure 6.1 was used to seed a relatively complex solution into the exposure-to-diagnosis database.

For the first seed, the LHS conditions (exposures) were *contm_watr*, *contm_food*, and *pq_after*, and the RHS conditions (diagnoses) were *a307_81* and *a692_9*. There were a total of 115 records in prim1 in which the three LHS attributes were "Y". The pre-seeded and post-seeded contingency tables are shown below:

	'T'	'F'
'T'	1	114
'F'	37	7746

modified j-measure
1.59

before seeding

	'T'	'F'
'T'	111	4
'F'	37	7594

modified j-measure
9.65

after seeding

A second solution was added in which the LHS conditions (exposures) were *microwaves*, *malaria*, and *botulism*, and the RHS conditions (diagnoses) were *a309_81* and *a780_71*. There were a total of 297 records in prim1 in which the three LHS attributes were "Y". The pre-seeded and post-seeded contingency tables are shown below:

	'T'	'F'
'T'	1	296
'F'	27	7422

modified j-measure
1.07

before seeding

	'T'	'F'
'T'	283	14
'F'	27	7422

modified j-measure
9.62

after seeding

The highest fitness measure located by the three production runs was 9.26. The seeded solutions' fitness measures of 9.65 and 9.62 were sufficiently higher than this figure to adequately test the hypothesis.

Nine experimental runs were performed on this modified database. The results are shown in Figure 6.4. Only weak reproducibility by the standards outlined in Jacobson (1996) was experienced, and the seeded solution was not located. This corresponded to a weak/negative result, tending to confirm the conjecture made by Jacobson. However, it was noted that in over 40 test runs leading up to the current experiment, the strong reproducibility described in Jacobson (1996) was *never* encountered. At this point, two questions arose:

- Since only a small portion of the solution space was affected by the insertion of the seeded solution, why didn't the test runs give a similar level of strong reproducibility as the production runs in Jacobson (1996)?
- Was there some other way of testing the hypothesis that did not require reproducing the strong reproducibility in the experimental runs as originally designed?

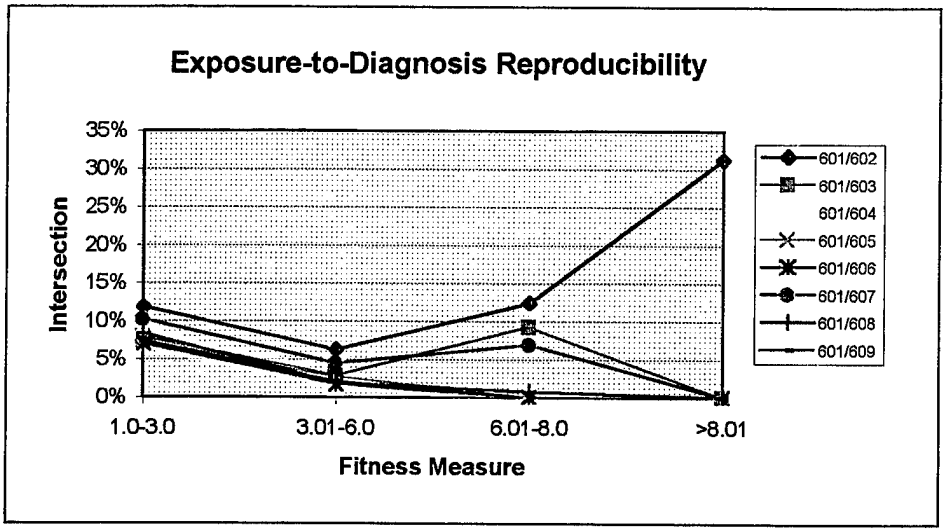

Figure 6.4: Testing the Sufficiency of Reproducibility

To answer these questions, a more detailed analysis of DaMI's results was necessary. Consider the level of reproducibility represented in Figure 6.4. Though this is the same method used by Jacobson (1996) to determine reproducibility, again it does not tell the whole story. Specifically, the graph tells nothing of the nature of the solutions found by DaMI. A manual analysis of the highest fitness measure (>8.01) solutions, though, yielded what we considered interesting conclusions. In the experimental runs, a total of 45 solutions with fitness measure >8.01 were discovered. Of these 45 solutions, 30 were found by the original production runs reported in Jacobson (1996), leaving 15 new solutions discovered in this experiment.

Of these 15 new solutions, 12 were present in the database (but not located by DaMI) during Jacobson's (1996) initial production runs. This was verified in two ways:

- The hypotheses did not involve any of the genes affected by the seed.
- The original prim1 table was manually queried for the new hypotheses and their actual presence in the database was verified.

Furthermore, a manual analysis of these solutions showed very high intersection of attributes between the 45 high fitness measure solutions. Of the 21 possible RHS attributes, only 6 are represented by these solutions. Furthermore, all 45 of the solutions have a RHS contribution that consists of some combination of two of these six attributes. Only five such combinations are represented, four of which contain the attribute *a296_20*. The fifth combination appeared five times and consists of the attribute *a780_7*, which, when paired with *a296_20*, represents 25 of the other 40 solutions. All 45 of the solutions have only one instance where both the LHS and RHS attributes were true. In other words, both the experimental runs and the production runs really are converging on the same solutions, though less consistently in the experimental runs.

To this point we have only discussed those solutions with fitness measures >8.01 for the following reasons:

- There are a relatively small number of hypotheses located with fitness measures >8.01, making manual verification feasible, and
- This is the only area with very strong reproducibility; therefore discussion of only these hypotheses is sufficient to address the sufficiency of strong reproducibility

Note that much of our discussion hinges on the lack of specificity in the term reproducibility. Jacobson (1996) defined reproducibility in terms of percent intersection of exact rules between solution sets. For a rule to be counted, both the LHS attributes and RHS attributes had to be exactly the same. No consideration was given to other possible similarities between the solution sets. It will be noted that 12 new hypotheses were located by the experimental runs that *could* have been located by the production runs, but were not. In this case, the three production runs located only 71% (30 of 42) of the

known good solutions in the solution space. It is conceivable that future runs could yield more new solutions, which would only cause this number to get worse. With this in mind, our findings to this point may be further supplemented with Figure 6.5.

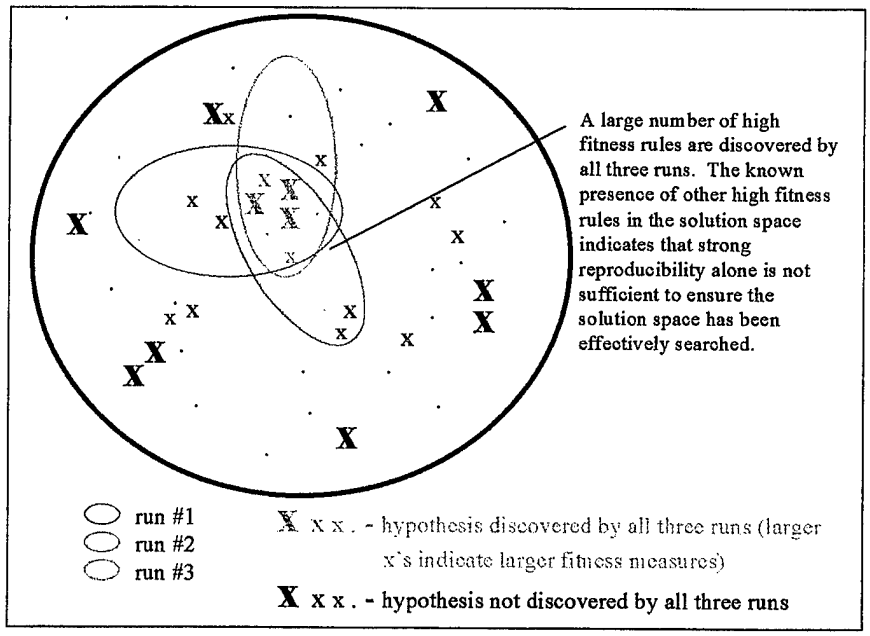

Figure 6.5: Alternate Explanation of Strong Reproducibility After Jacobson (1996)

Upon examining Figure 6.5, note that there is, again, no way of proving relatively how many of the large, black X's should be indicated on the diagram without performing an exhaustive search of the solution space. It is only the added information provided by the experimental runs that allows us to go back and redraw this figure as shown here, noting that at the time the strong reproducibility described in Jacobson (1996) was produced, there were, in fact, as yet unlocated high fitness measure hypotheses resident in the database.

Can this problem be solved by redefining the term reproducibility in a less strict manner? The term may be redefined, but this does not solve the problem. Using a loose definition of the term reproducibility, many of the production runs and experimental runs converged on the "same" solutions. With the high degree of similarities between these solutions, all these solutions may be considered to be on or near one "hill" in the fitness landscape. However, the experimental runs failed to locate the two seeded solutions, which had higher fitness measures than any of the solutions found by DaMI. With this loose definition of reproducibility, DaMI only located one of three solutions, still yielding negative results.

To this point, we have not discussed the three solutions located by the experimental runs that were not present in the original database. All three of these solutions were located in the same run. Was DaMI converging on the correct solution? To answer this question, examine the three solutions:

- LHS: *service, smoke_now, sex*; RHS: *a296_20, a692_9*
- LHS: *service, microwaves, sex*; RHS: *a296_20, a692_9*
- LHS: *service, carc_paint, sex*; RHS: *a296_20, a692_9*

All three contingency tables are identical and look like this:

	'T'	'F'
'T'	1	1
'F'	5	7739

modified j-measure
8.34

Furthermore, all three contingency tables appeared the same before the seed:

	'T'	'F'
'T'	0	2
'F'	3	7741

modified j-measure
0

Note that it is again the small number of RHS attributes when combined with *a296_20* that is the cause of the high fitness measure, not that DaMI is beginning to sniff the seeded solution.

It has now been demonstrated that just as strong reproducibility is not a necessary terminating condition for a data session, neither is it a sufficient terminating condition. It is a simplistic measure that does not consider the nature of the fitness landscape and the probabilistic nature of a genetic algorithm. In addition, the original results reported by Jacobson were misleading, as will now be discussed.

B. VERIFICATION RUNS

Note that we still have not determined why the algorithm gave such strong reproducibility during Jacobson's (1996) production runs, but not during the experimental runs performed in this study. At this point, it was noted that only three exposure-to-diagnosis runs were conducted by Jacobson in the original study--hardly enough to give statistical significance. Rather than speculate why the original production runs gave such strong reproducibility and the experimental runs in this thesis did not, it was decided that increasing the sample size of the original three production runs would be beneficial. Five more runs were performed identical to those performed in Jacobson (1996).

The results of the five verification runs are shown in Figure 6.6. The comparison is to the same table (run 20) as that made in Jacobson (1996). Four of the five verification runs showed weak reproducibility. Only the third run, run 503, showed reproducibility of a comparable level to that documented in Jacobson (1996).

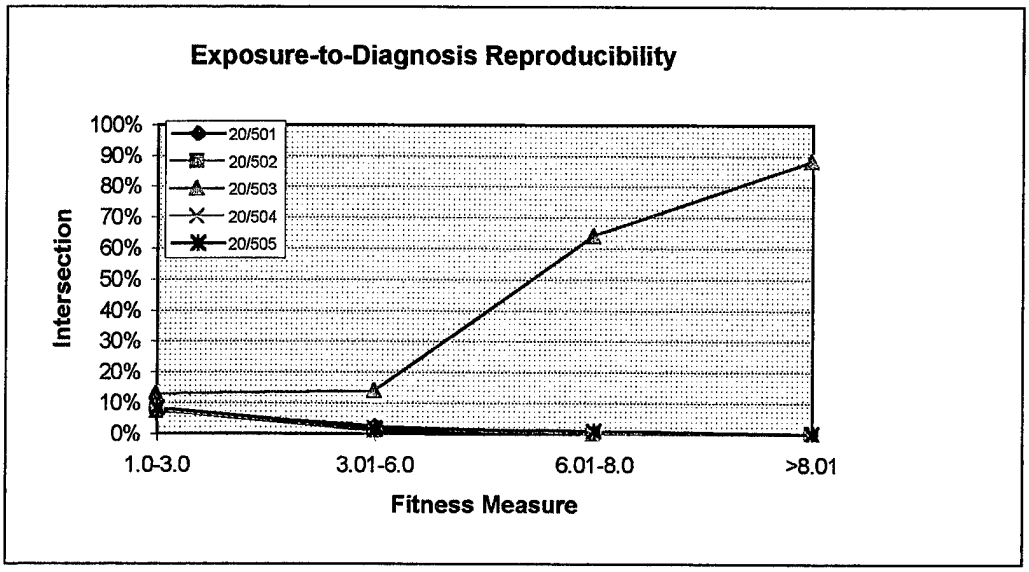

Figure 6.6: Verification Runs Reproducibility

For a scientific study's results to be conclusive, they must be reproducible by a third party. The results of the verification runs indicate that the findings reported in Jacobson (1996) are not reproducible in the strictly scientific usage of the term. A review of all the runs performed to this point will show that none actually gave consistently strong reproducibility. This is not only the case for the experimental runs, but for the original production runs as well.

The experimental runs demonstrated that reproducibility is neither a necessary nor a sufficient terminating condition of a genetic algorithm data session. The verification runs demonstrated that the *feasibility* of attaining consistent reproducibility is questionable. This is intuitively supported by an understanding of the probabilistic nature of genetic algorithms. Furthermore, it is supported by an absence of reference to its occurrence in the large body of genetic algorithm literature reviewed (see Chapter IV). In

any case, it has been demonstrated that DaMI does not give the consistent reproducibility necessary to terminate a data session according to the standards set forth by Bhargava and Jacobson (1997).

VII. CONCLUSIONS AND RECOMMENDATIONS

This study has examined the conjecture made by Bhargava and Jacobson (1997) that strong reproducibility is a necessary and sufficient terminating condition to ensure a genetic algorithm produces the best possible results. It is believed that this is the first study to examine this conjecture directly. First, the conjecture was examined from a theoretical standpoint. It was demonstrated that others had addressed the issue indirectly, particularly in their discussions of deceptive problems.

Furthermore, we have tested the conjecture in a scientific manner and have demonstrated practically that strong reproducibility is neither a necessary nor a sufficient terminating condition for a genetic algorithm data session. The necessity of reproducibility as a terminating condition was falsified by running the algorithm on a database modified by the surgical insertion of a solution with an infinite fitness measure. Because the algorithm located the seeded solution without producing strong reproducibility, the necessity of strong reproducibility as a terminating condition was rejected.

The sufficiency of strong reproducibility as a terminating condition for a genetic algorithm was tested by running the algorithm on a database modified by the insertion of a complex solution. Though we were not able to falsify the conjecture directly by producing strong/negative results, we were able to demonstrate its weakness in a secondary manner. This was accomplished by ex post facto analysis of Jacobson's (1996) results in light of the new knowledge of the solution space gained by this study.

To our knowledge, it remains to be shown that a probabilistic search technique such as a genetic algorithm can be expected to consistently produce the same results when run on a highly complex fitness landscape. Because a genetic algorithm is a probabilistic vice deterministic search technique, there remains no level of certainty in the outcome of a search on a complex database of unknown fitness landscape.

In addition, we have proposed a new method for testing the effectiveness of a particular genetic algorithm on a complex, unknown fitness landscape. This method involves the alteration of only a small portion of the fitness landscape to insert a solution of sufficient quality that the developer would be satisfied with the algorithm locating. The ability of the algorithm to locate this seeded solution would give a subjective indication of how well the algorithm performed on the unmodified database. While this method can yield no certain information about the landscape's quality, it can give an indication of the algorithm's ability to locate what the developer would otherwise consider solutions of interest.

This thesis also has practical implications for the search for a Persian Gulf War Syndrome. DaMI was adept at locating some high fitness measure scores in the unmodified fitness landscape during the original production runs. It was also shown in this study that DaMI could locate, with some regularity, simple solutions inserted by the authors, though not with the consistency discussed by Jacobson's (1996). However, DaMI proved inept at locating complex solutions of interest inserted by the author. Specific explanations of this phenomenon and recommendations for further DaMI research are contained in Appendix C.

APPENDIX A. REPRODUCIBILITY AS DEFINED BY JACOBSON

"Reproducibility gives a strong indication that the alternative space has been searched effectively. Ideally, we would like multiple independent runs of the genetic algorithm in order to test only a few of the same rules of low fitness but converge on the same rules of high fitness. A low intersection of low fitness rules between runs indicates that each approached convergence from different areas of the search space (i.e. they did not all follow the same path). A high intersection of high fitness rules suggests that, despite entering the search space from different directions, each independent run has arrived at the same answer. This reproducibility strongly suggests that the entire search space has been effectively, but not physically, examined.

DaMI achieves high reproducibility in spite of the rapid search time and tremendous space. In the exposure-to-diagnosis study, all three runs agree on the same 16 highest fitness hypotheses. Lower fitness hypotheses show steadily decreasing levels of intersection, as is theoretically predicted. This is particularly exciting, because each production run has achieved consensus by testing only 7,100 - 7,400 of the 1,041,000 possible attribute combinations. The probability of three independent runs randomly agreeing on the same sixteen hypotheses (especially since each run is testing only 0.7% of all possible attribute combinations) is infinitesimally small. The natural question is, "Did the three runs, by some streak of luck, enter the search space from the same starting point?" This is not the case, because the three runs only tested 14% of the same lower fitness rules, proving that they have entered the space from different points but converged

on the same answer. Note in Figure #18 that the percentage of rule intersection (Runs 20, 21, and 22 are the three runs conducted in the exposure-to-diagnosis study) between runs approaches 100% for rules with a fitness measure higher than 8.0. This intersection decreases steadily as the fitness measure decreases (going left on the graph)." (Jacobson, 1996)

APPENDIX B. TOP 20 SOLUTIONS FROM EXPERIMENTAL RUN 201

Fitness	LHS Rule	RHS Rule
12.20	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.FEVER="N"	KIN10="Y".and.KIN30="N"
12.20	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.DIABETES="N"	KIN10="Y".and.KIN30="N"
12.20	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.LIPID_ME="N"	KIN10="Y".and.KIN30="N"
12.20	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N"	KIN10="Y".and.KIN30="N"
11.51	FATIG="Y".and.DIARR="Y".and.BRONCHO="N"	KIN10="Y".and.KIN30="N"
11.51	FATIG="Y".and.DIARR="Y".and.BRONCHO="N".and.DIABETES="N"	KIN10="Y".and.KIN30="N"
11.11	FATIG="Y".and.DIARR="Y".and.BRONCHO="N".and.LUNG_AGT="N"	KIN10="Y".and.KIN30="N"
11.10	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.RHEUM_AR="N"	KIN10="Y".and.KIN30="N"
11.10	FATIG="Y".and.DIARR="Y".and.NAUSEA="N"	KIN10="Y".and.KIN30="N"
10.82	FATIG="Y".and.DIARR="Y".and.BRONCHO="N".and.RHEUM_AR="N"	KIN10="Y".and.KIN30="N"
10.81	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.NAUSEA="N"	KIN10="Y".and.KIN30="N"
10.59	FATIG="Y".and.DIARR="Y".and.BRONCHO="N".and.NAUSEA="N"	KIN10="Y".and.KIN30="N"
10.58	FATIG="Y".and.DIARR="Y".and.COUGH="N"	KIN10="Y".and.KIN30="N"
10.58	FATIG="Y".and.DIARR="Y".and.WEIGHT_L="N"	KIN10="Y".and.KIN30="N"
10.58	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.SARCOID="N"	KIN10="Y".and.KIN30="N"
10.48	DIARR="Y".and.LUNG_AGT="N"	KIN10="Y".and.KIN30="N"
10.40	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.DYSPHAG="N"	KIN10="Y".and.KIN30="N"
10.40	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.LYMPHAD="N"	KIN10="Y".and.KIN30="N"
10.40	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.WEIGHT_L="N"	KIN10="Y".and.KIN30="N"
10.40	FATIG="Y".and.DIARR="Y".and.LUNG_AGT="N".and.COUGH="N"	KIN10="Y".and.KIN30="N"

APPENDIX C. FINDINGS APPLICABLE ONLY TO DAMI RESEARCH

In the main body of the thesis, we discussed the testing of reproducibility in general terms which would extrapolate not only to the other production runs reported in Jacobson (1996), but to the testing of any genetic algorithm. We will now turn to the specifics of the DaMI algorithm which explain how the algorithm gives intuitively strong indications of effective search while at the same time producing disappointing results.

A. FITNESS LANDSCAPE CONSIDERATIONS

DaMI is a search algorithm designed to locate a syndrome (or syndromes), if one exists, within the CCEP database. If an undefined syndrome exists, it is likely to be a complex combination of common exposures, symptoms, and diagnoses, else it would have been easily located by medical professionals. Is a genetic algorithm well-suited to locate such a complex solution? Accepted genetic algorithm theory maintains that to answer that question, some estimate of the nature of the solution space is necessary.

For a genetic algorithm to be successful, a type of "learning" must take place from generation to generation. As stated in Chapter III, this requires a relatively high degree of correlation between neighbors on the fitness landscape to facilitate this learning process. So far, we have discussed fitness landscapes only in terms of three-dimensional space. To visualize the CCEP database solution space accurately would require the ability to comprehend many more dimensions, which is impossible for the human brain. Having said that, however, we will attempt to address the issue anyway.

Let us consider a hypothetical syndrome similar to the ones seeded into the CCEP database during the testing of the sufficiency of reproducibility as a terminating condition (see Chapter IV), where a combination of three exposures resulted in two medical diagnoses. Presumably, it is some interaction between the three of these exposures that causes the medical conditions in the patient. Examining the contingency table of this syndrome would show a high correlation between the LHS and RHS attributes which would also be born out in the fitness measure. If a patient were exposed to just one or any combination of two of the LHS attributes, no symptoms (and therefore no diagnoses) are expected. As these cases would clearly be neighbors of the actual syndrome, what correlation would we expect to see? This is a difficult question to answer in multi-dimensional space, but let us make an attempt.

Consider the hypothetical situation where 2000 people were exposed to the first attribute, 2000 to the second, and 2000 to the third. The intersection of any two of the three groups is 1000 people, and the intersection of all three is 150 people. Of these 150 people, 99% were diagnosed with the two RHS attributes. All of the others were diagnosed with these RHS attributes at the background rate of 10% for the rest of the population. This would result in the population of 7746 (identical to the number of records in prim1) represented in Table C.1, where E_1 is exposure 1, and so on.

E_1	E_2	E_3	# exposed	# diagnosed with D_1 and
Y	Y	Y	150	149
Y	Y	N	850	85
Y	N	Y	850	85
N	Y	Y	850	85
Y	N	N	150	15
N	Y	N	150	15
N	N	Y	150	15
N	N	N	4596	460

Table C.1. Hypothetical Population in CCEP Database

The contingency table and the fitness measure for the syndrome of interest ($E_1/E_2/E_3$ = 'Y'; D_1/D_2 = 'Y') would be:

	'T'	'F'
'T'	149	1
'F'	760	6836

modified j-measure
8.20

Now let us consider some of its neighbors. For example, (E_1/E_2 = 'Y'; D_1/D_2 = 'Y'):

	'T'	'F'
'T'	234	766
'F'	675	6071

modified j-measure
2.01

and (E_1 = 'Y'; D_1/D_2 = 'Y'):

	'T'	'F'
'T'	334	1666
'F'	575	5171

modified j-measure
1.59

These are only two of thousands of neighbors that the solution of interest could have. For instance, ($E_1/E_5/E_8$ = 'Y'; D_5/D_{11} = 'N') and ($E_1/E_5/E_8$ = 'Y'; D_5 = 'Y') would also be neighbors as they have the exposure E_1 in common. Calculation of all of these would be impractical. The two that were calculated should be two of the solution's nearest

neighbors, however. Even in this simple hypothetical example, it is easy to see that there is potentially little correlation between neighbors in the solution space, especially when it is remembered that the lowest possible fitness measure is 1.0. The fitness measures of 1.59 and 2.01 are in a range where thousands of other solutions reside, and are not large enough to alert the algorithm that it is approaching an interesting solution. A fitness landscape ideally suited for genetic algorithm search would be less rugged, with the optimum's nearest neighbors' fitness measures being only slightly less than its own, and a low slope drop-off as the solutions diverge.

There are numerous other variables to consider, however. It is likely that the modified j-measure proposed by the developers is not the best measure of a solution's fitness, and that some other measure would tend to smooth out the fitness landscape. The authors of this thesis did examine a number of other potential fitness measures, such as the chi-square and simple odds ratio, but all suffered from some weakness that made them undesirable. In any case, a visual examination of the three contingency tables above would show that there is still the very large potential for low correlation between neighbors on the fitness landscape no matter what fitness measure is used.

It is also possible that the hypothetical situation considered above is not representative of the CCEP database. If there were a smaller intersection between any two of the three exposures, or a larger intersection between the three, this would give a higher correlation between the three solutions.

B. DESIGN CONSIDERATIONS

Note in all the figures in this thesis labeled "Exposure-to-Diagnosis Reproducibility" that the ordinal x-axis range extends to >8.01. In the production, experimental, and verifications runs, many (but not all) of the runs produced results with values in this highest category (>8.01). However, upon examining the results in Appendix C of Jacobson (1996), entitled "Top 100 Hypotheses Discovered by Exposures-to-Diagnosis... Studies," we find that the #1 hypothesis reported has a fitness measure of only 3.24. The reason for this is that the raw results produced by DaMI are manually filtered by the author prior to inclusion in Appendix C according to the following criteria:

- Hypotheses applying to fewer than five individuals in the sample set were removed to prevent undue influence by single outliers. By definition, a syndrome is a medical condition shared by a number of individuals.
- Hypotheses were derived from a randomly selected 45% sample (without replacement) subset of the entire CCEP database. These hypotheses were tested against a separate 45% (independent) partition of the CCEP database. Hypotheses whose fitness measure in the second (verification) sample differed from the fitness measure from the original sample by more than 20% were eliminated. Fitness measures which remain constant over both the original and verification sample were called *duplicable*, suggesting they hold true for the entire database and were not a statistical anolmaly.

(Jacobson, 1996)

After this filtration process, all of the hypotheses in the >8.01 range, and all but one hypothesis in the 3.0-6.0 range have been intentionally eliminated due either to being outliers or being non-duplicable (according to the above standards).

Recall that the goal of the statistical package, was to return a value representing the interest of the given hypothesis, where "interesting" was defined as "combinations of RHS attributes (dependent variables) which are highly dependent on combinations of

LHS attributes (independent variables), or in other words, the candidate dependent variables are truly determined (not independent of) by the candidate independent variables." (Jacobson, 1996) Furthermore, in a genetic algorithm this interest value was to be represented by the fitness measure (modified j-measure, in this case).

Now, however, after the algorithm has completed running, whether intentionally or unintentionally, the authors have changed the definition of interesting. Interesting is no longer represented by simply the modified j-measure value, but the modified j-measure value subjected to the above filtration criteria. They are then left in the paradoxical situation that the best solutions on which the algorithm has converged are not interesting by the new definition. In a different paper (Bhargava and Jacobson, 1996), the authors write, "The problem in many forms of decision science is not whether a model performs accurately, but rather if it accurately represents the reality of the decision." Unfortunately, this problem has not yet been solved with DaMI. In other words, the algorithm may accurately search by the criteria in which it discriminates between competing solutions, but it does *not* accurately represent the reality of the decision.

This author performed another independent study in which the same LHS and RHS attributes were seeded, but a lower percentage was used for the seed. The pre-seeded and post-seeded contingency tables are shown below:

	'T'	'F'	modified j-measure
'T'	15	100	1.09
'F'	917	6714	

before seeding

	'T'	'F'	modified j-measure
'T'	84	31	3.98
'F'	917	6714	

after seeding

A total of nine runs was performed on this modified database. None of the runs located the seeded solution, and the algorithm yielded only weak reproducibility. Consider the seeded solution with a fitness measure of 3.98, still a very interesting solution both by fitness measure and by inspection of the contingency table. This value is sufficient to have placed it #1 in the top 100 hypotheses reported in Appendix C of Jacobson (1996), had it resided in the database at that time. Furthermore, the solution criteria were such that this hypothesis would survive the filtration outlined above. So far as the author knows, it was the best solution in the modified database by this criteria.

Let us now consider the hypothetical situation that DaMI had reproduced the results outlined in Jacobson (1996). Specifically, consider hypothetically that DaMI had converged similarly on the same high fitness measure (>8.01, though prior to the filtration) results upon which production runs 20-22 had converged. This would have yielded strong reproducibility according to the definition offered by Jacobson. What would this intuitively tell us about whether or not DaMI had located the seeded solution?

To answer this question, we performed a separate reproducibility analysis on the range of solutions offered in Appendix C of Jacobson (1996). As mentioned above, the seeded solution had a fitness measure of 3.98, higher than the #1 fitness measure reported in Appendix C (3.24). The #100 solution had a fitness measure of 2.15. The same program used to produce the "Exposure-to-Diagnosis Reproducibility" graphs was used to analyze production runs 20/21 and 20/22 for the range of fitness measures 2.15-3.98. The reproducibilities in this range were 8.61% and 9.01%, respectfully. While DaMI could hypothetically give reproducibility on the order of 90%-100% in the range of fitness measures >8.01, it was giving very low reproducibility in the range of fitness measures that could very likely contain the most interesting solutions. Consequently, even if strong reproducibility was a good indication of DaMI's effectiveness, it does not yield a lot of confidence that the solution space has been adequately searched *in the area where the most interesting hypotheses could likely reside.*

C. CONCLUSIONS AND RECOMMENDATIONS

It has been theoretically demonstrated that DaMI suffers from potentially severe limitations depending on the nature of the fitness landscape, and that a genetic algorithm may not be well suited for problems of this type. The inability of DaMI to locate a complex seeded solution in any of 19 experimental runs lends practical support to this conclusion. Though it is possible that a different fitness measure could overcome this weakness, none were found by these authors that did not suffer from other debilitating weaknesses.

It has also been demonstrated that DaMI does not accurately model the decision process. It is recommended that any criteria that will ultimately be used to determine the level of interest of a particular hypothesis also be included within the generational operation of the algorithm, so that DaMI is not biased towards uninteresting solutions.

It is the opinion of these authors that a the "brute force" method be reconsidered. The calculations reported by Jacobson (see Chapter II) involved two worst-case assumptions. Specifically, 1) all combinations of attributes were considered, no matter how unreasonable (e.g. 29 exposures combining to yield 15 diagnoses, 31 exposures combining to give 17 diagnoses, etc.), and 2) a relatively slow machine was used to perform the calculations. More reasonable assumptions about the nature of possible syndromes coupled with a more powerful machine would bring the feasibility of this method well within acceptable bounds, especially considering the months of effort that will be necessary to improve DaMI as it stands. This would also eliminate any uncertainty in the results.

LIST OF REFERENCES

Bhargava, Hemant K. and David L. Jacobson, "Genetic Algorithms in the Search for Desert Storm Syndrome," *Proceedings of the 1997 Conference of HealthCare Information and Management Systems Society (San Diego ,CA)*, February 1997.

Bhargava, Hemant K. and David L. Jacobson, "Exploratory Data Analysis with Genetic Algorithms: Is There a Gulf War Syndrome?" *Proceedings of the Second INFORMS Conference on Information Systems and Technology*, May 4-5, 1997.

CCEP, "CCEP Report on 18,598 Participants," *Comprehensive Clinical Evaluation Program for Gulf War Veterans*, April 2, 1996a.

CCEP, "Summary of Interview with Director, Deployment Surveillance Team and CCEP Epidemiologist," September 11, 1996b.

Corcoran, Arthur L. and Roger L. Wainwright "Using LibGA to Develop Genetic Algorithms for Solving Combinatorial Optimization Problems," *Practical Handbook of Genetic Algorithms: Applications*, 1994.

Dasgupta, Dipankar, "Handling Deceptive Problems Using a Different Genetic Search," *IEEE World Congress on Computational Intelligence*, pp. 807-11, Vol. 2, 1994.

Ermakov, S.M., and A.A. Zhiglyavskij, "On Random Search of Global Extremum," *Probability Theory and Applications*, Vol. 183, No. 1, 1983.

Goldberg, David E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989.

Holland, John H., *Adaptation in Natural and Artificial Systems*, The MIT Press, 1975.

Jacobson, David L., "Using Genetic Algorithms to Search Large, Unstructured Databases: The Search for Desert Storm Syndrome," *Unpublished Master's Thesis from Naval Postgraduate School*, September 1996.

Kauffman, S. A., "Adaptation on Rugged Fitness Landscapes," *Lectures in the Sciences of Complexity: Volume I in the Santa Fe Institute Studies in the Sciences of Complexity*, Addison-Wesley Publishing Company, 1989.

Kingdon, Jason, and Laura Dekker, "The Shape of Space," *Genetic Algorithms in Engineering Systems: Innovations and Applications*, 12-14 September 1995.

Kinnear, Kenneth E., Jr. "Fitness Landscapes and Difficulty in Genetic Programming," *IEEE World Congress on Computational Intelligence*, pp. 142-7, Vol. 1, 1994.

Koza, John R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, The MIT Press, 1990.

Qi, X.F., and F. Palmieri, "Theoretical Analysis of Evolutionary Algorithms with an Infinite Population Size in Continuous Space, Part I: Basic Property of Selection and Mutation," *IEEE Transactions on Neural Network*, Vol. 115, No. 1, 1994.

Renders, Jean-Michel, and Hugues Bersini, "Hybridizing Genetic Algorithms with Hill-Climbing Methods for Global Optimization: Two Possible Ways," *IEEE World Congress on Computational Intelligence*, pp. 312-17, Vol. 1, 1994.

Weishui, Wan, and Xiong Chen, "Convergence Theorem of Genetical Algorithm," *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 1996.

Wright, Sewell, "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution," *Proceedings of the Sixth International Congress on Genetics* **1**, 1932.

BIBLIOGRAPHY

Agapie, Alexandru, and Horia Dediu, "GA for Deceptive Problems: Inverting Schemata by a Statistical Approach," *Proceedings of 1996 IEEE International Conference on Evolutionary Computation*, pp. 336-40, 1996.

Bandyopadhyay, S., C.A. Murthy, and S.K. Pal, "GA-Based Pattern Classification: Theoretical and Experimental Studies," *Proceedings of ICPR '96*, 1996.

Behe, Michael J., *Darwin's Black Box*, The Free Press, 1996.

Chatfield, Chris, "Exploratory Data Analysis," *European Journal of Operational Research*, No. 23, 1986.

Courcelle, Richard J., and Leonard J. Tashman, "Box Plots: Another Graphical Aid in Forecasting," *The Journal of Business Forecasting*, Winter '88-'89.

Deutsch, Michael S., "An Exploratory Analysis Relating the Software Project Management Process to Project Success," *IEEE Transactions on Engineering Management*, Vol. 38, No. 4, November 1991.

Drazin, Robert, and Robert K. Kazanjian, "Applying the Del Technique to the Analysis of Cross-Classification Data: A Test of CEO Succession and Top Management Team Development," *Academy of Management Journal*, Vol. 36, No. 6, 1993.

Freisleben, Bernd, and Peter Merz, "A Genetic Local Search Algorithm for Solving Symmetric and Asymmetric Traveling Salesman Problems," *Proceedings of 1996 IEEE Conference on Evolutionary Computation*, pp. 616-21, 1996.

Gordon, V. Scott, "Locality in Genetic Algorithms," *IEEE World Congress on Computational Intelligence*, pp. 428-32, Vol. 1, 1994.

Horrocks, D.H., and M.C. Spittle, "Genetic Algorithms," *IEE Colloquium on Circuit Theory*, 1992.

Huang, Yiadong, and Shu-Park Chan, "Optimizing the Performance of the Genetic Algorithms for Finding the Optimal Value of a Given Function," *Proceedings of the 34th Midwest Symposium on Circuits and Systems*, pp. 819-22, Vol. 2.

Johnson, Phillip, "Darwin on Trial," *InterVarsity Press*, 2nd Edition, 1993.

Kozlov, Alexander V., and Jaswinder Pal Singh, "Parallel Implementations of Probabilistic Inference," *Computer*, December, 1996.

Krovi, Ravindra, "Genetic Algorithms for Clustering: A Preliminary Investigation," *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, pp. 540-4, Vol. 4, 1992.

Lopez, Luis R., "Are Deception and Complexity Conjugate Variables in Genetic Learning?" *IEEE World Congress on Computational Intelligence*, pp. 607-12, Vol. 2, 1994.

Mastracchio, Nicholas J., Jr., and Josef Schmee, "Analytical Procedures that Work," *The CPA Journal*, November, 1992.

Matsui, Kazuhiro, and Yukio Kosugi, "An Analysis on Genetic Algorithms Using Markov Process with Rewards," *Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pp. 129-38, 1996.

Potter, Walter D., John A. Miller, Ravi Gandham, and Chito N. Lapena, "Improving the Reliability of Heuristic Diagnosis," *Decision Aiding for Complex Systems*, pp. 725-30, Vol. 2, 1991.

Reeves, Colin R., and Christine C. Wright, "Genetic Algorithms and Statistical Methods: A Comparison," *Genetic Algorithms in Engineering Systems: Innovations and Applications*, September, 1995.

Smith, Richard, Stephen Warrington, and Frank Mill, "Shape Representation for Optimization," *Genetic Algorithms in Engineering Systems: Innovations and Applications*, IEEE, September, 1995.

Srinivas, M., and L.M. Patnaik, "Genetic Search: Analysis Using Fitness Moments," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 1, February, 1996.

Stuckman, B., G. Evans, and M. Mollaghasemi, "Comparison of Global Search Methods for Design Optimization Using Simulation," *Proceedings of the 1991 Winter Simulation Conference*, 1991.

Talbi, E-G., and T. Muntean, "Hill-Climbing, Simulated Annealing and Genetic Algorithms: A Comparative Study and Application to the Mapping Problem," *Proceedings of the 26th Hawaii International Conference on System Sciences*, pp. 565-73, Vol. 2, 1993.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center 2
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, Virginia 22060-6218

2. Dudley Knox Library 2
   Naval Postgraduate School
   411 Dyer Rd.
   Monterey, California 93943-5101

3. United States General Accounting Office 2
   National Security and International Affairs Division
   Washington, D.C. 20548

4. Prof. Hemant Bhargava, Code SM/BH 4
   Department of Systems Management
   Naval Postgraduate School
   Monterey, California 93943

5. Dr. William J. Haga 1
   Department of Systems Management
   Naval Postgraduate School
   Monterey, California 93943

6. LT Jason W. Carter. 3
   870 Northridge
   Pacific, MO 63069

7. LT David L. Jacobson. 1
   Naval Medical Information Management Center
   8901 Wisconsin Avenue
   Building 27
   Bethesda, MD 20889-5605